РАЗБЕРИСЬ В

DATA SCIENCE

КАК ОСВОИТЬ НАУКУ О ДАННЫХ И НАУЧИТЬСЯ ДУМАТЬ КАК ЭКСПЕРТ

- МЕТОДЫ АНАЛИТИКИ ДАННЫХ
- ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ
- РАСПРОСТРАНЕННЫЕ ОШИБКИ ПРИ РАБОТЕ С ДАННЫМИ

АЛЕКС ДЖ. ГАТМАН
ДЖОРДАН ГОЛДМЕЙЕР

БОМБОРА
ИЗДАТЕЛЬСТВО

ALEX J. GUTMAN
JORDAN GOLDMEIER

BECOMING A DATA HEAD

HOW TO THINK, SPEAK, AND UNDERSTAND DATA SCIENCE, STATISTICS, AND MACHINE LEARNING

АЛЕКС ДЖ. ГАТМАН
ДЖОРДАН ГОЛДМЕЙЕР

РАЗБЕРИСЬ В DATA SCIENCE

КАК ОСВОИТЬ НАУКУ О ДАННЫХ И НАУЧИТЬСЯ ДУМАТЬ КАК ЭКСПЕРТ

- **МЕТОДЫ АНАЛИТИКИ ДАННЫХ**
- **ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ**
- **РАСПРОСТРАНЕННЫЕ ОШИБКИ ПРИ РАБОТЕ С ДАННЫМИ**

БОМБОРА
ИЗДАТЕЛЬСТВО
Москва 2023

УДК 004.4
ББК 32.973.2-018
Г23

Jordan Goldmeier, Alex J. Gutman
BECOMING A DATA HEAD:
How to Think, Speak and Understand Data Science, Statistics and Machine Learning

Гатман, Алекс Дж.

Г23 Разберись в Data Science : как освоить науку о данных и научиться думать как эксперт / Алекс Дж. Гатман, Джордан Голдмейер ; [перевод с английского М. А. Райтман]. — Москва : Эксмо, 2023. — 304 с. — (Мировой компьютерный бестселлер).

ISBN 978-5-04-174810-4

Перед вами исчерпывающее руководство по основам Data Science. С помощью него вы сможете научиться мыслить статистически и понимать, какую роль в вашей работе играет аналитика, пользоваться языком науки о данных, избегать распространенных ошибок при работе с ними и, наконец, разобраться в полезных инструментах, которые используют эксперты.

УДК 004.4
ББК 32.973.2-018

ISBN 978-5-04-174810-4

Посвящается моим детям Элли, Уильяму и Эллен.
Элли было три года, когда она узнала, что ее папа — «доктор».
Озадаченно посмотрев на меня, она сказала: «Но ведь ты
не помогаешь людям…»
Памятуя об этом, я также посвящаю эту книгу вам, читатель.
Надеюсь, что она вам поможет.
— Алекс

Посвящается Стивену и Мелиссе.
— Джордан

Оглавление

Предисловие 15

Введение 19
Промышленный комплекс науки о данных 19
Почему нам это важно 20
Кризис субстандартного ипотечного кредитования 20
Всеобщие выборы в США 2016 года 22
Наша гипотеза 23
Данные на рабочем месте 24
Сцена в зале заседаний 24
Вы можете понять общую картину 26
Классификация ресторанов 26
Что дальше? 29
Для кого написана эта книга? 30
Зачем мы написали эту книгу 32
Что вы узнаете 33
Как организована эта книга 34
Прежде чем мы начнем 35

ЧАСТЬ I
Думайте как главный по данным

ГЛАВА 1
В чем суть проблемы? 39
Вопросы, которые должен задать главный по данным 40
Почему эта проблема важна? 41
Кого затрагивает эта проблема? 42
Что, если у нас нет нужных данных? 43
Когда проект будет завершен? 44
Что, если нам не понравятся результаты? 44
Причины провала проектов по работе с данными 45
Клиентское восприятие 45
Обсуждение 47
Работа над значимыми проблемами 48
Подведение итогов 49

ГЛАВА 2
Что такое данные? 51
Данные и информация 51
Пример набора данных 52
Типы данных 53
Сбор и структурирование данных 55
Данные наблюдений и экспериментальные данные 55
Структурированные и неструктурированные данные 56
Основы сводной статистики 57
Подведение итогов 58

ГЛАВА 3
Готовьтесь мыслить статистически 60
Задавайте вопросы 61
Во всем есть вариации 63
Сценарий: Клиентское восприятие (продолжение) 64
Анализ реальной ситуации: показатели заболеваемости раком почки 67
Вероятности и статистика 69
Вероятность и интуиция 71
Открытия с помощью статистики 73
Подведение итогов 75

ЧАСТЬ II
Говорите как главный по данным

ГЛАВА 4
Сомневайтесь в данных 79
Что бы вы сделали? 80
Катастрофа, вызванная недостатком данных 82
Расскажите мне историю происхождения данных 87
Кто собирал данные? 87
Как собирались эти данные? 88
Являются ли данные репрезентативными? 89
Имеет ли место предвзятость выборки? 90
Что вы сделали с выбросами? 90
Какие данные я не вижу? 91
Как вы поступили с отсутствующими значениями? 91
Позволяют ли данные измерить то, что вас интересует? 92
Сомневайтесь в данных любого размера 93
Подведение итогов 93

ГЛАВА 5
Исследуйте данные 94
Разведочный анализ данных и вы 95
Освоение исследовательского образа мышления 96
Направляющие вопросы 96
Сценарий 97
Позволяют ли данные ответить на поставленный вопрос? 97
Определитесь с ожиданиями и руководствуйтесь здравым смыслом 97
Имеют ли данные интуитивный смысл? 98
Осторожно: выбросы и отсутствующие значения 102
Обнаружили ли вы какие-либо взаимосвязи? 103
Корреляция 104
Осторожно: неверная интерпретация корреляции 105
Осторожно: корреляция не означает причинность 107
Обнаружили ли вы новые возможности в данных? 108
Подведение итогов 109

ГЛАВА 6
Изучайте вероятности 110
Попробуйте угадать 111
Правила игры 112
Нотация 112
Условная вероятность и независимые события 114
Вероятность наступления множества событий 115
Одновременное наступление двух событий 115
Наступление одного или другого события 117
Мысленное упражнение на определение вероятности 119
Дальнейшие шаги 120
Будьте осторожны, делая предположения о независимости событий 121
Не допускайте ошибку игрока 122
Все вероятности являются условными 123
Не меняйте зависимости местами 123
Теорема Байеса 125
Убедитесь, что вероятности имеют смысл 128
Калибровка 128
Редкие события могут случаться и случаются 129
Подведение итогов 130

ГЛАВА 7
Бросайте вызов статистике **131**
Краткие уроки по статистическому выводу 131
Обеспечьте себе простор для маневра 132
Больше данных — больше доказательств 133
Бросьте вызов статус-кво 133
Доказательства обратного 135
Сбалансируйте ошибки, допускаемые при принятии решений 137
Процесс построения статистического вывода 139
Вопросы, позволяющие бросить вызов статистическим показателям 140
Каков контекст этой статистики? 140
Каков размер выборки? 141
Что вы тестируете? 142
Какова нулевая гипотеза? 142
Допущение эквивалентности 144
Каков уровень значимости? 144
Сколько тестов вы проводите? 145
Каковы доверительные интервалы? 146
Имеет ли это практическое значение? 147
Предполагаете ли вы наличие причинно-следственной связи? 148
Подведение итогов 149

ЧАСТЬ III
Освойте набор инструментов дата-сайентиста

ГЛАВА 8
Ищите скрытые группы **153**
Обучение без учителя 154
Снижение размерности 155
Создание составных признаков 155
Анализ главных компонент 157
Главные компоненты спортивных способностей 158
Анализ главных компонент. Резюме 162
Потенциальные ловушки 163
Кластеризация 164
Кластеризация методом k-средних 165
Кластеризация точек продаж 166
Потенциальные ловушки 168
Подведение итогов 169

ГЛАВА 9
Освойте модели регрессии 171
Обучение с учителем 171
Линейная регрессия: что она делает 174
Регрессия методом наименьших квадратов: больше, чем умное название 175
Линейная регрессия: что она дает 179
Включение множества признаков 180
Линейная регрессия: какую путаницу она вызывает 181
Пропущенные переменные 182
Мультиколлинеарность 183
Утечка данных 184
Ошибки экстраполяции 185
Многие взаимосвязи не являются линейными 186
Вы объясняете или предсказываете? 186
Производительность регрессионной модели 187
Прочие модели регрессии 189
Подведение итогов 189

ГЛАВА 10
Освойте модели классификации 191
Введение в классификацию 191
Чему вы научитесь 192
Постановка задачи классификации 193
Логистическая регрессия 194
Логистическая регрессия: что дальше? 197
Деревья решений 199
Ансамблевые методы 203
Случайные леса 203
Деревья решений с градиентным усилением 204
Интерпретируемость ансамблевых моделей 206
Остерегайтесь ловушек 206
Неправильное определение типа задачи 207
Утечка данных 207
Отсутствие разделения данных 208
Выбор неправильного порогового значения для принятия решения 208
Неправильное понимание точности 209
Матрицы ошибок 210
Подведение итогов 212

ГЛАВА 11
Освойте текстовую аналитику 214
Ожидания от текстовой аналитики 214
Как текст превращается в числа 216
Большой мешок слов 216
N-граммы 221
Векторное представление слов 222
Тематическое моделирование 225
Классификация текстов 227
Наивный байесовский алгоритм 229
Анализ настроений 232
Практические соображения при работе с текстом 233
Преимущества технологических гигантов 234
Подведение итогов 235

ГЛАВА 12
Концептуализируйте глубокое обучение 237
Нейронные сети 238
Чем нейронные сети похожи на мозг? 238
Простая нейронная сеть 239
Как учится нейронная сеть 241
Чуть более сложная нейронная сеть 242
Применение глубокого обучения 245
Преимущества глубокого обучения 247
Как компьютеры «видят» изображения 249
Сверточные нейронные сети 250
Глубокое обучение для обработки языка и последовательностей 252
Глубокое обучение на практике 254
Есть ли у вас данные? 254
Являются ли ваши данные структурированными? 256
Как будет выглядеть сеть? 256
Искусственный интеллект и вы 257
Преимущества технологических гигантов 258
Этический аспект глубокого обучения 259
Подведение итогов 261

ЧАСТЬ IV

Гарантируйте успех

ГЛАВА 13
Остерегайтесь ловушек **265**
Предвзятости и странности в данных 266
Систематическая ошибка выжившего 266
Регрессия к среднему 267
Парадокс Симпсона 268
Предвзятость подтверждения 270
Ловушка невозвратных затрат 270
Алгоритмическая предвзятость 271
Прочие предубеждения 272
Большой список ловушек 272
Ловушки статистики и машинного обучения 272
Ловушки проекта 274
Подведение итогов 277

ГЛАВА 14
Знайте людей и типы личностей **278**
Семь сцен коммуникативного сбоя 279
Постмортем 280
Время историй 280
Игра «Телефон» 281
В дебри 282
Проверка реальности 282
Захват власти 283
Хвастун 283
Отношение к данным 284
Энтузиасты 284
Циники 285
Скептики 285
Подведение итогов 286

ГЛАВА 15
Что дальше? **287**

Об авторах **290**

О технических редакторах **291**

Благодарности **293**

Предметный указатель **296**

Предисловие

Книга «Разберись в Data Science» вышла очень своевременно, учитывая текущую ситуацию с данными и аналитикой в организациях. Давайте кратко пробежимся по последним событиям. Начиная с 1970-х годов лишь немногие передовые компании эффективно использовали данные и аналитику для принятия решений и обоснования своих действий. Большинство игнорировало этот ценный ресурс или не придавало ему особого значения.

В 2000-х годах ситуация стала меняться, и компании начали понимать, как они могут изменить свою ситуацию с помощью данных и аналитики. К началу 2010-х годов интерес стал смещаться в сторону «больших данных», которые изначально появились в интернет-компаниях, а затем распространились по всей экономике. В связи с возросшим объемом и сложностью данных в компаниях возникла роль «дата-сайентиста», опять же, сначала в Силиконовой долине, а затем повсюду.

Однако как только фирмы начали приспосабливаться к большим данным, в период с 2015 по 2018 год акцент во многих фирмах снова сместился, на этот раз в сторону искусственного интеллекта. Сбор, хранение и анализ больших данных уступили место машинному обучению, обработке естественного языка и автоматизации.

В основе этих быстрых сдвигов фокуса лежал ряд допущений относительно данных и аналитики, распространенных внутри организаций. Я рад сообщить, что книга «Разберись в Data Science» разрушает многие из них и делает это весьма своевременно. Многие люди, внимательно наблюдающие за этими тенденциями, уже начинают признавать, что эти допущения направляют нас по непродуктивному пути. В оставшейся части этого предисловия я опишу пять взаимосвязанных допущений и то, как изложенные в этой книге идеи обоснованно опровергают их.

Допущение 1. Аналитика, большие данные и ИИ — совершенно разные явления.

Многие полагают, что «традиционная» аналитика, большие данные и ИИ — это отдельные явления. Однако авторы книги «Разберись в Data Science»

справедливо считают, что эти вещи тесно связаны друг с другом. Все они требуют статистического мышления, использования традиционных аналитических подходов, вроде регрессионного анализа, а также методов визуализации данных. Предиктивная аналитика — это, по сути, то же самое, что и контролируемое машинное обучение. Кроме того, большинство методов анализа данных работают с наборами данных любого размера. Короче говоря, главный по данным может эффективно работать во всех трех областях, так что заострять внимание на различиях между ними не очень продуктивно.

Допущение 2. В этой песочнице могут играть только дата-сайентисты. Мы часто прославляли дата-сайентистов, полагая, что только они способны эффективно работать с данными и аналитикой. Тем не менее в настоящее время зарождается важная тенденция к демократизации этих идей, и все больше организаций расширяют полномочия «гражданских специалистов по работе с данным». Автоматизированные инструменты машинного обучения упрощают создание моделей, которые отлично справляются с прогнозированием. Разумеется, нам все еще нужны профессиональные дата-сайентисты для разработки новых алгоритмов и проверки работы гражданских специалистов, занимающихся сложным анализом. Однако организации, которые демократизируют занятие аналитикой и наукой о данных, привлекая к этому «любителей», способны значительно расширить использование этих важных возможностей.

Допущение 3. Дата-сайентисты — это единороги, обладающими всеми необходимыми навыками.
Мы привыкли полагать, что дата-сайентисты, умеющие разрабатывать модели, также способны решать все остальные задачи, связанные с внедрением этих моделей. Другими словами, мы считаем их своеобразными «единорогами», которые могут все. Но таких «единорогов» нет вообще, или они существуют лишь в небольшом количестве. Главные по данным, которые понимают не только основы науки о данных, но и особенности бизнеса, а также способны эффективно управлять проектами и выстраивать деловые отношения, будут чрезвычайно ценны как участники проектов по работе с данными. Они могут стать продуктивными членами команд дата-сайентистов и повысить вероятность того, что проекты по работе с данными принесут бизнесу пользу.

Допущение 4. Чтобы преуспеть в работе с данными и аналитикой, вам необходимы выдающиеся математические способности и много тренировок.

Еще одно похожее допущение сводится к тому, что для работы с данными человек должен быть очень хорошо подготовлен в этой области, а также хорошо разбираться в математике. Математические способности и подготовка, безусловно, очень важны, но авторы книги «Разберись в Data Science» утверждают (и я с ними согласен), что мотивированный ученик способен освоить необходимые навыки в достаточной степени для того, чтобы стать полезным участником проектов по работе с данными. Во-первых, общие принципы статистического анализа далеко не так сложны, как может показаться. Во-вторых, для того, чтобы «быть полезным» участником проектов по работе с данными, ваш уровень владения аналитикой не обязательно должен быть чрезвычайно высоким. Работа с профессиональными дата-сайентистами или автоматизированными ИИ-программами требует лишь любознательности и умения задавать хорошие вопросы, находить взаимосвязи между бизнес-проблемами и количественными результатами, а также обращать внимание на сомнительные предположения.

Допущение 5. Если в колледже или аспирантуре вы не занимались в основном количественными предметами, вам слишком поздно осваивать навыки, необходимые для работы с данными и аналитикой.

Это предположение подтверждается данными опросов. Согласно результатам опроса, проведенного компанией Splunk в 2019 году, в котором приняли участие около 1300 руководителей по всему миру, практически каждый респондент (98%) согласен с тем, что навыки работы с данными важны для специалистов будущего[1]. А 81% респондентов считает, что навыки работы с данными необходимы для того, чтобы стать старшим руководителем в их компаниях, а 85% согласны с тем, что ценность таких навыков в их фирмах будет расти. Тем не менее 67% респондентов заявили, что им неудобно получать доступ к данным или использовать их самостоятельно, 73% считают, что навыки работы с данными труднее освоить, чем другие бизнес-навыки, а 53% — что они слишком стары для освоения навыков работы с данными. Подобное пораженчество наносит ущерб как отдельным лицам,

[1] Splunk Inc., “The State of Dark Data,” 2019, **www.splunk.com/en_us/form/the-state-of-dark-data.html**.

так и организациям в целом, и ни авторы этой книги, ни я не считаем его оправданным. В ходе чтения этой книги вы увидите, что в этом нет ничего сложного!

Итак, отбросьте эти ложные допущения и станьте главным по данным. Это позволит вам повысить свою ценность как сотрудника и сделать свою организацию более успешной. Именно по этому пути движется мир, так что пришло время узнать больше о данных и аналитике. Я уверен, что процесс чтения книги «Разберись в Data Science» окажется гораздо более полезным и приятным, чем вы можете себе представить.

Томас Х. Дэвенпорт
Заслуженный профессор Бэбсон-колледжа, приглашенный профессор Бизнес-школы Саида при Оксфордском университете, научный сотрудник инициативы Массачусетского технологического института в сфере цифровой экономики, автор книг «Аналитика как конкурентное преимущество», «Внедрение искусственного интеллекта в бизнес-практику: Преимущества и сложности» и «Big Data @ Work»

Введение

Данные — это, пожалуй, важнейший аспект вашей работы, нравится вам это или нет. И, скорее всего, вы решили прочитать эту книгу, чтобы лучше в них разобраться.

Для начала стоит констатировать то, что уже почти превратилось в клише: в настоящее время мы создаем и потребляем больше информации, чем когда-либо прежде. Мы, без сомнения, живем в эпоху данных, которая породила массу обещаний, модных словечек и продуктов, многие из которых вы, ваши менеджеры, коллеги и подчиненные уже используете или будете использовать. Однако, несмотря на распространение этих обещаний и продуктов, проекты по работе с данными терпят неудачу с пугающей регулярностью[2].

Разумеется, мы не утверждаем, что все обещания пусты, а продукты — ужасны. Скорее, чтобы по-настоящему разобраться в этой области, вы должны принять фундаментальную истину: работа с данными очень сложна и сопряжена с нюансами и неопределенностью. Данные, безусловно, важны, но работать с ними совсем не просто. И все же существует целая индустрия, которая заставляет нас думать иначе, обещает определенность в мире неопределенности и играет на страхе компаний упустить выгоду. Мы называем это промышленным комплексом науки о данных.

ПРОМЫШЛЕННЫЙ КОМПЛЕКС НАУКИ О ДАННЫХ

Эта проблема касается всех. Компании бесконечно ищут продукты, которые думали бы за них. Менеджеры нанимают профессионалов в области аналитики, которые на самом деле таковыми не являются. Дата-сайентистов нанимают для работы в компаниях, которые к ним не готовы. Руководители вынуждены слушать техническую болтовню и делать вид, что понимают, о чем идет речь. Работа над проектами стопорится. Деньги тратятся впустую.

---

[2] Venture Beat. “87% of data science projects failing”: **venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production**

Тем временем промышленный комплекс науки о данных штампует новые концепции быстрее, чем мы можем определить и сформулировать порождаемые ими возможности (и проблемы). Стоит моргнуть, и обязательно что-нибудь пропустишь. Когда авторы этой книги начали работать вместе, все говорили о больших данных. Со временем популярной новой темой стала наука о данных. Затем внимание общественности сосредоточилось на машинном обучении, глубоком обучении и искусственном интеллекте.

Но самых любознательных и критически мыслящих из нас что-то не устраивает. Действительно ли эти проблемы новые? Или они просто переосмысление старых?

Ответ на оба вопроса утвердительный.

Однако мы надеемся, что вы задаетесь более важным вопросом — «Как научиться критически мыслить и говорить о данных?»

Мы вас этому научим.

В этой книге вы познакомитесь с инструментами, терминами и образом мышления, необходимыми для навигации по промышленному комплексу науки о данных. Вы научитесь понимать данные и связанные с ними проблемы на более глубоком уровне, критически относиться к данным и результатам, с которыми сталкиваетесь, а также разумно говорить обо всем, что касается данных.

Короче говоря, вы станете главным по данным.

ПОЧЕМУ НАМ ЭТО ВАЖНО

Прежде чем мы начнем, стоит сказать, почему авторов этой книги, Алекса и Джордана, так волнует эта тема. В этом разделе мы опишем два важных примера того, как данные повлияли на общество в целом и на нас лично.

Кризис субстандартного ипотечного кредитования

Мы едва закончили колледж, когда разразился кризис субстандартного ипотечного кредитования. Мы оба устроились на работу в BBC в 2009 году, когда найти работу было очень трудно. Нам повезло, поскольку мы обладали востребованным навыком — мы умели работать с данными. Мы каждый день работали над преобразованием результатов исследований, проведенных аналитиками и учеными BBC, в продукты, которые могло бы использовать правительство. Наш прием на работу стал предвестником грядущего роста важности тех ролей, которые мы исполняли. Будучи специалистами

по работе с данными, мы наблюдали за развитием ипотечного кризиса с интересом и любопытством.

У кризиса субстандартного ипотечного кредитования было множество причин[3]. Приводя его здесь в качестве примера, мы не отрицаем прочие факторы, однако, по нашему мнению, важнейшим из них была серьезная проблема с данными. Банки и инвесторы создали модели для оценки ценности обеспеченных ипотекой долговых обязательств (CDO) — инвестиционных инструментов, ставших причиной обвала рынка США.

Облигации с ипотечным покрытием считались безопасными инструментами, поскольку распределяли риск дефолта по кредиту между несколькими инвестиционными единицами. Идея заключалась в том, что если лишь некоторые активы в портфеле ипотечных кредитов окажутся убыточными, это не окажет существенного влияния на стоимость всего портфеля.

И все же, если поразмыслить, становится очевидно, что некоторые фундаментальные предположения были неверны. В первую очередь речь идет о допущении независимости между возможными дефолтами, то есть предположении о том, что если заемщик А не выполнит обязательства по кредиту, это не повлияет на риск неплатежа заемщика Б. Впоследствии мы узнали о том, что дефолты происходят по принципу домино, то есть предыдущий дефолт может предсказать вероятность дальнейших дефолтов. Дефолт по одному ипотечному кредиту приводил к снижению стоимости находящейся поблизости недвижимости, что способствовало росту риска дефолта по соответствующим кредитам. По сути, один дом утягивал за собой соседние.

Допущение независимости фактически связанных между собой событий — распространенная ошибка в статистике.

Но давайте углубимся в эту историю. Инвестиционные банки создали модели, которые переоценили эти инвестиции. Модели, о которых мы поговорим далее в книге, — это упрощенные версии реальности. Они используют предположения о реальном мире для понимания и предсказания определенных явлений.

А кто создавал эти модели? Это были люди, которые заложили основы будущей профессии дата-сайентиста. Люди вроде нас. Статистики, экономисты, физики — люди, которые занимались машинным обучением, искусственным интеллектом и статистикой. Они работали с данными. И они были умны. Невероятно умны.

---

[3] **www.brookings.edu/wp-content/uploads/2016/06/11_origins_crisis_baily_litan.pdf**

И все же что-то пошло не так. Может быть, они не сумели задать правильные вопросы? Или информация о риске и неопределенности не была должным образом донесена до лиц, принимающих решения, в результате чего у них возникла иллюзия совершенно предсказуемого рынка недвижимости? А может быть, кто-то откровенно соврал о результатах?

Но больше всего нас интересовало то, как избежать подобных ошибок в нашей собственной работе?

У нас было много вопросов, и об ответах мы могли лишь гадать, но одно было ясно — это была крупномасштабная катастрофа с данными. И она обещала быть не последней.

Всеобщие выборы в США 2016 года

8 ноября 2016 года кандидат от республиканцев Дональд Дж. Трамп победил на всеобщих выборах в Соединенных Штатах, обойдя предполагаемого лидера и кандидата от демократической партии Хиллари Клинтон. Для политических социологов это стало настоящим шоком, поскольку их модели не предсказывали его победу. А год был самым подходящим для подобных предсказаний.

В 2008 году Нейт Сильвер, автор блога FiveThirtyEight, тогда бывшего частью газеты *The New York Times*, проделал фантастическую работу и предсказал победу Барака Обамы. В то время эксперты скептически относились к способности его алгоритма прогнозирования точно предсказывать результаты выборов. В 2012 году Нейт Сильвер снова оказался в центре внимания, предсказав очередную победу Обамы.

К этому моменту деловой мир уже начал осваивать работу с данными и нанимать дата-сайентистов. Успешное предсказание переизбрания Барака Обамы Нейтом Сильвером лишь подчеркнуло важность и оракулоподобные возможности прогнозирования на основе данных. Статьи в деловых журналах предостерегали руководителей о том, что если они не освоят работу с данными, то проиграют в конкурентной борьбе. Промышленный комплекс науки о данных заработал в полную силу.

К 2016 году каждое крупное новостное издание вложило средства в алгоритм предсказания исхода всеобщих выборов. Подавляющее большинство из них прогнозировали сокрушительную победу кандидата от демократической партии Хиллари Клинтон. Как же они ошибались.

Давайте сравним эту ошибку с кризисом субстандартного ипотечного кредитования. Можно было бы утверждать, что мы многому научились и что

интерес к науке о данных должен был бы позволить избежать ошибок прошлого. Действительно, начиная с 2008 года, новостные организации стали нанимать дата-сайентистов, вкладывать средства в проведение опросов общественного мнения, формировать команды аналитиков и тратить большое количество денег на сбор качественных данных.

Что же произошло, учитывая все это время, деньги, усилия и образование?[4]

Наша гипотеза

Почему возникают подобные проблемы с данными? Мы видим три причины: сложность проблемы, недостаток критического мышления и плохая коммуникация.

Во-первых (как мы уже говорили), работа с данными зачастую очень сложна. Даже при наличии большого количества данных, подходящих инструментов, методик и умнейших аналитиков случаются ошибки. Прогнозы могут и будут оказываться ошибочными. И это не критика данных и статистики. Такова реальность.

Во-вторых, некоторые аналитики и заинтересованные стороны перестали критически относиться к проблемам данных. Промышленный комплекс науки о данных в своем высокомерии нарисовал картину уверенности и простоты, и некоторые люди на нее купились. Возможно, такова человеческая природа: люди не хотят признавать, что не знают будущего. Однако ключевым аспектом правильного осмысления и использования данных является признание возможности принятия неверного решения. Это означает понимание и распространение информации о рисках и неопределенностях. Но эта идея где-то затерялась. Мы надеялись, что колоссальный прогресс в исследованиях и методах анализа и работы с данными обострит критическое мышление каждого человека, но, судя по всему, некоторые люди его, наоборот, отключили.

Третья причина возникновения проблем с данными, по нашему мнению, — плохая коммуникация между дата-сайентистами и лицами, принимающими решения. Даже при наличии самых лучших намерений результаты зачастую доносятся с искажениями. Лица, принимающие решения,

[4] Нейт Сильвер написал по этому поводу целую серию статей (**fivethirtyeight.com/tag/the-real-story-of-2016**). Одна из ошибок социологов заключалась в допущении независимости событий, как и в случае с ипотечным кризисом.

не говорят на языке данных, потому что никто не удосужился их этому научить. Кроме того, специалисты по работе с данными далеко не всегда способны понятно объяснить те или иные вещи. Итак, существует пробел в общении.

ДАННЫЕ НА РАБОЧЕМ МЕСТЕ

Ваши проблемы с данными, скорее всего, не грозят обрушением мировой экономики или неправильным предсказанием результатов следующих президентских выборов в США, но контекст этих историй имеет значение. Если недопонимание и ошибки в критическом мышлении случаются на глазах у всего мира, то, вероятно, это происходит на вашем рабочем месте. В большинстве случаев эти микросбои укрепляют культуру безграмотности в отношении данных.

Это происходило и на нашем рабочем месте и отчасти по нашей вине.

Сцена в зале заседаний

Поклонникам научной фантастики и приключенческих фильмов хорошо знакома такая сцена: герой сталкивается, казалось бы, с нерешаемой задачей, и мировые лидеры и ученые собираются вместе, чтобы обсудить ситуацию. Один из ученых, самый занудный среди всей группы, предлагает идею, используя непонятный жаргон, а генерал обрывает его, требуя «говорить по-человечески». После этого зритель получает некоторое объяснение того, что имелось в виду. Суть этого момента — преобразование критически важной для миссии информации в то, что способен понять не только наш герой, но и зритель.

Мы часто обсуждали этот сюжет в контексте нашей роли исследователей для федерального правительства. Почему? Потому что нам казалось, что ситуация никогда не разворачивалась таким образом. На ранних этапах нашей карьеры мы часто наблюдали нечто противоположное.

Мы представляли нашу работу людям, смотревшим на нас пустыми глазами, которые вяло кивали, а иногда почти засыпали. Мы наблюдали за тем, как сбитые с толку зрители воспринимали все, что мы говорили, без единого вопроса. Их либо впечатляло то, какими умными мы казались, либо им было скучно, потому что они ничего не понимали. Никто не просил повторить сказанное на понятном всем языке. Очень часто ситуация разворачивалась следующим образом:

Мы: «Проведя анализ бинарной переменной отклика методом контролируемого обучения с использованием множественной логистической регрессии, мы получили вневыборочную производительность со специфичностью 0,76 и несколько статистически значимых независимых переменных с использованием значений альфа равных 0,05».

*Бизнес-профессионал: *неловкое молчание**

Мы: «Это понятно?»

*Бизнес-профессионал: *снова тишина**

Мы: «Есть вопросы?»

Бизнес-профессионал: «В данный момент вопросов нет».

Внутренний монолог бизнес-профессионала: «О чем, черт возьми, они говорят?»

Увидев подобную сцену в кино, вы могли бы подумать: надо перемотать назад, возможно, я что-то упустил. Но в реальной жизни, когда принимаемые решения имеют огромное влияние на результат миссии, такое случается редко. Мы не перематываем. Мы не просим разъяснений.

Оглядываясь назад, мы понимаем, что наши презентации были слишком техническими. Отчасти причина заключалась в банальном упрямстве: до ипотечного кризиса технические детали чрезмерно упрощались; аналитиков приглашали для того, чтобы они говорили руководителям то, что те хотели услышать, но мы не собирались играть в эту игру. Мы хотели, чтобы наши зрители понимали нас.

Но мы перестарались. Наша аудитория не могла критически осмыслить результаты нашей работы, потому что не понимала, о чем мы говорили.

Мы подумали, что должен быть способ получше. Мы хотели повлиять на ситуацию с помощью своей работы, поэтому начали практиковаться в объяснении сложных статистических концепций друг другу и нашим зрителям, а также исследовать то, как наши объяснения воспринимают другие люди.

Нам удалось обнаружить точку соприкосновения между специалистами по работе с данными и бизнес-профессионалами, в которой могут иметь место честные дискуссии о данных, не будучи при этом слишком техническими или слишком упрощенными. Это предполагает более критическое

отношение обеих сторон к проблемам данных вне зависимости от их масштаба. Именно об этом и пойдет речь в этой книге.

ВЫ МОЖЕТЕ ПОНЯТЬ ОБЩУЮ КАРТИНУ

Для лучшего понимания данных и работы с ними вам необходимо быть готовым к изучению сложных концепций. И даже если вы уже знакомы с ними, мы научим вас тому, как донести их до вашей аудитории.

Вам также предстоит принять такой редко обсуждаемый факт, что во многих компаниях работа с данными оказывается неэффективной. Вы разовьете интуицию, понимание и здоровый скептицизм в отношении чисел и терминов, с которыми сталкиваетесь. Эта задача может показаться сложной, но эта книга поможет вам ее решить. И для этого вам не понадобятся ни навыки программирования, ни докторская степень.

С помощью четких объяснений, мысленных упражнений и аналогий вы сможете выстроить ментальную модель для понимания науки о данных, статистики и машинного обучения.

В следующем примере мы сделаем именно это.

Классификация ресторанов

Представьте, что вы идете по улице и видите пустую витрину с вывеской «Новый ресторан: скоро открытие». Вы устали питаться в сетевых ресторанах и постоянно ищете новые местные заведения, поэтому задаетесь вопросом: «Появится ли здесь новый независимый ресторан?»

Давайте поставим этот вопрос более формально: как вы думаете, будет ли новый ресторан сетевым или независимым?

Угадайте. (Серьезно, подумайте об этом, прежде чем двигаться дальше.)

В реальной жизни вы сделали бы довольно хорошее предположение за доли секунды. Находясь в модном районе с множеством местных пабов и закусочных, вы бы предположили, что ресторан будет независимым. А если бы речь шла о межштатной автомагистрали с расположенным рядом торговым центром, вы бы предположили, что ресторан будет сетевым.

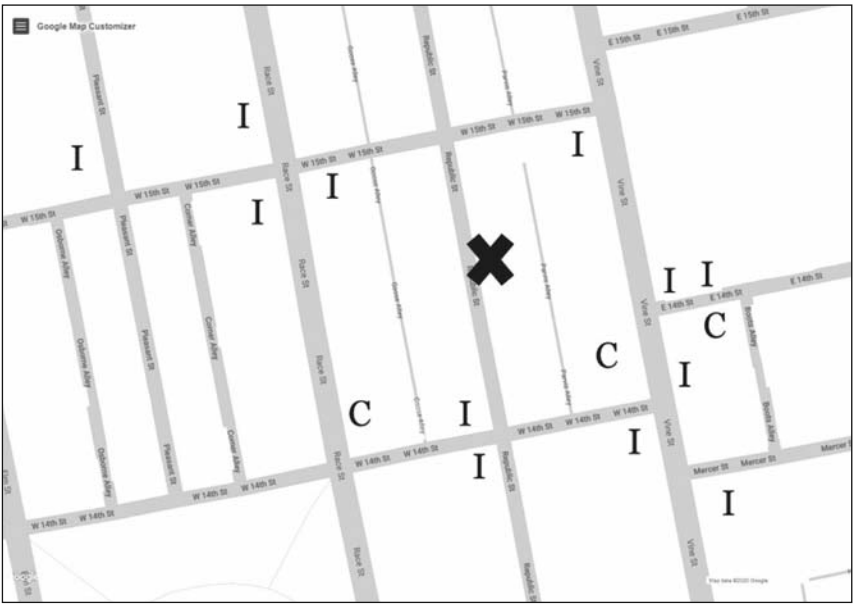
Но когда мы задали вопрос, вы заколебались. Вы подумали, что мы предоставили недостаточно информации. И вы были правы. Мы не предоставили вам никаких данных для принятия решения.

Мораль: для принятия обоснованных решений требуются данные.

Теперь посмотрите на первое изображение на следующей странице. Новый ресторан отмечен крестиком (X), буквой *C* обозначены сетевые рестораны (chain), а буквой *I* — независимые (independent) местные закусочные. Какое предположение вы сделали бы на этот раз?

Большинство людей предполагает, что ресторан будет независимым (I), потому что такова большая часть близлежащих ресторанов. Однако обратите внимание на то, что независимыми являются далеко не все из них. Если бы мы попросили вас оценить уровень достоверности[5] вашего прогноза в диапазоне от 0 до 100, то она, скорее всего, была бы высокой, но не равной 100, поскольку по соседству вполне может появиться еще один сетевой ресторан.

Мораль заключается в следующем: предсказания никогда не могут быть на 100% достоверными.

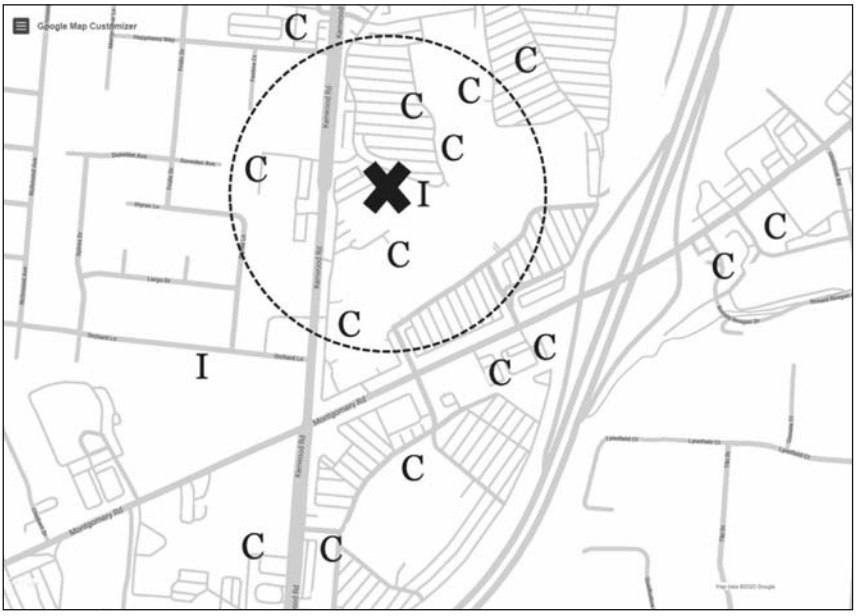


Район Овер-Райн, Цинциннати, штат Огайо

Теперь взгляните на следующее изображение. В этом районе есть большой торговый центр, и большинство ресторанов здесь — сетевые. Когда людям

[5] Примечание для коллег-статистиков: мы имеем в виду обычную, а не статистическую достоверность.

предлагается предсказать, каким будет новый ресторан в этом районе — сетевым или независимым, большинство выбирает вариант (C). Но нам нравится, когда кто-то выбирает вариант (I), потому что это подчеркивает несколько важных моментов.



Кенвуд Таун Центр, Цинциннати, штат Огайо

В ходе этого мысленного эксперимента каждый участник создает в своей голове слегка отличающийся алгоритм. Разумеется, все смотрят на маркеры, окружающие интересующую нас точку X, чтобы понять особенности района, но в какой-то момент необходимо решить, что ресторан находится слишком далеко, чтобы повлиять на прогноз. Иногда человек видит единственный ближайший ресторан, в данном случае — независимый (I), и основывает на этом свой прогноз: «Ближайшим соседом ресторана X является независимый ресторан (I), поэтому мой прогноз — (I)».

Однако большинство людей учитывают несколько соседних ресторанов. На втором изображении вокруг нового ресторана нарисована окружность, включающая семь его ближайших соседей. Вероятно, вы выбрали другое число, но мы выбрали 7. Шесть из семи ресторанов сетевые (C), поэтому мы прогнозируем, что новый ресторан тоже будет сетевым.

Что дальше?

Если вы поняли пример с рестораном, значит, вы уже на пути становления главным по данным. Давайте пройдемся по тому, что вы узнали.

— Вы выполнили классификацию, предсказав метку для нового ресторана (сетевой или независимый), обучив алгоритм на наборе данных (содержащем местоположения ресторанов и соответствующие метки).
— В этом состоит суть машинного обучения! Просто для разработки алгоритма вы использовали не компьютер, а собственную голову.
— Данный тип машинного обучения называется контролируемым обучением, потому что вы знали, что существующие рестораны были сетевыми (C) или независимыми (I). Эти метки направляли (то есть контролировали) ход ваших мыслей при размышлении о том, как расположение ресторана связано с его типом (сетевой или независимый).
— Если еще конкретнее, то вы использовали алгоритм контролируемой классификации под названием метод *k-ближайших соседей*[6]. Если K = 1, посмотрите на ближайший ресторан и получите свой прогноз. Если K = 7, посмотрите на 7 ближайших ресторанов и сделайте предсказание на основе их большинства. Это интуитивно понятный и мощный алгоритм. И в нем нет никакого волшебства.
— Вы также узнали о том, что для принятия обоснованных решений вам нужны данные. Однако помимо них вам необходимо кое-что еще. В конце концов, в этой книге много внимания уделяется критическому мышлению. Мы хотим показать не только то, как работают те или иные вещи, но и то, почему иногда они не срабатывают. Если бы мы попросили вас спрогнозировать, опираясь на приведенные в этом разделе изображения, будет ли новый ресторан ориентирован на детей, вы бы не смогли ответить. Для принятия обоснованных решений подходят далеко не любые данные. Для этого нужно достаточное количество точных и релевантных данных.
— Помните технические термины, которые мы упоминали ранее, говоря об «…анализе бинарной переменной отклика методом контролируемого обучения?..» Поздравляем, вы только что выполнили такой анализ. Переменная отклика — это просто еще одно название метки,

[6] Метод *k*-ближайших соседей можно использовать для предсказания не только классов, но и чисел. Эти так называемые задачи регрессии мы рассмотрим далее в книге.

и она является бинарной, потому что в нашем примере их было две — (C) и (I).

В этом разделе вы многое узнали, причем даже не осознавая этого.

ДЛЯ КОГО НАПИСАНА ЭТА КНИГА?

Как говорится в начале этой книги, данные затрагивают жизни многих сотрудников современных корпораций. Мы придумали нескольких аватаров, представляющих людей, которые могут выиграть от становления главными по данным.

Мишель — специалист по маркетингу, которая работает бок о бок с аналитиком данных. Она разрабатывает маркетинговые инициативы, а ее коллега собирает данные и измеряет влияние, оказываемое этими инициативами. Мишель считает, что их работа должна быть более инновационной, но не может донести до коллеги свои потребности в данных и их анализе. Общение между ними затруднено. Она поискала в Google некоторые специальные термины (машинное обучение и прогностическая аналитика), но в большинстве найденных ею статей использовались чрезмерно технические определения, неразборчивый компьютерный код, реклама аналитического программного обеспечения или консультационных услуг. В результате поисков она почувствовала еще большую тревогу и растерянность, чем раньше.

Даг имеет докторскую степень в области наук о жизни и работает в отделе исследований и разработок крупной корпорации. Скептик по натуре, он задается вопросом о том, не является ли шумиха вокруг данных очередным хайпом. Однако Даг старается не демонстрировать свой скептицизм на рабочем месте (особенно в присутствии нового директора, который носит футболку с надписью «Данные — это новая нефть»), поскольку не хочет, чтобы его считали дата-луддитом. В то же время он чувствует себя не у дел и решает узнать, из-за чего весь этот шум.

Реджина — топ-менеджер компании и хорошо осведомлена о последних тенденциях в области науки о данных. Она курирует новое подразделение своей компании, занимающееся наукой о данных, и регулярно взаимодействует со старшими дата-сайентистами. Реджина доверяет

своим специалистам, но ей хотелось бы иметь более глубокое понимание сути их деятельности, потому что ей часто приходится представлять и отстаивать результаты работы своей команды перед советом директоров компании. Реджине также поручена проверка нового технологического программного обеспечения. Она подозревает, что некоторые заявления поставщиков относительно «искусственного интеллекта» слишком хороши, чтобы быть правдой, и хочет получить дополнительные технические знания, чтобы отделить маркетинговые заявления от реальности.

Нельсон руководит работой трех дата-сайентистов в рамках своей новой должности. Будучи специалистом по компьютерным наукам, он знает, как писать программы и работать с данными, но плохо разбирается в статистике (поскольку прошел в колледже только один курс) и машинном обучении. Учитывая наличие технического образования, он хочет и может разобраться в деталях, но просто не находит на это времени. Руководство также побуждает его команду «больше заниматься машинным обучением», но на данный момент это кажется ей волшебным черным ящиком. Нельсон приступает к поиску материала, который поможет ему завоевать доверие команды и понять, какие проблемы можно решить с помощью машинного обучения, а какие — нет.

Мы надеемся, вы узнали себя в одном или нескольких из этих персонажей. Общим для них и, вероятно, для вас является желание стать лучшим «потребителем» данных и аналитики, с которыми вы сталкиваетесь.

Мы также создали аватар, представляющий людей, которым следует прочитать эту книгу, но которые, скорее всего, не станут этого делать (потому что в каждой истории должен быть злодей).

Джордж — менеджер среднего звена, читает последние деловые статьи об искусственном интеллекте и рассылает понравившиеся вверх и вниз по своей цепочке управления, как доказательство своей технической подкованности. Однако в зале заседаний он предпочитает «прислушиваться к своей интуиции». Джорджу нравится, когда его дата-сайентисты представляют ему цифры с помощью одного или двух слайдов. Когда результаты анализа согласуются с тем, что подсказала его интуиция, прежде чем он заказал исследование, он передает их вверх по цепочке и хвастается перед коллегами «внедрением искусственного интеллекта». Если результаты анализа не согласуются с его интуицией, он задает своим

дата-сайентистам ряд туманных вопросов и отправляет их на поиски «доказательств», необходимых для продвижения его проекта.

Не будьте такими, как Джордж. Если вы знаете «Джорджа», порекомендуйте ему эту книгу и скажите, что он похож на «Реджину».

ЗАЧЕМ МЫ НАПИСАЛИ ЭТУ КНИГУ

Мы считаем, что многие люди, похожие на описанные выше аватары, хотят больше узнать о данных, но не знают, с чего начать. Существует широкий спектр книг, посвященных науке о данных и статистике. На одном конце этого спектра находятся нетехнические книги, превозносящие достоинства и перспективы работы с данными. Какие-то из них лучше, чем другие. Самые лучшие из них напоминают современные бизнес-книги. Однако многие написаны журналистами, которые стремятся драматизировать начало эпохи данных.

В этих книгах описывается то, как те или иные бизнес-проблемы были решены путем их рассмотрения через призму данных. И в них даже встречаются такие понятия, как искусственный интеллект, машинное обучение и тому подобное. Не поймите нас неправильно, эти книги способствуют созданию осведомленности. Однако они не позволяют глубоко вникнуть в соответствующие темы, вместо этого сосредотачиваясь на высокоуровневом описании конкретной проблемы и ее решения.

На другом конце спектра находятся узкотехнические книги — 500-страничные тома в твердом переплете, пугающие как своим объемом, так и содержанием.

На противоположных сторонах этого спектра сосредоточены горы книг, что усугубляет разрыв в общении, — большинство людей читают либо только бизнес-книги, либо только технические книги, а не то и другое.

К счастью, между этими двумя крайностями есть много отличных книг. Нашими любимыми являются следующие:

— «Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking», Фостер Провост и Том Фосетт (Издательство: O'Reilly Media, 2013 год);
— «Много цифр. Анализ больших данных при помощи Excel», Джон Форман (Издательство: Альпина Паблишер, 2016 год).

Мы хотим добавить к этому списку еще одну книгу, которую вы сможете прочитать, не имея под рукой ни компьютера, ни блокнота. Если вам понравится наша книга, мы настоятельно рекомендуем прочитать одну или обе из указанных выше книг, чтобы углубить свое понимание. Вы не пожалеете.

Кроме того, мы очень любим эту тему. Если с помощью своей книги нам удастся побудить вас узнать больше о данных и аналитике, мы будем считать, что достигли успеха.

ЧТО ВЫ УЗНАЕТЕ

Эта книга поможет вам построить ментальную модель для понимания науки о данных, статистики и машинного обучения. Ментальная модель — это «упрощенное представление наиболее важных элементов некоторой предметной области, достаточное для решения проблем»[7]. Думайте о ней как о хранилище в вашем мозгу, в которое вы можете поместить информацию.

Некоторые книги и статьи начинаются со списка определений: «Машинное обучение — это…», «Глубокое обучение — это…» и так далее. Чтение технических определений в отсутствие ментальной модели, в которую эту информацию можно было бы вписать, похоже на скупку одежды, которую вам негде хранить. Рано или поздно вся она окажется на свалке.

Однако с помощью ментальной модели вы научитесь понимать, думать и говорить на языке данных. Вы станете главным по данным.

В частности, прочитав эту книгу, вы сможете:

— Думать статистически и понимать, какую роль вариации играют в вашей жизни и процессе принятия решений.
— Разбираться в данных — разумно говорить и задавать правильные вопросы о статистике и результатах, с которыми сталкиваетесь на рабочем месте.
— Осознавать истинное положение вещей в сфере машинного обучения, текстовой аналитики, глубокого обучения и искусственного интеллекта.
— Избегать распространенных ловушек при работе с данными и их интерпретации.

[7] Эта идея обсуждается в чрезвычайно полезной книге Г. Уилсона «Teaching tech together» (CRC Press, 2019).

КАК ОРГАНИЗОВАНА ЭТА КНИГА

Главный по данным — это тот, кто способен критически осмыслять данные вне зависимости от своей официальной роли. Это может быть аналитик, сидящий за компьютером, или топ-менеджер, наблюдающий за работой других. В этой книге вам как главному по данным предстоит сыграть разные роли.

Хотя «сюжет» книги выстроен в хронологическом порядке, каждая глава — это отдельный урок, который может быть изучен сам по себе. Однако мы рекомендуем прочитать книгу от начала до конца, чтобы выстроить свою ментальную модель и перейти от основ к более глубокому пониманию.

Книга состоит из четырех частей.

Часть I. Думайте как главный по данным. В этой части вы научитесь мыслить критически и задавать правильные вопросы о проектах по работе с данными, реализуемых в вашей организации; вы узнаете, что такое данные, а также освоите специальную терминологию и научитесь смотреть на мир через призму статистики.

Часть II. Говорите как главный по данным. Главные по данным — активные участники важных обсуждений. Эта часть научит вас «спорить» с данными и задавать правильные вопросы для понимания статистики, с которой вы сталкиваетесь. В ней вы познакомитесь с основными понятиями статистики и теории вероятностей, необходимыми для понимания и оспаривания предоставляемых вам результатов.

Часть III. Освойте набор инструментов дата-сайентиста. Главные по данным должны понимать фундаментальные концепции, лежащие в основе работы статистических моделей и моделей машинного обучения. В этой части вы получите интуитивное представление о неконтролируемом обучении, регрессии, классификации, текстовой аналитике и глубоком обучении.

Часть IV. Гарантируйте успех. Главные по данным знают о распространенных ошибках, допускаемых при работе с данными. В этой части вы узнаете о технических ловушках, которые приводят к провалу проектов, а также о людях и типах личностей, участвующих в соответствующих проектах. Наконец, мы дадим вам несколько рекомендаций о том, как добиться успеха в качестве главного по данным.

ПРЕЖДЕ ЧЕМ МЫ НАЧНЕМ

Мы не раз отмечали, что объем данных растет гораздо быстрее, чем наша способность формулировать порождаемые этим проблемы и возможности. Мы показали, что прошлое как всего общества, так и авторов этой книги наполнено неудачами, связанными с данными. И только поняв это прошлое, мы можем понять будущее. Для начала мы познакомили вас с несколькими важными концепциями в примере с классификацией ресторанов.

Для более глубокого понимания данных вам необходимо прорваться сквозь шум, критически осмыслить связанные с данными проблемы и научиться эффективно взаимодействовать с соответствующими специалистами. Мы уверены, что, вооружившись этими знаниями, вы добьетесь успеха.

Готовы? Ваш путь становления главным по данным начинается на следующей странице.

ЧАСТЬ I

Думайте как главный по данным

Многие компании спешат попробовать «что-нибудь новенькое», не останавливаясь для того, чтобы задать правильные бизнес-вопросы, изучить базовую терминологию или научиться смотреть на мир сквозь призму статистики.

У главных по данным не будет такой проблемы. Часть I, «Думайте как главный по данным», подготовит вас к предстоящему пути и поможет сформировать правильный настрой для размышлений о данных и их понимания. Эта часть состоит из следующих глав:

Глава 1. В чем суть проблемы?
Глава 2. Что такое данные?
Глава 3. Готовьтесь мыслить статистически.

ГЛАВА 1

В чем суть проблемы?

«Хорошо сформулированная проблема — это наполовину решенная проблема»
_ *Чарльз Кеттеринг, изобретатель и инженер*

Первый шаг на пути становления главным по данным заключается в том, чтобы помочь своей организации выбрать для решения те проблемы, которые действительно важны.

Это может показаться очевидным, однако многие из вас наверняка были свидетелями того, как компании говорили, насколько замечательные у них данные, а затем преувеличивали их влияние, неправильно интерпретировали результаты или инвестировали в технологии работы с данными, которые не создавали ценности для бизнеса. Часто кажется, что компании запускают проекты по работе с данными просто потому, что им нравится, как это звучит, не вполне понимая важность самих проектов.

Такой подход оборачивается напрасной тратой времени и денег и может породить негативное отношение к будущим проектам. Действительно, стремясь найти скрытую ценность в имеющихся данных, многие компании часто терпят неудачу на самом первом этапе процесса, связанном с определением стоящей перед бизнесом проблемы[8]. Итак, в этой главе нам предстоит вернуться к началу.

[8] Надежная стратегия работы с данными способна смягчить эти проблемы. Разумеется, важным компонентом любой подобной стратегии является решение значимых проблем, и именно на этом мы сосредоточим внимание в этой главе. Если вы хотите узнать больше о высокоуровневой стратегии работы с данными, обратитесь к книге *Jagare, U.* Data science strategy for dummies. (John Wiley & Sons, 2019).

В следующих разделах мы рассмотрим полезные вопросы, которые следует задать главному по данным, чтобы убедиться в важности его работы. Затем мы рассмотрим примеры того, как игнорирование этих вопросов оборачивается провалом проекта. Наконец, мы обсудим некоторые скрытые издержки, связанные с недостатком четкости в исходном определении проблемы.

ВОПРОСЫ, КОТОРЫЕ ДОЛЖЕН ЗАДАТЬ ГЛАВНЫЙ ПО ДАННЫМ

Мы по опыту знаем, что вернуться к основным принципам и задать фундаментальные вопросы гораздо сложнее, чем кажется на первый взгляд. Каждая компания имеет уникальную культуру, и командная динамика не всегда позволяет открыто задавать вопросы — особенно те, которые могут заставить других почувствовать свою несостоятельность. Многие главные по данным не могут даже начать задавать важные вопросы, способствующие реализации проектов. Вот почему иметь культуру, которая поощряет постановку таких вопросов, так же важно, как и сами вопросы.

Не существует универсальной формулы, подходящей для всех компаний и главных по данным. Если вы руководитель, мы призываем вас создать открытую среду, позволяющую задавать такие вопросы. (Начните с привлечения к обсуждению технических экспертов.) И задавайте вопросы сами. Это позволит вам продемонстрировать такую ключевую черту лидерства, как смирение, а также побудит других включиться в процесс. Если вы не руководитель, мы все равно рекомендуем вам задавать эти вопросы, не боясь нарушить статус-кво. Наш совет — просто делать все от себя зависящее. Исходя из опыта, мы считаем, что задавание правильных вопросов всегда позволяет получить гораздо больше, чем отказ от этого.

Мы хотим научить вас вовремя замечать предупреждающие знаки и сообщать о возникающих проблемах. Вот пять вопросов, которые вам следует задать, прежде чем приступать к работе с данными:

1. Почему эта проблема важна?
2. Кого затрагивает эта проблема?
3. Что, если у нас нет нужных данных?
4. Когда проект будет завершен?
5. Что, если нам не понравятся результаты?

Давайте подробно разберем каждый из них.

Почему эта проблема важна?

Несмотря на кажущуюся простоту, этот фундаментальный вопрос часто упускают из виду. Зачастую еще до начала реализации проекта мы сосредоточиваем внимание на способах решения проблемы и на потенциальных выгодах от этого. В конце главы мы поговорим об истинных последствиях оставления этот вопроса без ответа. Как минимум этот вопрос позволяет определиться с ожиданиями относительно результатов проекта. Это важно, поскольку проекты по работе с данными требуют затрат времени и сил, а зачастую и дополнительных инвестиций в технологии и данные. Простое определение важности проблемы до запуска проекта поможет повысить эффективность использования ресурсов компании.

Вы можете задать этот вопрос по-разному:

— Что мешает вам (нам) спокойно спать по ночам?
— Почему это важно?
— Эта проблема новая или она уже была решена ранее?
— Какой приз на кону? (Какова отдача от инвестиций?)

Вам нужно понять, как эту проблему видят другие. Это, в свою очередь, поможет понять, как разные люди будут поддерживать проект и согласятся ли они на его запуск.

Во время первоначальных обсуждений вам следует сосредоточиться на центральной бизнес-проблеме и пристально следить за разговорами о последних технологических тенденциях: они могут легко отвлечь участников от основной темы совещания. Обращайте особое внимание на два предупреждающих знака:

— Фокус на методологии. Это когда компании кажется, будто использование какого-то нового метода анализа данных или технологии даст ей некое преимущество. Вы наверняка сталкивались с маркетинговыми уловками наподобие: «Если вы не используете искусственный интеллект (ИИ), то вы отстаете…» Или когда компания привязывается к какому-то понравившемуся ей модному термину (вроде «анализа настроений»).
— Фокус на конечном результате. Некоторые проекты сбиваются с пути, потому что компании уделяют слишком много внимания тому, каким должен быть конечный результат. Например, они говорят

> о необходимости создания в рамках проекта интерактивной информационной панели. Вы приступаете к реализации проекта и оказываетесь перед выбором между созданием новой информационной панели и установкой системы бизнес-аналитики. Проектные группы должны быть готовы сделать шаг назад и понять, как именно то, что они собираются создать, принесет пользу организации.

То, что оба предупреждающих знака касаются технологии, а также то, что ее не следует упоминать на этапе определения проблемы, может показаться неожиданностью или облегчением. На более позднем этапе реализации проекта методологиям и результатам, безусловно, придется уделить внимание. Однако в самом начале проблема должна быть изложена в ясных и понятных каждому терминах. Вот почему мы рекомендуем вам отказаться от технической терминологии и маркетинговой риторики. Начните с описания проблемы, которую требуется решить, а не технологии, которую планируется использовать.

Почему это важно? Дело в том, что проектные команды обычно состоят из тех, кто обожает данные, и тех, кто их боится. Как только в ходе обсуждения проблемы разговор заходит о методах анализа или технологиях, могут произойти две вещи. Люди, которых пугают данные, перестают участвовать в определении бизнес-проблемы. А те, кто их обожает, быстро разбивают проблему на технические подзадачи, которые могут соответствовать или не соответствовать реальной бизнес-цели. После превращения бизнес-проблемы в набор подзадач, связанных с обработкой данных, на обнаружение допущенной ошибки могут уйти недели и даже месяцы, потому что после начала работы над проектом никто не захочет пересматривать формулировку основной проблемы.

По сути, команды должны ответить на вопрос: «Действительно ли это реальная бизнес-проблема, которую необходимо решить, или мы занимаемся анализом данных ради него самого?» Это хороший и прямолинейный вопрос, который следует задавать именно сейчас, когда вокруг науки о данных и смежных областей такой ажиотаж и путаница.

Кого затрагивает эта проблема?

В данном случае важно понять не только то, кого затрагивает проблема, но и то, как может измениться работа соответствующих специалистов в будущем.

Вы должны подумать обо всех уровнях организации (а также о ее клиентах, если таковые имеются). Мы не имеем в виду дата-сайентиста, работающего над проблемой, или команду инженеров, которым придется поддерживать программное обеспечение. Речь идет об установлении конечных пользователей. Зачастую это не только те люди, которые участвуют в определении проблемы. Поэтому очень важно понять, чья повседневная работа будет затронута в случае реализации проекта, и привлечь этих людей *к* его обсуждению.

Мы рекомендуем перечислить имена тех, чья работа изменится в случае решения поставленной проблемы. Если таких людей много, соберите небольшую группу из их представителей. Составьте список этих людей и поймите, как на них повлияет результат проекта — а затем свяжите полученные ответы с последним вопросом.

Вы можете выполнить пробный запуск решения в рамках мысленного эксперимента. Допустите возможность ответа на вопрос, а затем спросите свою команду:

— Можем ли мы использовать полученный ответ?
— Чья работа от этого изменится?

Разумеется, это предполагает, что у вас есть нужные данные для ответа на вопрос. (Как мы увидим в главе 4, это предположение может оказаться чрезмерно оптимистичным.) Тем не менее вы должны ответить на эти вопросы и рассмотреть несколько сценариев, предполагающих успешное решение проблемы. Во многих случаях ответы на эти вопросы позволяют либо усилить влияние предложенного проекта, либо установить тот факт, что его реализация не предвещает коммерческой выгоды.

Что, если у нас нет нужных данных?

Каждый набор данных содержит ограниченное количество информации, и на каком-то этапе уже никакая технология или метод анализа не помогут вам двигаться дальше.

Мы по опыту знаем, что некоторые из самых больших ошибок компании совершают тогда, когда не задаются этим вопросом. А ведь этих ошибок можно было бы избежать, если бы их учли до начала реализации проекта. Дело в том, что люди, которые до сих пор работали над проектом, хотят довести его до конца любой ценой. Главные по данным изначально допускают

вероятность отсутствия необходимых данных. На этот случай они предусматривают возможность сбора более качественных данных для ответа на вопрос. А если таких данных не существует, они возвращаются к исходному вопросу и пытаются пересмотреть содержание проекта.

Когда проект будет завершен?

Многим из нас доводилось участвовать в проектах, которые длились слишком долго. Если ожидания относительно длительности проекта неясны с самого начала, то со временем команды начинают посещать совещания просто по привычке и генерировать отчеты, которые никто не читает. Чтобы переломить подобные тенденции, следует ответить на вопрос: «Когда проект будет завершен?» еще до начала его реализации.

Этот вопрос позволяет сосредоточиться на причине, по которой проект был инициирован, и согласовать ожидания всех участников. Серьезные проблемы могут возникнуть в связи с тем, что в будущем вам может потребоваться некоторая информация или продукт, которых пока не существует. Определитесь с окончательным результатом. Это позволит возобновить процесс обсуждения потенциальной отдачи от инвестиций в реализацию проекта и убедиться в наличии у команды согласованной метрики для измерения его воздействия.

Итак, соберите всех участников проекта и определите причины, по которым он может быть завершен. Некоторые из них довольно очевидны, например, проект может быть свернут из-за отсутствия финансирования или снижения интереса. Отбросьте эти очевидные неудачи и сосредоточьтесь на том, что нужно сделать для решения бизнес-проблемы и успешного завершения проекта. В случае с проектами по работе с данными конечным результатом обычно является понимание (например, того, насколько эффективной была последняя маркетинговая кампания) или применение (например, прогностической модели, которая предсказывает объем поставок на неделю вперед). Многие проекты потребуют дополнительной работы, например, в виде дальнейшей поддержки и обслуживания, и об этом команде необходимо сообщить заранее.

Не думайте, что знаете ответ на этот вопрос, пока не задали его.

Что, если нам не понравятся результаты?

Ответ на последний вопрос готовит участников проекта к тому исходу, о котором они предпочли бы не думать, — к вероятности установления того,

что их изначальные предположения были ошибкой. Чтобы ответить на этот вопрос, вы должны представить, что находитесь в точке невозврата. После многих часов, потраченных на реализацию проекта, вы понимаете, что его результаты показывают совсем не то, на что вы рассчитывали. Обратите внимание, что речь идет не о том, что данные не позволяют ответить на поставленный вопрос. Данные отвечают на вопрос и, возможно, довольно уверенно — но полученный ответ не устраивает заинтересованные стороны.

Осознавать то, что результаты проекта оказались не такими, как вы ожидали, всегда нелегко. К сожалению, такое случается гораздо чаще, чем можно было бы предположить. Заранее обдумав возможность неудовлетворительного результата проекта, вы сможете разработать план действий на тот случай, если вам потребуется сообщить участникам плохие новости.

Задав этот вопрос, вы также сможете обнаружить различия в восприятии результатов проекта разными людьми. Например, вспомните нашего аватара Джорджа из введения. Джордж относится к тому типу людей, которые склонны игнорировать результаты, не отвечающие их убеждениям, и наоборот — отстаивать результаты, которые им соответствуют. Постановка данного вопроса позволяет выявить подобную предвзятость еще до начала реализации проекта.

Не стоит приступать к работе над проектом, если вы знаете, что у него есть только один приемлемый результат.

ПРИЧИНЫ ПРОВАЛА ПРОЕКТОВ ПО РАБОТЕ С ДАННЫМИ

Проекты могут провалиться по многим причинам, среди которых нехватка финансирования, ограниченные сроки, отсутствие необходимой экспертизы, необоснованные ожидания и тому подобное. А еще существуют проблемы, связанные с данными и методами их анализа. Например, проектная группа может применить метод анализа, который она не может объяснить, к данным, которые она не понимает, чтобы решить не имеющую значения проблему — и все равно считать полученный результат успехом.

Давайте рассмотрим такой сценарий.

Клиентское восприятие

Вы работаете в компании X, входящей в список Fortune 10, чья социально нечувствительная маркетинговая кампания вызвала негативную реакцию со стороны средств массовой информации. Вас назначили ответственным

за мониторинг «клиентского восприятия». В команду проекта входят следующие люди:

— менеджер проекта (вы);
— спонсор проекта (лицо, оплачивающее его реализацию);
— два специалиста по маркетингу (не имеющие опыта работы с данными);
— молодой дата-сайентист (только что окончивший колледж и стремящийся применить изученные методы на практике).

В ходе установочного совещания спонсор проекта и дата-сайентист бодро и взволнованно обсуждают то, что называется «анализом настроений». Спонсор проекта услышал об этом методе на недавней технической конференции, когда конкурирующая компания заявила о его использовании. Дата-сайентист сказал, что научился проводить анализ настроений в ходе работы над своим дипломным проектом, и предложил применить эту технику к комментариям клиентов на страницах компании в Twitter. Маркетологи думают, что эта техника позволяет интерпретировать эмоциональные реакции людей, используя данные из социальных сетей, но практически не участвуют в обсуждении.

Вам говорят, что анализ настроений позволяет автоматически пометить твит или пост как «положительный» или «отрицательный». Например, фраза «Спасибо за спонсорство Олимпиады» является положительным комментарием, а «Ужасное обслуживание клиентов» — отрицательным. Дата-сайентист сказал, что мог бы ежедневно подсчитывать количество положительных и отрицательных комментариев, строить графики, отражающие текущие тенденции, и публиковать результаты на информационной панели для всеобщего ознакомления. А самое главное, никому больше не нужно читать комментарии клиентов. Машина сделает это сама. Итак, решено. Проект запущен.

Месяц спустя дата-сайентист с гордостью демонстрирует информационную панель с результатами анализа клиентского восприятия, проведенного на базе клиентов компании X. Результаты обновляются каждый день с учетом последних данных и включают в себя ряд «положительных» комментариев, опубликованных за прошедшую неделю. На рис. 1.1 показан главный график этой информационной панели, отражающий динамику настроений клиентов. На нем показаны только доли положительных и отрицательных комментариев, хотя большинство их — нейтральные.

Спонсор проекта доволен. Неделю спустя эта информационная панель выводится на монитор в комнате отдыха на всеобщее обозрение.

Это успех.

Шесть месяцев спустя монитор убирается из комнаты отдыха в связи с ее ремонтом.

Никто этого не замечает.

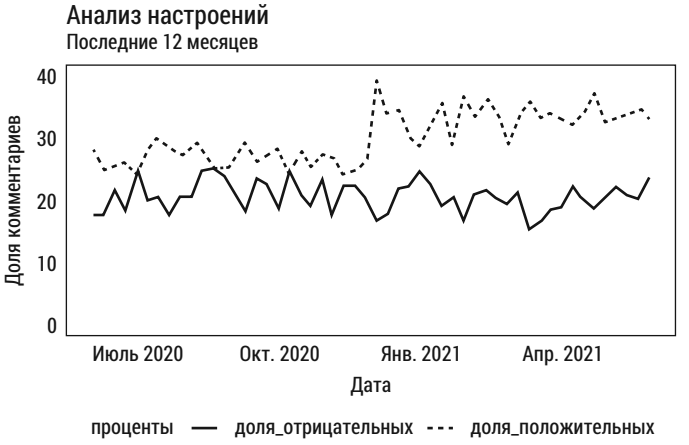


Рис. 1.1. Тенденции, выявленные в результате анализа настроений

Критический анализ проекта показал, что никто в компании не использовал полученные результаты анализа настроений — даже маркетологи, входящие в команду проекта. Когда их спросили, почему, они признались, что им не очень понравилась изначальная идея. Да, каждое сообщение можно было пометить как положительное или отрицательное. Но мысль о том, что никому больше не нужно читать комментарии, казалась принятием желаемого за действительное. Маркетологи сомневались в полезности такого метода маркировки. Кроме того, по их мнению, клиентское восприятие не может быть проанализировано на основе лишь онлайн-взаимодействия, хотя этот набор данных — самый доступный для проведения анализа настроений.

Обсуждение

На первый взгляд кажется, что в этом сценарии все прошло хорошо. Однако фундаментальный вопрос о важности проекта так и не был поднят. Вместо этого команда проекта попыталась ответить на другой вопрос: «Можем ли мы создать информационную панель для отслеживания настроения клиентов, исходя из содержания их комментариев на страницах компании?»

Разумеется, ответ на этот вопрос был положительным. Однако результаты этого проекта для организации оказались бесполезны и даже не важны.

Казалось бы, маркетологи могли бы сказать больше — однако они не были изначально включены в список людей, которых могут затронуть результаты проекта. Кроме того, в подходе команды к решению проблемы проявились два описанных ранее предупреждающих знака — фокус на методологии (анализ настроений) и фокус на конечном результате (информационная панель).

Более того, проектная группа из вышеописанного сценария могла бы выполнить пробный запуск своего решения, предположив, что у нее уже есть информационная панель, которая ежедневно обновляется с учетом результатов анализа комментариев в социальных сетях:

— Можем ли мы использовать полученный ответ? Команда могла бы подумать об актуальности анализа настроений с точки зрения изучения клиентского восприятия. Как команда может использовать полученную информацию? Какую выгоду может извлечь бизнес из знания настроений клиентов, основанного на анализе их комментариев в социальных сетях?
— Чья работа от этого изменится? Предположим, команда убедила себя в том, что знание настроений важно для эффективного управления бизнесом. Но будет ли кто-то следить за показателями на этой информационной панели? Будем ли мы принимать какие-либо меры, если вдруг наметится тенденция к снижению? А к повышению?

На этом этапе в разговор, возможно, вступили бы маркетологи. Но изменили бы они свою повседневную деятельность с учетом полученной информации? Скорее всего, нет. Проект в его нынешнем виде зашел в тупик.

Все было бы иначе, задай они те пять вопросов.

РАБОТА НАД ЗНАЧИМЫМИ ПРОБЛЕМАМИ

До сих пор мы объясняли провалы проектов неправильным определением основополагающей проблемы и связывали их с потерей денег, времени и энергии. Однако в мире науки о данных есть более серьезная и весьма неожиданная проблема.

В настоящее время отрасль сосредоточена на подготовке специалистов по работе с данными. Чтобы удовлетворить существующий спрос, всевозможные учебные учреждения выпускают множество критически мыслящих людей. А поскольку работа с данными заключается в нахождении истины, то главные по данным стремятся именно к этому.

Что же происходит, когда они вынуждены браться за проект, который их не вдохновляет, когда им приходится работать над плохо определенной проблемой или когда их навыки становятся всего лишь поводом для хвастовства руководителей?

В этих случаях многие специалисты по работе с данными разочаровываются в своей профессии. Работа над проблемами, которые чрезмерно сосредоточены на технологиях или имеют неоднозначные результаты, вызывает у них раздражение и неудовлетворенность. Сайт **Kaggle.com**, на котором дата-сайентисты со всего мира соревнуются друг с другом и изучают новые методы анализа данных, провел опрос относительно того, с какими препятствиями эти специалисты сталкиваются на работе[9]. Некоторые из перечисленных далее препятствий напрямую связаны с плохо сформулированными проблемами и неправильным планированием:

— Отсутствие четко поставленного вопроса, на который необходимо дать ответ (с этим столкнулись 30,4% респондентов).
— Результаты не используются лицами, принимающими решения (24,3%).
— Отсутствие вклада со стороны экспертов в предметной области (19,6%).
— Ожидания относительно воздействия проекта (15,8%).
— Интеграция результатов в решения (13,6%).

Все это имеет очевидные последствия. Те, кто не удовлетворен своей работой, уходят.

ПОДВЕДЕНИЕ ИТОГОВ

Цель данной книги — научить вас задавать больше вопросов. Этот процесс начинается с важнейшего, а иногда и сложнейшего вопроса: «В чем суть проблемы?»

---

[9] 2017 Kaggle Machine Learning & Data Science Survey. Результаты доступны по адресу: **www.kaggle.com/kaggle/kaggle-survey-2017**. Доступ получен 12 января 2021.

В этой главе вы узнали о том, как можно уточнить и прояснить центральный бизнес-вопрос, а также о том, почему проблемы, связанные с данными и их анализом, особенно важны. Мы перечислили пять важнейших вопросов, которые должен задать главный по данным при определении проблемы, а также предупреждающие знаки, говорящие о риске сбиться с пути. Если при обсуждении вопроса вы замечаете, что фокусируетесь (1) на методологии или (2) на конечных результатах, значит, пришло время взять паузу.

Ответив на все эти вопросы, вы можете приступать к работе.

ГЛАВА 2

Что такое данные?

«Если у нас есть данные, давайте смотреть на данные.
Если все, что у нас есть, — это мнения, давайте придерживаться моего»

_ Джим Барксдейл, бывший генеральный директор компании Netscape

Многие люди работают с данными, не владея соответствующим языком. Чтобы упростить понимание материала, изложенного в остальной части книги, в этой главе мы поговорим о данных и их типах. Если вы уже проходили базовый курс по статистике или аналитике, термины будут вам знакомы, однако некоторые фрагменты изложенного далее материала могут выходить за рамки вашего обучения.

ДАННЫЕ И ИНФОРМАЦИЯ

Термины «данные» и «информация» часто взаимозаменяемы. Однако в этой книге мы проводим между ними различие.

Информация — это извлеченное знание. Вы можете извлекать знания разными способами — например, путем измерения показателей процесса, размышлений о чем-то новом, изучения произведений искусства и обсуждения некоего предмета. Информация создается постоянно, и ее источниками является множество вещей, начиная с датчиков спутников и заканчивая нейронами в нашем мозге. Однако передать и зафиксировать эту информацию не всегда бывает легко. Некоторые вещи довольно просто измерить, а другие — нет. И все же мы стараемся передавать знания

другим и сохранять то, чему научились. Один из способов передачи и хранения информации — ее кодирование. В процессе кодирования мы создаем данные. Таким образом, данные представляют собой закодированную информацию.

Пример набора данных

Содержимое табл. 2.1 рассказывает историю компании, которая каждый месяц проводит различные маркетинговые мероприятия в Интернете, на телевидении или в печатных СМИ (газетах и журналах). Этот процесс каждый месяц генерирует новую информацию. Созданная компанией таблица представляет собой результат кодирования этой информации и, следовательно, содержит данные.

Таблица с данными, подобная табл. 2.1, называется набором данных.

Обратите внимание на то, что эта таблица содержит строки и столбцы, которые играют определенные роли в процессе интерпретации ее содержимого. Каждая горизонтальная строка таблицы представляет собой измеренный экземпляр связанной информации. В данном случае — информации о маркетинговой кампании. Каждый вертикальный столбец таблицы представляет собой список интересующих нас фрагментов информации, имеющих общую кодировку, что позволяет нам сравнивать экземпляры между собой.

Строки подобных таблиц обычно называются наблюдениями, записями, кортежами или испытаниями. Столбцы в наборах данных часто называются признаками, полями, атрибутами, предикторами или переменными.

Знайте свою аудиторию

Работа с данными ведется во множестве предметных областей, в каждой из которых используется профессиональный сленг, поэтому для одних и тех же вещей существует несколько названий. Одни специалисты по работе с данными могут называть столбцы в наборе данных «признаками», а другие — «переменными» или «предикторами». Поэтому главному по данным важно уметь ориентироваться в предпочтениях разных групп.

Точка данных — это место пересечения наблюдения и признака. В данном случае примером точки данных является *150* единиц товара, проданного 01 февраля 2021 года.

Табл. 2.1. Пример набора данных о рекламных расходах и прибыли

Дата	Рекламные расходы	Количество проданного товара	Прибыль	Медиа
2021–01–01	2000	100	10452	Печать
2021–02–01	1000	150	15349	Интернет
2021–03–01	3000	200	25095	Телевидение
2021–04–01	1000	175	12443	Интернет

Таблица 2.1 имеет заголовок (фрагмент нечисловых данных), который помогает нам понять, что означает каждый признак. Обратите внимание, что строка заголовка не обязательна. В таких случаях заголовок подразумевается, и человек, работающий с набором данных, должен знать, что означает каждый из признаков.

ТИПЫ ДАННЫХ

Существует множество способов кодирования информации, однако специалисты по работе с данными используют несколько видов кодировки для хранения информации и передачи полученных результатов. Два наиболее распространенных типа данных — числовые и категориальные.

Числовые данные в основном состоят из чисел, но могут включать дополнительные символы для обозначения единиц. К категориальным данным относятся слова, символы, фразы и (как ни странно) иногда числа — например, почтовые индексы. И числовые, и категориальные данные делятся на дополнительные подкатегории.

Существуют два основных типа числовых данных:

— Непрерывные данные могут принимать любое значение в некотором числовом диапазоне. Они представляют собой принципиально неисчисляемый набор значений. Возьмем, к примеру, погоду. Температура воздуха на улице, преобразованная в данные, будет представлять

собой непрерывную переменную. Допустим, она составляет 65,62 градуса по Фаренгейту (18,67 °C). Местная новостная станция может передать это значение как 65 °F (18 °C), 66 °F (19 °C) или 65,6 °F (18,7 °C).
— Счетные (или дискретные) данные, в отличие от непрерывных, ограничивают точность целым числом. Например, количество автомобилей, которыми вы владеете, может быть равно 0, 1, 2 и так далее, но не 1,23. Это отражает основополагающую реальность измеряемой вещи[10].

Категориальные данные также делятся на два основных типа:

— Упорядоченные (или порядковые) данные — это категориальные данные, которым присущ определенный порядок. Такие данные используют, например, организаторы опросов, когда предлагают вам оценить свой опыт по шкале от 1 до 10. Хотя эти данные напоминают счетные, мы не можем приравнять разницу между оценками 10 и 9 к разнице между 1 и 0. Разумеется, порядковые категориальные данные не обязательно кодировать в виде чисел. Например, размер рубашки относится к порядковым данным, но его можно закодировать с помощью слов: маленький, средний, большой, очень большой.
— Неупорядоченные (или номинальные) категориальные данные не имеют присущего им порядка. Например, табл. 2.1 содержит признак «Медиа» со значениями «Печать», «Интернет» и «Телевидение». Другие примеры номинальных переменных — ответы «Да» и «Нет», а также принадлежность к демократической или республиканской партии. Порядок их перечисления всегда является произвольным — нельзя сказать, что одна категория «важнее» другой.

В табл. 2.1 также есть признак «Дата», представляющий собой дополнительный тип данных, который является последовательным и может использоваться в арифметических выражениях в качестве числовых данных.

[10] Существуют дополнительные уровни непрерывных данных, называемые отношением и интервалом. Вы можете ознакомиться с ними самостоятельно, однако, согласно нашим наблюдениям, эти термины довольно редко используются в бизнес-среде. Кроме того, бывают ситуации, когда различие между непрерывными и счетными данными не имеет особого значения. Такие большие числа, как количества посещений веб-сайтов, часто считаются при анализе данных непрерывными, а не счетными. Это различие оказывается важным лишь тогда, когда речь идет о близких к нулю значениях. Мы поговорим об этом подробнее в следующих главах.

СБОР И СТРУКТУРИРОВАНИЕ ДАННЫХ

В предыдущем разделе мы говорили о типах данных в наборах, однако существуют более крупные категории для описания способа сбора и структурирования данных.

Данные наблюдений и экспериментальные данные

В зависимости от способа сбора данные могут называться экспериментальными или данными наблюдений.

— Данные наблюдений собираются в процессе пассивного наблюдения человека или компьютера за каким-либо процессом.
— Экспериментальные данные собираются в соответствии с научным методом с использованием предписанной методологии.

Большая часть данных в вашей компании и в мире вообще относится к данным наблюдений. Их примеры — число посещений веб-сайта, объем продаж на определенную дату и количество электронных писем, которые вы получаете каждый день. Иногда такие данные сохраняются с определенной целью, а иногда — просто так. Порой данные этого типа называют «обнаруженными»; очень часто они являются побочным продуктом продаж, платежей, сделанных с помощью кредитных карт, публикации сообщений в Twitter, лайков и тому подобного. То есть они находятся где-то в базе данных, ожидая, когда их обнаружат и используют с какой-то целью. Иногда данные наблюдений собираются потому, что их сбор ничего не стоит. Но иногда их собирают специально — например, с помощью опросов.

Экспериментальные данные собираются не пассивно, а намеренно и методично, чтобы ответить на конкретные вопросы. По этим причинам экспериментальные данные — золотой стандарт для статистиков и исследователей. Чтобы собрать экспериментальные данные, вы должны оказать воздействие на случайным образом выбранный объект. Распространенным примером в данном случае являются клинические испытания лекарств, в ходе которых пациентов случайным образом делят на две группы — группу активного воздействия и контрольную группу. При этом пациенты из первой группы получают настоящее лекарство, а пациенты из второй группы — плацебо. Случайное распределение пациентов позволяет сбалансировать

информацию, не представляющую важность для исследования (такую как возраст, социально-экономический статус, вес и так далее), чтобы две группы были максимально похожи во всех отношениях, за исключением факта применения лекарства. Это позволяет исследователям изолировать и измерить эффект препарата, не беспокоясь о потенциальном смешении признаков, способном исказить результат эксперимента[11].

Такой подход может применяться в разных сферах, начиная с клинических испытаний лекарств и заканчивая проведением маркетинговых кампаний. В сфере цифрового маркетинга веб-дизайнеры часто проводят над нами эксперименты, разрабатывая различные макеты веб-страниц или рекламные баннеры. Когда мы делаем покупки в Интернете, за кулисами происходит своеобразное подбрасывание монеты, от результатов которого зависит то, какой именно вариант из двух рекламных объявлений (назовем их А и Б) будет нам показан. После того как сайт посетят несколько тысяч ничего не подозревающих «морских свинок», веб-дизайнеры увидят, какой из вариантов обеспечил больше «кликов». А поскольку объявления А и Б показывались случайным образом, они могут определить, какое из объявлений более эффективно с точки зрения числа кликов, потому что все остальные потенциально смешивающиеся признаки (время суток, тип веб-пользователя и так далее) были сбалансированы путем рандомизации. Подобный метод часто называется «А/Б-тестированием» или «А/Б-экспериментом».

Подробнее о важности этого различия мы поговорим в главе 4 «Сомневайтесь в данных».

Структурированные и неструктурированные данные

Данные также могут быть структурированными и неструктурированными. Пример структурированных данных — содержимое таблиц, упорядоченное в виде строк и столбцов.

К неструктурированным данным относятся тексты обзоров на Amazon, изображения в социальных сетях, видео на YouTube, аудиофайлы и тому

[11] Пример таких искажающих результаты признаков можно найти в сфере клинических испытаний лекарств. Если группа активного воздействия состоит только из детей и никто из них не заболел, вам останется только гадать, чем это обусловлено — эффективным лекарством или особенностью детского организма. Эффект от использования препарата будет смешан с возрастом. Случайное распределение участников эксперимента на две группы позволяет этого избежать.

подобное. Преобразование неструктурированных данных в структурированные с целью дальнейшего анализа требует применения специальных методов (см. часть III данной книги).

Data — это один или много?

Настало время уточнить, какой позиции мы придерживаемся в споре, о котором вы, вероятно, даже не слышали.

На самом деле слово *data* (данные) в английском языке является множественным числом слова *datum*. (Как в случае со словами *criteria* (критерии) — *criterion* (критерий), *agenda* (повестка дня) — *agendum* (пункт повестки дня).)

Мы пытались придерживаться правил языка, говоря *the data are…* вместо *the data is…* но быстро поняли, что это не для нас. Нам кажется, что это звучит странно. И не только нам. Автор популярного блога **FiveThirtyEight.com**[12] предлагает использовать слово *data* в качестве неисчисляемого существительного, вроде *water* (вода) или *grass* (трава).

ОСНОВЫ СВОДНОЙ СТАТИСТИКИ

Данные не всегда выглядят как набор или электронная таблица. Часто они бывают представлены в виде сводной статистики. Сводная статистика позволяет получить информацию о наборе данных.

Три самых распространенных понятия сводной статистики — среднее значение, медиана и мода, с которыми вы, вероятно, уже хорошо знакомы. Тем не менее мы хотим потратить несколько минут на обсуждение этих понятий, поскольку часто замечаем, что в разговорной речи слова «нормальный», «обычный», «типичный» и «средний» используются в качестве синонимов для них. Чтобы избежать путаницы, давайте проясним, что же означают эти понятия.

— Среднее значение — это сумма всех имеющихся у вас чисел, деленная на их количество. Нахождение среднего значения дает вам

[12] "Data Is" vs. "Data Are": **fivethirtyeight.com/features/data-is-vs-data-are**

представление о том, какой вклад в общую сумму вносит каждое из наблюдений, когда все они имеют одно и то же значение.
— Медиана — это средняя точка диапазона значений, отсортированных по порядку.
— Мода — это число, которое встречается в наборе данных чаще всех остальных.

Среднее значение, медиана и мода называются мерами положения или мерами центральной тенденции. Меры вариации — дисперсия, размах и стандартное отклонение — являются мерами разброса. Номер положения указывает, где именно в числовом ряду находится типичное значение, а разброс говорит о том, насколько другие числа отклоняются от этого значения.

В качестве примера возьмем числа 7, 5, 4, 8, 4, 2, 9, 4 и 100. В данном случае среднее значение равно 15,89, медиана — 5, а мода — 4. Обратите внимание на то, что среднее значение 15,89 не присутствует среди исходных значений. Такое случается очень часто: среднее количество людей в домохозяйстве в США в 2018 году составляло 2,63 человека; звезда баскетбола Леброн Джеймс набирает в среднем 27,1 очка за игру.

Распространенная ошибка — использование среднего значения как средней точки данных, которой является медиана. Может показаться, что половина значений должна быть выше среднего, а половина — ниже. Но это не так. Чаще всего большинство значений находятся либо ниже, либо выше среднего. Например, у подавляющего большинства людей количество пальцев превышает среднее значение (которое составляет 9 с чем-то).

Чтобы избежать путаницы и недоразумений, мы рекомендуем использовать среднее значение, медиану и моду вместо таких понятий, как «обычный», «типичный» или «нормальный».

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы преподали вам основы языка, на котором вы можете говорить о данных на рабочем месте. В частности, мы обсудили:

— данные, наборы данных и различные названия строк и столбцов в них;
— числовые данные (непрерывные и дискретные);
— категориальные данные (порядковые и номинальные);
— экспериментальные данные и данные наблюдений;

— структурированные и неструктурированные данные;
— меры центральной тенденции.

Теперь, когда вы освоили терминологию, пора приступать к статистическому осмыслению имеющихся данных.

ГЛАВА 3

Готовьтесь мыслить статистически

«Статистическим называется особый стиль мышления, который сочетает в себе элементы детективной работы и скептицизма, а также предполагает использование альтернативных подходов к решению проблемы»[13]

_ *Фрэнк Харрелл, статистик и профессор*

Эта глава научит вас критически воспринимать и осмыслять данные, с которыми вы сталкиваетесь на рабочем месте и в повседневной жизни. Она закладывает основу для понимания остальной части книги, и если какое-либо из описанных далее понятий окажется для вас новым, то вскоре вы, вероятно, обнаружите, что смотрите новости или читаете научно-популярные статьи сквозь новый статистический объектив.

Прежде чем мы начнем, стоит сделать два важных замечания.

Во-первых, в этой главе мы коснемся лишь поверхности. Ее чтение не заменит семестр изучения статистики и не позволит разобраться во всех аспектах процесса «мышления», как это позволяет сделать уже ставшая классической книга «Думай медленно… решай быстро»[14]. Но мы все-таки введем несколько понятий, чтобы заложить основы для освоения статистического образа мышления, насколько это возможно.

Во-вторых, существует риск того, что при чтении следующих нескольких глав у вас сформируется довольно циничное отношение к данным. Вы

---

[13] Ф. Харрелл, профессор и заведующий кафедрой биостатистики Университета Вандербильта: **www.fharrell.com/post/introduction**

[14] «Думай медленно… решай быстро», Даниэль Канеман (Издательство: АСТ, 2014).

можете вскинуть руки и заявить, что вся эта статистическая чепуха скрывает правду под сложными уравнениями и цифрами и начать воспринимать в штыки любые результаты анализа, попадающиеся вам на глаза. А может быть, вы начнете бросаться помидорами в каждую прочитанную статью только потому, что вы узнали несколько статистических приемов и сомневаетесь в компетентности авторов.

Пожалуйста, воздержитесь от этого. Мы хотим, чтобы вы не отвергали предложенную вам информацию, а ставили ее под сомнение, вникали в ее смысл, осознавали имеющиеся ограничения — и, возможно, даже ее ценность.

ЗАДАВАЙТЕ ВОПРОСЫ

Основной принцип статистического мышления — «задавать вопросы».

Многие из нас делают это в повседневной жизни. Мы предполагаем, что вы как читатель книги о работе с данными не воспринимаете всерьез громкие заявления рекламодателей («Похудей на 5 килограммов за месяц!» или «Эти акции скоро будут стоить как акции Amazon!») и странные сообщения в социальных сетях. Итак, эта мышца у вас уже натренирована. Когда вы только наблюдаете со стороны, разбирать очевидную ложь может быть очень весело.

Однако все становится гораздо сложнее, когда заявления и данные касаются нас лично. Это демонстрируют любые политические выборы. Попытайтесь честно ответить себе на вопрос о том, насколько быстро утверждения или цифры, озвучиваемые представителями другой политической партии, начинают вызывать у вас подозрения[15]. Какие мысли приходят вам на ум? «У них плохие источники. Мои источники хорошие. Их информация ложная. Моя информация верна. Они просто не понимают, что происходит».

Совершенно очевидно, что эта дискуссия может очень быстро превратиться в философский спор. Мы не стремимся разжигать политические дебаты или углубляться в те факторы, которые определяют нашу личную и политическую идеологию. Мы лишь хотим подчеркнуть тот факт, что человеку трудно подвергать сомнению то, что затрагивает сам процесс его мышления и рассуждения.

[15] В США существуют две политические партии.

А теперь подумайте об информации, с которой вы сталкиваетесь на рабочем месте. Действительно ли вы способны скептически воспринимать содержимое электронных таблиц и презентаций PowerPoint, влияющее на успех вашей компании, результативность вашей работы и, возможно, даже на размер вашей премии? Наши наблюдения говорят о том, что зачастую это не так. В зале заседаний совета директоров цифры воспринимаются как неопровержимые факты, как истина, написанная черными чернилами и округленная до ближайшего десятичного знака.

Почему? Вероятно, это связано с тем, что у вас нет времени задавать вопросы или собирать дополнительную информацию. У вас есть ограниченное количество данных, на основе которых вы принимаете решения и на которые в случае необходимости можете списать неудачу. В условиях подобных ограничений скептицизм отключается почти рефлекторно. Еще одна причина может заключаться в том, что даже если вы понимаете связанные с данными проблемы, это не всегда можно сказать о вашем начальнике. Цепная реакция запускается тогда, когда все полагают, что остальные звенья управленческой цепочки принимают предоставляемые им цифры за чистую монету. И это предположение распространяется на всех, включая тех из нас, кто работает с электронной таблицей. Руководство не будет подвергать информацию сомнению, поэтому мы будем действовать так, будто она правдива.

Главные по данным смогут противостоять этой тенденции, если поймут суть вариации.

Комментарий по поводу «статистического мышления»

В понятие «статистическое мышление» мы вкладываем смысл из цитаты, приведенной в начале данной главы. Вы можете называть это вероятностным мышлением, статистической грамотностью или математическим мышлением. Вне зависимости от того, какую фразу вы предпочитаете, все эти понятия связаны с оценкой данных или доказательств.

Некоторые могут задаться вопросом о том, чем обусловлена важность этого стиля мышления. В конце концов, и бизнес, и жизнь в целом до сих пор обходились без него. Так почему сейчас? Почему это должно волновать главных по данным?

Ответ на эти вопросы можно найти в статье под названием «Data Science: What the Educated Citizen Needs to Know» («Наука о данных: что нужно знать образованному гражданину»), написанной гарвардским экономистом и врачом Аланом Гарбером:[16]

Преимущества использования науки о данных реальны и как никогда заметны и важны. Рост точности прогнозов сделает продукты этой науки более ценными и повысит интерес к ней. Однако ее успехи также могут порождать самоуспокоенность и заставлять нас закрывать глаза на ее недостатки. Специалисты будущего должны осознавать не только то, как наука о данных помогает им в работе, но и то, где и когда она оказывается бесполезной… Более глубокое освоение вероятностного мышления и оценки фактов — это тот навык, который пригодится всем.

ВО ВСЕМ ЕСТЬ ВАРИАЦИИ

Результаты наблюдений различаются между собой, и это вряд ли может кого-то удивить.

Цены на фондовом рынке колеблются ежедневно, результаты политических опросов меняются в зависимости от недели (и от того, кто именно проводит эти опросы), цены на бензин то растут, то снижаются, а ваше кровяное давление резко повышается, когда вы видите врача (при этом на медсестру вы так не реагируете). Даже ваши ежедневные поездки на работу, если разбить их на части и измерить с точностью до секунды, каждый день будут немного отличаться в зависимости от загруженности дорог, погоды, необходимости подвозить детей до школы или останавливаться, чтобы выпить кофе. Вариации есть во всем. Насколько вам комфортно от этой мысли?

Вероятно, вы давно приняли или, по крайней мере, смирились с вариациями в своей повседневной жизни, а, возможно, они вам даже нравятся. (Ну, за исключением колебаний фондового рынка.) Однако в целом мы понимаем, что некоторые вещи меняются по причинам, которые мы не всегда можем объяснить. Когда дело доходит до таких вещей, как накачка шин,

[16] Ссылка на статью в *Harvard Data Science Review*: **hdsr.mitpress.mit.edu/pub/pjl0jtkp**

заправка бензобака или оплата счетов за электричество, мы готовы мириться с постоянным изменением цифр при условии, что они имеют для нас интуитивно понятный смысл. Но, как было сказано в предыдущем разделе, нам гораздо сложнее относиться столь же беспристрастно к данным, затрагивающим нашу карьеру или бизнес.

Объем продаж компаний колеблется ежедневно, еженедельно, ежемесячно и ежегодно. Результаты опроса на тему удовлетворенности клиентов могут сильно различаться в разные дни. Если мы признаем реальность вариаций в нашей жизни, нам не нужно объяснять каждый пик и каждую впадину на графике. Однако именно к этому стремится любой бизнес. «Что делалось иначе в течение недели высоких продаж? — спрашивает руководство. — Давайте повторим все хорошее и устраним плохое». Вариации заставляют людей чувствовать себя беспомощными в отношении тех самых вещей, за знание которых им платят деньги и на которые они должны оказывать влияние.

Вероятно, когда дело касается бизнеса, вариации вызывают у нас гораздо больше дискомфорта, чем нам хотелось бы думать.

Существуют два типа вариаций. Один из них связан со способом сбора данных или проведения измерений и называется вариацией измерений. Второй тип связан со случайностью, лежащей в основе самого процесса, и называется случайной вариацией. На первый взгляд разница между ними может показаться незначительной, однако именно здесь проявляется важность статистического мышления. Принимаются ли решения в ответ на случайные вариации, которые невозможно контролировать? Или имеющаяся вариация отражает какой-то основополагающий процесс, который можно контролировать при условии его правильного выявления? Все мы надеемся на последнее.

Проще говоря, вариации порождают неопределенность.

Давайте рассмотрим один гипотетический сценарий и один исторический пример таких вариаций.

Сценарий: Клиентское восприятие (продолжение)[17]

Вы — менеджер розничного магазина, и ваше руководство внимательно отслеживает данные об удовлетворенности ваших клиентов, которые собираются, когда те звонят по номеру 1–800… указанному в нижней части

[17] Мы уделяем так много внимания клиентскому восприятию потому, что (1) его трудно измерить точно, (2) небольшая группа предвзятых людей оказывает сильное влияние на результаты и (3) руководство очень тщательно его анализирует.

квитанции. В ходе опроса клиентам предлагается оценить свою удовлетворенность по шкале от 1 до 10, где 10 означает «полностью удовлетворен». (Опрос включает ряд дополнительных вопросов, но первый — самый важный.)

При этом руководство устраивают только оценки 9 и 10. Оценка 8 для него равнозначна 0. Данные собираются еженедельно и отправляются лично вам и в корпоративный офис в файле PDF с красочными графиками, в котором слишком много страниц для представленной в нем информации. Тем не менее эти значения влияют на размер вашей премии и на размер премии вашего начальника, поэтому каждую неделю вы нервно и одержимо подсчитываете среднюю оценку удовлетворенности клиентов, надеясь, что вам удастся достичь показателя в 85%.

Здесь нам следует остановиться и поговорить об одном из источников вариаций — о способе измерения результатов опроса. Общеизвестно, что оценить что-либо по шкале от 1 до 10 весьма проблематично. Оценка 10, выставленная одним человеком («У них не было того, что я искал, но сотрудник помог мне найти замену!»), равнозначна оценке 5, выставленной другим («У них не было того, что я искал! Сотруднику пришлось помочь мне найти замену»). Мы проигнорируем другие потенциальные источники вариаций, такие как грубость сотрудника, переполненный магазин, экономический спад, заставляющий всех нервничать, то, что покупателю пришлось отправиться за покупками вместе с детьми и так далее.

Мы вовсе не предлагаем отказываться от таких опросов. Мы лишь хотим показать, что сам способ измерения данных является источником вариаций, часто упускаемых из виду. Из-за игнорирования вариаций может показаться, что отклонения от наших ожиданий отражают некачественное обслуживание, а не те различия, которые присущи самому вопросу. И все же компании продолжают гнаться за высокими целевыми показателями (в данном случае это оценки 9 и 10), не понимая, что главная причина вариации — выбранный способ измерения.

Вот как это может развернуться. Предположим, 50 человек оставляют отзывы каждый день на протяжении 52 недель. Это значит 350 опросов в неделю или 18 200 в год. Может показаться, что такое количество участников позволяет получить хорошее представление о клиентском восприятии. В конце каждой недели происходит подсчет результатов: руководство складывает все оценки 9 и 10, делит полученную сумму на общее количество опросов за неделю (350) и наносит результаты на график, показанный на рис. 3.1. Если

показатель превышает отметку 85%, вас одобрительно похлопывают по спине, а если нет, то вы покрываетесь холодным потом.

Каждый понедельник вы получаете отчет и звоните в компанию, чтобы обсудить результаты. Представьте, какой стресс вызывают эти разговоры на 5–9 неделях, когда результаты оказались чуть ниже порогового значения. На 10-й неделе вам наконец удается превысить пороговое значение (несомненно, благодаря мотивации со стороны вашего начальника), но наступает 11-я неделя, и вы достигаете нового минимума. И так происходит снова и снова.

Однако то, что вы видите на рис. 3.1 — чистая случайность. Мы сгенерировали 18 200 случайных чисел, которые были равны 8, 9 или 10, чтобы симулировать результаты опроса об удовлетворенности клиентов, и перетасовали их, как колоду карт[18]. Каждую «неделю» мы получали 350 оценок и рассчитывали на их основе значение метрики. Средний процент оценок 9 и 10 в наборе данных составил 85,3% (очень близко к истинному значению в 85%), что соответствовало корпоративному стандарту, но каждую неделю отклонялось от этого порогового значения просто из-за случайных вариаций.

Рис. 3.1. Результаты еженедельного опроса клиентов: процент положительных отзывов. Горизонтальная линия на уровне 85% соответствует целевому показателю

[18] В нашей симуляции вероятность получения оценки 8 составляла 15%, вероятность получения оценки 9–40%, а вероятность получения оценки 10–45%. Поскольку мы сами сгенерировали эти данные, мы точно знаем, что истинное значение показателя удовлетворенности клиентов, то есть вероятность получения оценки 9 или 10, составляет ровно 85%.

Из-за того, что никто не мыслил статистически, вы, ваш начальник и руководство компании старались добиться роста произвольного показателя, значение которого в принципе не зависело от чьих-либо действий.

Подобное стремление управлять метриками, не имея четкого статистического обоснования того, что они означают, мы называем иллюзией квантификации.

Сталкиваетесь ли вы с такой иллюзией на рабочем месте?

Анализ реальной ситуации: показатели заболеваемости раком почки

Самые высокие показатели заболеваемости раком почки в США, измеряемые как число случаев на 100 000 человек, наблюдаются в сельских округах, разбросанных по Среднему Западу, Южному и Западному регионам страны.

Остановитесь на мгновение и подумайте, чем это обусловлено.

Вы можете подумать, что жители сельской местности не имеют доступа к качественному медицинскому обслуживанию. Или, может быть, это результат нездорового образа жизни, диеты с высоким содержанием мяса, соли и жира или злоупотребления алкоголем. На самом деле строить предположения на основе фактов вполне естественно. Вы уже наверняка представляете, как исследователи начинают разрабатывать меры, необходимые для решения этой проблемы.

Однако есть еще один факт: самые низкие показатели заболеваемости раком почки в Соединенных Штатах также отмечаются в сельских округах, находящихся на Среднем Западе, а также в Южном и Западном регионах страны, которые часто соседствуют с округами с самыми высокими показателями заболеваемости[19].

Как такое может быть? Как в двух городах с похожей демографией могут наблюдаться столь разные результаты? Любая причина, которую вы могли бы предложить для объяснения высокого уровня заболеваемости раком почки в сельских округах, наверняка (в некоторой степени) применима и к соседним округам. Значит, дело в чем-то еще.

Возьмем два соседних сельских округа на Среднем Западе, округ А и округ Б, и предположим, что в каждом из них проживает всего 1000 жителей. Если в округе А отсутствуют случаи заболевания, то соответствующий уровень

[19] Представьте, что мы описали обратную ситуацию и сказали вам, что в сельской местности наблюдается самый низкий уровень заболеваемости раком почки. Какие причины вы бы назвали? Попробуйте поразмышлять о них, и вы увидите, как легко сочинить историю на основе имеющихся данных.

будет равен 0, а значит, этот округ будет относиться к категории с самым низким уровнем заболеваемости. Но если в округе Б есть хотя бы один случай заболевания раком почки, то соответствующий уровень там будет составлять 100 случаев на 100 000 жителей, что является самым высоким показателем в стране. Именно низкая численность населения в подобных округах обусловливает вариацию, которая одновременно приводит к самым высоким и самым низким показателям заболеваемости. И наоборот, один дополнительный случай заболевания в округе Нью-Йорк (в который входит Манхэттен) с населением более 1,5 миллиона человек вряд ли может повлиять на этот показатель. Увеличение количества случаев с 75 до 76 изменило бы число случаев на 100 000 человек с 5 на 5,07.

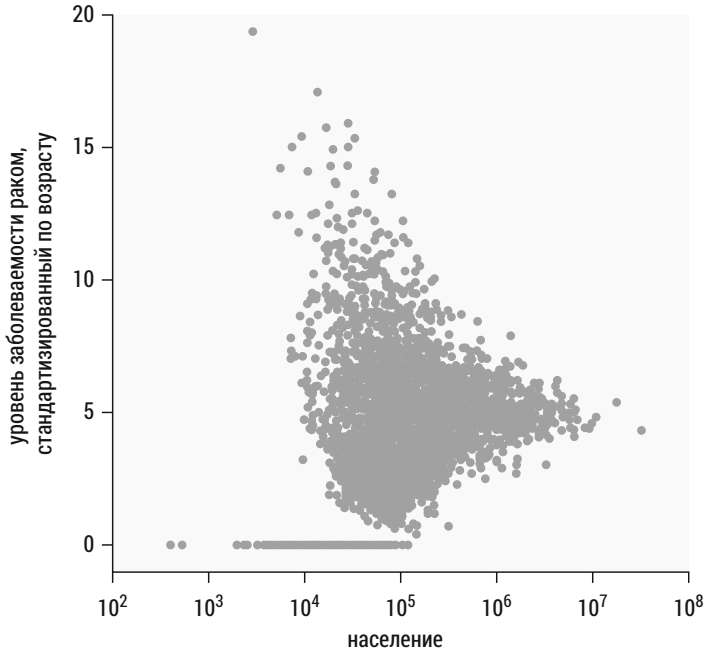


Рис. 3.2. График из статьи в журнале *American Scientist*

Все эти результаты реальны и взяты из статьи в журнале *American Scientist* под названием «The Most Dangerous Equation» («Самое опасное уравнение»)[20]. Результаты измерения уровня заболеваемости в разных округах

[20] Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95(3), 249.

США показаны на рис. 3.2. Малонаселенные округа в левой части графика демонстрируют гораздо более высокую вариацию уровня заболеваемости раком — от 0 до 20 (самый высокий показатель в стране). По мере движения слева направо с ростом численности населения вариация уменьшается, что придает графику треугольную форму. В правой части вариация совсем небольшая. Это значит, что в густонаселенных округах уровень заболеваемости стабилизируется у отметки 5 случаев на 100 000 человек и практически не меняется при выявлении дополнительных случаев.

В этой же статье приводятся и другие примеры того, как небольшие числа приводят к большой вариации. Например, были бы вы удивлены, узнав о том, что маленькие школы демонстрируют как лучшие, так и худшие результаты тестов? Один или два ученика, провалившие экзамен, могут очень сильно повлиять на общий процент. Экстремальные результаты часто обусловливаются именно небольшими числами.

ВЕРОЯТНОСТИ И СТАТИСТИКА

В нескольких предыдущих разделах мы говорили о вариации и о том, что она — источник неопределенности для многих бизнесов. Однако неопределенностью можно управлять, и именно здесь в игру вступают вероятность и статистика.

При описании математики, лежащей в основе результатов, мы часто используем термины «вероятность» и «статистика» как взаимозаменяемые. Давайте немного глубже разберемся в этих понятиях, чтобы по-настоящему осознать разницу между ними.

Представьте большой мешок со стеклянными шариками. Вы не знаете, какого они цвета. Вы не знаете ни их формы, ни размера. Вы даже не знаете, сколько их. Вы опускаете руку в мешок и вслепую берете горсть шариков.

Давайте остановимся на мгновение. У вас есть мешок, в который вы не заглядывали, и горсть стеклянных шариков в руке, которые вы никогда не видели. У вас нет никакой информации о том, что находится у вас в руке или в мешке.

И вот в чем разница. Теория вероятности позволяет вам угадать, что находится у вас в руке, если вам точно известно содержимое мешка. А статистика позволяет вам узнать о содержимом мешка на основании того, что оказалось у вас в руке.

Теория вероятности позволяет двигаться от общего к частному, а статистика — от частного к общему. Надеемся, так понятнее.

А теперь давайте рассмотрим два примера из реальной жизни.

— В основе работы казино Лас-Вегаса лежит вероятность. Каждый раз, когда вы играете в азартные игры, вы вытаскиваете из принадлежащего казино мешка шарики, которые являются либо выигрышами, либо проигрышами. Количество выигрышных шариков в этом мешке достаточно ровно для того, чтобы вы не утратили интерес к игре. Владельцы казино хорошо понимают суть вариации; более того, они ее коммерциализировали, оптимизировав выигрыши и проигрыши, чтобы поддерживать в вас определенный уровень интереса и возбуждения. Однако владельцы точно знают, что в долгосрочной перспективе казино окажется в выигрыше: именно они создали мешок, из которого игроки достают шарики, поэтому они точно знают, что внутри. Когда вы делаете ставку, кладете фишку на стол или дергаете за рычаг игрового автомата, казино точно знает вероятность вашего выигрыша. Если вы подумаете о том, каким количеством данных располагает казино, вы поймете, что они буквально живут в мире вариаций и при этом имеют четкое представление о возможных результатах.
— В основе политических опросов лежит статистика. В случае с казино содержимое мешка с шариками тщательно продумано, и из него постоянно делается выборка. Что касается выборов, то политики не знают, что на самом деле находится внутри всего мешка, вплоть до дня голосования, когда все шарики (то есть голоса) вытаскиваются наружу[21]. Только тогда политики могут узнать, что в мешке, и достаточно ли в нем выигрышных для них шариков. До выборов политики и политические партии имеют доступ лишь к небольшому набору случайных шариков (результатов опросов), и за этот доступ они платят огромные деньги. На основании анализа результатов опросов они делают выводы о закономерностях распределения шариков внутри мешка и соответственно корректируют свои предвыборные кампании. Поскольку их информация является неполной (и поскольку они часто допускают предвзятость и ошибки), они не всегда правильно ее понимают.

[21] Здесь мы немного упрощаем. Перед выборами политические партии пытаются повлиять как на количество шариков в мешке, так и на их цвет. Но даже это не позволяет им узнать все о содержимом мешка, поэтому им приходится полагаться на выборку.

> Но когда им это удается, полученный результат определяет разницу между их победой и поражением на выборах.

Некоторые важные концепции теории вероятности и статистики мы кратко рассмотрим в следующих разделах.

Вероятность и интуиция

Ранее в этой главе мы говорили о том, что случайная вариация не поддается контролю. Однако ее можно измерить, и теория вероятности дает нам для этого инструменты.

Иногда вероятности для нас вполне понятны. Если вы бросили честный кубик, то вы знаете, какова вероятность выпадения того или иного числа (1 из 6) или буквы (1 из 4). При игре в простые азартные игры вероятности кажутся нам интуитивно понятными. Однако это интуитивное понимание зачастую скрывает сложность, лежащую в основе этих вероятностей. Например, рекламные ролики часто апеллируют к простым вероятностям, сводя их к тому, что кажется нам интуитивно понятным.

Табл. 3.1. Вероятность того, что стоматологи согласятся с рекламным утверждением

	Стоматологи				
	1	2	3	4	5
Согласие	Да	Да	Да	Да	Нет
Вероятность	0,8	0,8	0,8	0,8	0,2

Вы наверняка видели рекламные ролики, в которых говорилось что-то вроде: «4 из 5 стоматологов согласны» с рекламным утверждением X (X может быть чем угодно, начиная с того, что жевательная резинка снижает риск развития кариеса, и заканчивая тем, что пищевая сода отбеливает зубы).

Теперь предположим, что перед вами сидят пять стоматологов. Если вы знаете, что 80% всех стоматологов согласны с утверждением X, насколько вероятно, что с ним согласны ровно четыре из пяти сидящих перед вами стоматологов?[22] 100%, 90% или 80%?

[22] Данный пример взят с сайта **www.johndcook.com/blog/2008/01/25/example-of-the-law-of-small-numbers**

На самом деле ответ равен 41%.

Интуитивно он может показаться слишком маленьким, но он правильный. Давайте разберемся, почему. Таблица 3.1 отражает одну из комбинаций ответов пяти стоматологов на вопрос о том, согласны ли они с утверждением X.

$$\textit{Вероятность такой комбинации} = 0{,}8 \times 0{,}8 \times 0{,}8 \times 0{,}8 \times 0{,}2 = 0{,}08192$$

Или, если кратко,

$$p = 0{,}8^4 \times 0{,}2 = 0{,}08192$$

Однако ответ «Нет», показанный в табл. 3.2, может быть дан пятью разными стоматологами, поэтому существуют пять комбинаций ответов.

Таким образом, мы должны умножить исходную вероятность на пять: $0{,}08192 \times 5 = 0{,}4096$, что примерно равно 41%.

Мы знаем, что с утверждением X соглашаются в среднем четверо из пяти стоматологов, но это не гарантирует того, что такой результат будет наблюдаться в каждой выборке, состоящей из пяти стоматологов. Вернемся к нашей аналогии с шариками. Если 80% шариков в мешке соответствует ответу «да», а 20% — ответу «нет», то иногда все пять шариков, оказавшихся у вас в руке, будут соответствовать положительному ответу, а в очень редких случаях — отрицательному. (Так проявляется вариация.)

Мы привели этот пример, чтобы еще раз подчеркнуть то, что люди часто недооценивают значение вариации, особенно когда имеют дело с небольшими числами. Их ожидания, основанные на интуиции, редко совпадают с реальными результатами расчета вероятностей. Недооценка вариации заставляет людей переоценивать свою уверенность в тех случаях, когда они имеют дело с небольшими значениями. Эта «склонность преувеличивать вероятность того, что малая выборка точно отражает свойства генеральной совокупности»[23] получила название «закона малых чисел».

Мыслить статистически, как и подобает главному по данным, значит помнить о том, что интуиция может сыграть с нами злую шутку. Мы рассмотрим еще несколько подобных примеров и заблуждений в следующих главах.

---

[23] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Табл. 3.2. Возможные комбинации из пяти стоматологов, среди которых четверо согласны с рекламным утверждением

	Стоматологи: согласны ли вы с утверждением X?				
Комбинация	1	2	3	4	5
1	Да	Да	Да	Да	**Нет**
2	Да	Да	Да	**Нет**	Да
3	Да	Да	**Нет**	Да	Да
4	Да	**Нет**	Да	Да	Да
5	**Нет**	Да	Да	Да	Да

Открытия с помощью статистики

Статистика часто делится на описательную и индуктивную. Скорее всего, вы уже знакомы с описательной статистикой, даже если не используете это выражение. Описательная статистика — это числа, обобщающие некие данные, значения, которые вы видите в газете или на проекционном экране в офисе. Средние объемы продаж за последний квартал, рост по сравнению с прошлым годом, уровень безработицы и так далее. Такие показатели, как среднее значение, медиана, размах, дисперсия и стандартное отклонение, относятся к описательной статистике, и для их расчета требуются специальные формулы, которые во множестве встречаются в соответствующих учебниках.

Описательная статистика предполагает преднамеренное упрощение данных и позволяет, например, свести всю электронную таблицу с данными о продажах компании в несколько ключевых показателей. В аналогии с шариками описательная статистика предполагает простое суммирование шариков, оказавшихся в вашей руке.

Несмотря на полезность этой операции, мы редко на ней останавливаемся. Мы хотим сделать дополнительный шаг и понять, как мы можем сделать предположение о содержимом мешка на основании информации о шариках, оказавшихся в нашей руке. В этом заключается суть индуктивной статистики, которая позволяет «перейти от мира к данным, а затем от данных обратно к миру»[24]. (Подробнее об этом мы поговорим в главе 7.)

[24] О'Нил Кэти, Шатт Рэйчел. «*Data Science.* Инсайдерская информация для новичков» (Издательство: Питер, 2019).

А пока давайте рассмотрим пример. Представьте, как бы вы отреагировали на заголовок «75% американцев верят в существование НЛО!», зная о том, что этот результат был получен в ходе опроса 20 посетителей Международного музея и исследовательского центра НЛО в Розуэлле, штат Нью-Мексико. Как вы думаете, можно ли на основе подобного исследования сделать вывод об истинном проценте американцев, верящих в НЛО?

Главный по данным отнесся бы к такому результату весьма скептически, поскольку в данном случае показатель 75% основан на:

— Предвзятой выборке. Люди, посещающие Розуэлл, с гораздо большей вероятностью верят в НЛО, чем среднестатистические жители США.
— Небольшой выборке. Вы уже знаете, какая значительная вариация может наблюдаться в выборке небольших размеров. Нет смысла делать выводы о том, что думают миллионы, на основе мнений 20 человек.
— Основополагающих допущениях. В заголовке говорится о том, что «американцы» верят в НЛО просто потому, что опрос был проведен в Америке. Однако данный музей — международная достопримечательность. Вы не можете быть уверены в том, что участники опроса были американцами.

Такие понятия, как предвзятость и размер выборки, — инструменты статистического вывода, помогающие нам понять, заслуживают ли доверия те статистические данные, которые мы видим или получаем в результате вычисления. Они — важная часть нашего инструментария. Основополагающие допущения также важно учитывать. Если вы хотите мыслить как главный по данным, не стоит принимать за чистую монету допущения, лежащие в основе высказанного вывода.

Сталкиваясь с какими-либо данными в своей работе, старайтесь не принимать предложенную информацию на веру и не прислушиваться к собственной интуиции.

Думайте статистически. Задавайте вопросы. Именно это делают главные по данным. В следующих главах вы найдете вопросы, которые помогут вам освоить статистический образ мышления.

Ресурсы для освоения статистического образа мышления

Ранее в этой главе мы сказали о том, что в ходе дальнейшего обсуждения статистического мышления мы собираемся лишь коснуться поверхности. К счастью, есть несколько отличных книг, в которых эта тема рассматривается более подробно. Больше всего нам нравятся следующие:

- «Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists», Joel Best (University of California Press, 2001);
- «Как не ошибаться. Сила математического мышления», Джордан Элленберг (Издательство: Манн, Иванов и Фербер, 2021);
- «Как лгать при помощи статистики», Дарелл Хафф (Издательство: Альпина Паблишер, 2015);
- «Голая статистика. Самая интересная книга о самой скучной науке», Чарльз Уилан (Издательство: Манн, Иванов и Фербер, 2022);
- «Proofiness: How You're Being Fooled by the Numbers», Charles Seife (Penguin Books, 1994);
- «(Не)совершенная случайность. Как случай управляет нашей жизнью», Леонард Млодинов (Издательство: Livebook, 2021);
- «Сигнал и Шум. Почему одни прогнозы сбываются, а другие — нет», Нейт Сильвер (Издательство: КоЛибри, 2016);
- «Думай медленно… решай быстро», Даниэль Канеман (Издательство: АСТ, 2014).

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы заложили основы для освоения статистического образа мышления, от которых будем отталкиваться в следующих главах книги.

В частности, мы поговорили о важности вариаций и понимания их существования в контексте измеряемых нами вещей. Мы показали, что результаты опросов клиентов могут иметь широкий разброс не потому, что

обслуживание было плохим (хотя и это возможно), а потому, что сам вопрос предрасполагает к даче совершенно разных ответов, которые до измерения могут характеризоваться как похожие.

Мы также поговорили о вероятности и статистике, которые помогают нам управлять вариациями, демонстрируя то, что некоторые из этих вариаций являются предсказуемыми, а некоторые не имеют значения в долгосрочной перспективе.

Теория вероятности позволяет нам двигаться от общего к частному, то есть делать выводы о небольшом фрагменте данных на основе знаний о совокупности информации. А статистика позволяет нам двигаться от частного к общему, то есть делать выводы о совокупности информации на основе доступных нам фрагментов. И теория вероятности, и статистика — инструменты, которые помогают нам узнать больше о полной картине, пока она остается для нас неясной. Наконец, мы поговорили об использовании знаний о теории вероятности и статистике для оттачивания навыка критического мышления.

ЧАСТЬ II

Говорите как главный по данным

Часть II, «Говорите как главный по данным», так же, как и первая, побуждает вас мыслить статистически и подвергать все сомнению. В ней вы найдете вопросы, которые следует задать, и вещи, которые следует обдумать независимо от того, о чьем проекте по работе с данными идет речь — о вашем или о чужом. Многие из этих вопросов отражены в названиях будущих разделов. Считайте это своеобразной подсказкой. Данная часть книги состоит из следующих глав:

Глава 4. Сомневайтесь в данных.
Глава 5. Исследуйте данные.
Глава 6. Изучайте вероятности.
Глава 7. Бросайте вызов статистике.

Прочитав эти главы, вы научитесь задавать правильные вопросы относительно данных и аналитики, с которыми будете сталкиваться на работе.

ГЛАВА 4

Сомневайтесь в данных

«Для извлечения разумного ответа из имеющейся совокупности данных одного страстного желания недостаточно»

_ *Джон Тьюки, известный статистик*

Как главный по данным, именно вы должны подвергать сомнению данные, используемые в рамках того или иного проекта.

Мы говорим о необработанных данных — исходном материале, на основе которого рассчитываются все статистические показатели, строятся модели машинного обучения и создаются визуализации, отображаемые на информационных панелях. Это данные, которые хранятся в ваших электронных таблицах или базах данных. Если эти необработанные данные плохие, то никакие методы очистки, статистической обработки или машинного обучения не помогут это скрыть. В качестве резюме для данной главы лучше всего подходит фраза, которую вы, вероятно, уже слышали: «Мусор на входе, мусор на выходе». В этой главе мы перечислим те типы вопросов, которые вам следует задать, чтобы оценить качество имеющихся у вас данных.

Мы выделили три основных и несколько уточняющих вопросов, которые помогут вам поспорить с имеющимися данными.

— Какова история происхождения этих данных?
  - Кто собирал данные?
  - Как собирались эти данные?

— Являются ли данные репрезентативными?
  - Имеет ли место предвзятость выборки?
  - Что вы сделали с выбросами?
— Какие данные я не вижу?
  - Как вы поступили с отсутствующими значениями?
  - Позволяют ли данные измерить то, что вас интересует?

В следующих разделах мы подробно рассмотрим каждый вопрос, поговорим о причинах, по которым его следует задавать, и о том, какие проблемы он обычно позволяет обнаруживать.

Однако прежде, чем это сделать, мы предлагаем вам выполнить одно мысленное упражнение.

ЧТО БЫ ВЫ СДЕЛАЛИ?

Вы отвечаете за крупный проект в технологической компании, которая находится на пороге прорыва в области создания беспилотных автомобилей. Это важный момент для вас и вашей работы, не говоря уже о карьере. Успешная демонстрация вашего продукта обещает искупить все сверхурочные часы работы, чрезмерно оптимистичные обещания, данные руководству, задержки в реализации проекта и бюджетные затраты на исследования и разработки.

И сейчас вечер накануне презентации прототипа нового автомобиля.

Руководители компании, десятки сотрудников, потенциальные инвесторы и представители СМИ проехали сотни километров, чтобы засвидетельствовать то, что может стать переломным моментом в истории автомобилестроения. Однако поздно вечером ваш старший инженер сообщает, что на завтра синоптики прогнозируют 31 °F (–1 °C). По словам инженера, низкие температуры могут поставить под угрозу жизненно важные компоненты инновационной системы автономного вождения прототипа автомобиля. Дело не в том, что он уверен в непременном возникновении проблемы. Просто система, которую в будущем планируется адаптировать и испытать при отрицательных температурах, еще не была опробована на морозе, так что демонстрация рискует превратиться в публичную и дорогостоящую катастрофу.

Однако перенести подобное мероприятие дорого и непросто. Если презентация не состоится завтра, то идеальных условий придется ждать месяцами. Ваша компания потратила большую часть предыдущего года на создание

ажиотажа вокруг этого момента. Если презентацию перенести, уровень заинтересованности уже не будет таким высоким.

Вы просите инженера предоставить данные, заставляющие его беспокоиться о возможном повреждении внутренних компонентов автомобиля из-за низких температур. Он показывает вам график, представленный на рис. 4.1.

По словам инженера, компания провела 23 тест-драйва при различных температурах, и в ходе семи из них (отмеченных на графике) имел место выход из строя критической части системы самонавигации. В ходе двух тест-драйвов из строя вышли сразу два критических компонента.

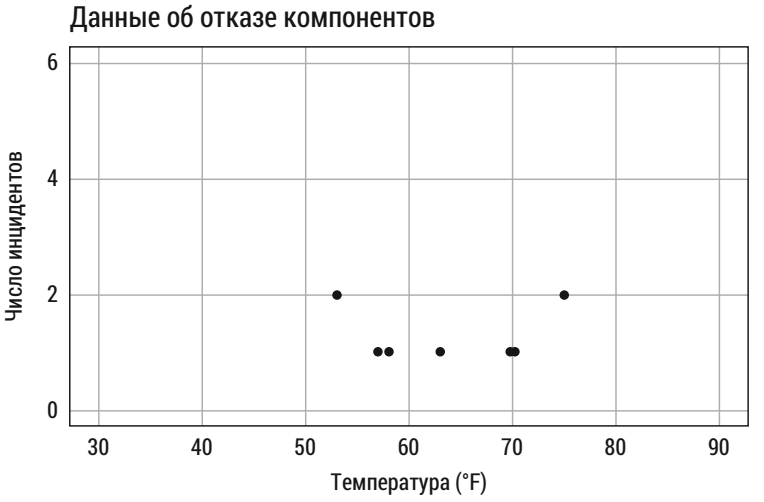


Рис. 4.1. График зависимости числа отказов критических компонентов от температуры во время тест-драйвов

Ваши инженеры учли вероятность подобных отказов, поэтому они обеспечили избыточность. Каждая система предусматривает шесть критических компонентов (вот почему максимальное значение на вертикальной оси — 6). Наличие запасных частей означает, что даже в случае поломки некоторых из них машина продолжит функционировать. В ходе 23 тест-драйвов из строя ни разу не вышло сразу более двух компонентов, поэтому и проблем с использованием автомобиля ни разу не возникло. В обоих случаях, имевших место при температуре 53 °F (12 °C) и 75 °F (24 °C), машина так и не остановилась. Минимальная температура, при которой проводилось испытание, составляла 53 °F (12 °C), а максимальная — 81 °F (27 °C).

«Однако мы не тестировали систему при более низких температурах», — говорят инженеры. И вы понимаете, что они обеспокоены.

Но как бы вы ни старались, вы не можете заметить связь между температурой и вероятностью отказа компонентов за исключением того, что все они имели место при температурах значительно выше 30 °F (–1 °C). Вам трудно представить сценарий, при котором низкие температуры могут вывести из строя более двух компонентов из шести, учитывая данные, полученные в ходе 23 тест-драйвов. Кроме того, машина вполне может продолжать движение и при наличии четырех исправных критических компонентов. Если во время демонстрации выйдет из строя максимум два, узнает ли об этом кто-нибудь вообще?

Что бы вы сделали? Отложили бы презентацию или провели ее в запланированный день?

Остановитесь на мгновение и подумайте о том, есть ли какие-нибудь недостающие данные, которые вы захотели бы учесть.

Катастрофа, вызванная недостатком данных

28 января 1986 года на глазах у всего мира НАСА запустило космический шаттл «Челленджер» из Космического центра им. Кеннеди во Флориде при отрицательных температурах.

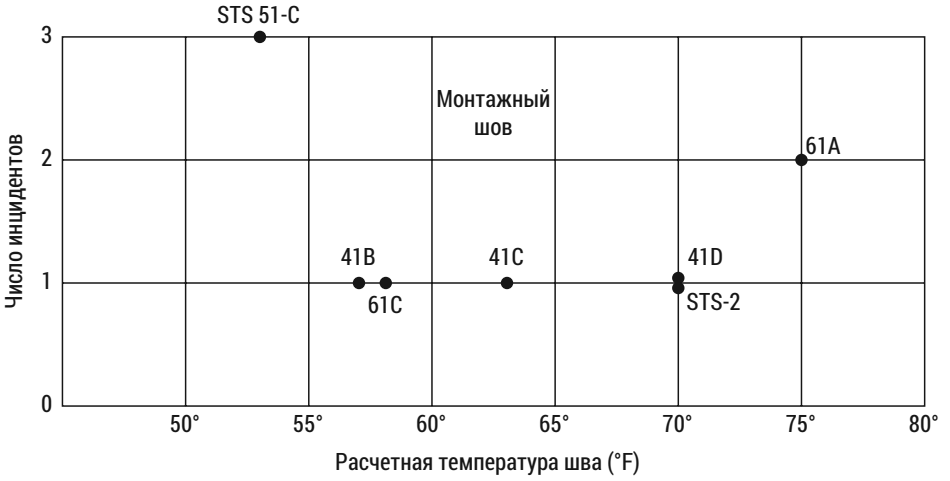


Рис. 4.2. График зависимости числа неисправностей уплотнительных колец от температуры во время полетов. График взят из отчета Президентской комиссии, занимавшейся расследованием катастрофы космического челнока «Челленджер»

Многие из нас знают эту часть истории «Челленджера», однако мало кто знаком со стоящими за ней данными. Дело в том, что у «Челленджера» тоже было шесть критически важных компонентов, известных как уплотнительные кольца, которые «предотвращают утечку горящего ракетного топлива из соединений ускорителя»[25]. До запуска в ходе 23 испытаний имели место семь инцидентов с этими уплотнительными кольцами.

Знакомый сценарий?

Вечером накануне запуска НАСА оказалось перед тем же трудным выбором, что и вы в ходе выполнения своего мысленного упражнения. Согласно отчету комиссии Роджерса (который был заказан президентом Рональдом Рейганом после аварии «Челленджера»), в ночь перед запуском состоялось совещание по этому вопросу.

> Менеджеры сравнили только те полеты, в ходе которых наблюдались тепловые повреждения уплотнительных колец, вместо того, чтобы проанализировать частоту возникновения этой неисправности с учетом всех полетов (рис. 4.2)[26].

«При таком сравнении, — говорилось в отчете, — в распределении «повреждений» уплотнительных колец в диапазоне температур швов между 53 и 75 градусами по Фаренгейту, фиксируемых при запуске, нет ничего необычного».

Проанализировав эти неисправности, НАСА осуществило запуск. Но из-за необычно холодных условий уплотнительные кольца не сработали должным образом, и на 73-й секунде полета шаттл развалился на части. Погибли все семь астронавтов на борту.

Как вы думаете, какие данные упустили специалисты космического агентства?

Как насчет тех 16 испытательных запусков, в ходе которых не возникло никаких неисправностей, отмеченных на рис. 4.3 и задокументированных комиссией Роджерса?

---

[25] Цитата из статьи NRP. “Challenger engineer who warned of shuttle disaster dies.” **www.npr.org/sections/thetwo-way/2016/03/21/470870426/challenger-engineer-who-warned-of-shuttle-disaster-dies**

[26] Цитата из отчета Президентской комиссии, занимавшейся расследованием катастрофы космического челнока «Челленджер» (Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident). С. 146. **sma.nasa.gov/SignificantIncidents/assets/rogers_commission_report.pdf**

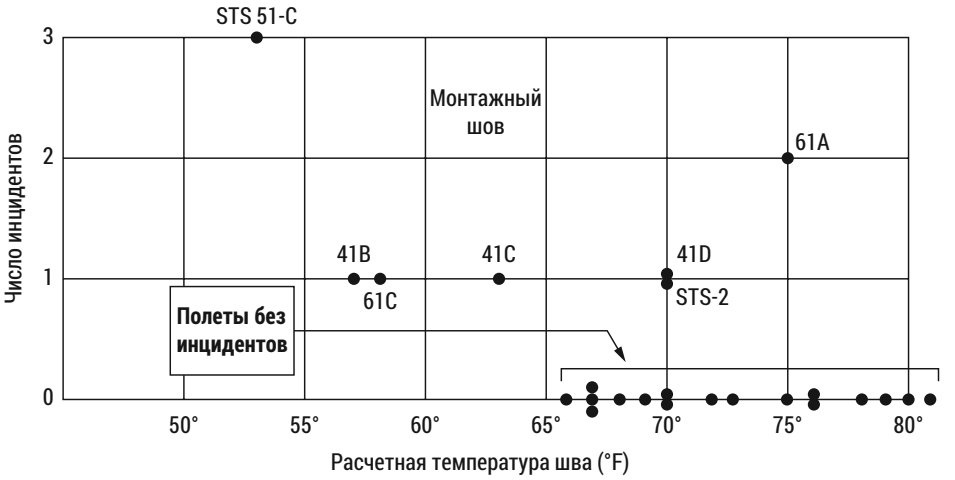


Рис. 4.3. График зависимости числа неисправностей уплотнительных колец от температуры во время полетов, включая испытательные запуски без инцидентов. График взят из отчета Президентской комиссии, занимавшейся расследованием катастрофы космического челнока «Челленджер»

Примечание

В главе 2 «Что такое данные?» мы говорили о том, как тип данных диктует выбор метода анализа. Это как раз один из таких случаев. Количество инцидентов — это числовые счетные данные, которые требуют применения специального типа моделирования, называемого биномиальной регрессией. Поскольку речь идет о счетных, а не о непрерывных данных, вы не можете использовать линейную регрессию, о которой мы поговорим в главе 9. Описание биномиальной регрессии выходит за рамки этой книги, но тип данных, о которых идет речь, диктует использование именно этого метода анализа. Если бы вы использовали линейную регрессию, чтобы провести прямую линию через точки данных, вы бы предсказали отрицательные значения количества отказов для высоких температур, что не имеет никакого смысла.

Вернемся к мысленному упражнению. Запросили бы вы какие-нибудь недостающие данные? Если бы вы это сделали, а возможно, и привлекли бы к анализу статистиков, вы могли бы заметить тенденцию, предупреждающую о возможном отказе компонентов при более низких температурах. На рис. 4.4

показаны испытания нашего гипотетического беспилотного автомобиля, в том числе те, в ходе которых критические компоненты не выходили из строя.

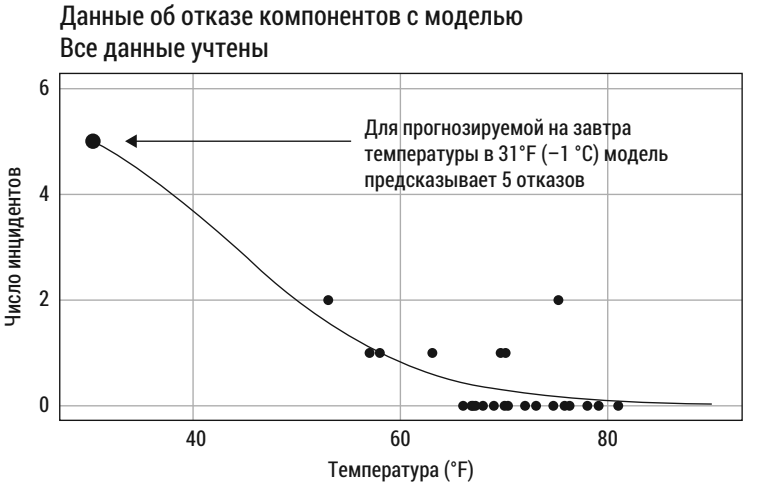


Рис. 4.4. График зависимости числа отказов критических компонентов от температуры во время тест-драйвов. Линия представляет собой модель биномиальной регрессии

В последующие десятилетия статистики, инженеры и исследователи тщательно изучали данные[27], связанные с катастрофой «Челленджера». С помощью этого реального сценария мы хотели продемонстрировать вам те вопросы, с которыми приходится сталкиваться специалистам по работе с данными. В статье, опубликованной в престижном журнале *Journal of the American Statistical Association (JASA)*, издаваемом Американской статистической ассоциацией, был представлен анализ, который мы воссоздали на рис. 4.4. Он говорит о том, что при отрицательных температурах пять из шести основных уплотнительных колец могут выйти из строя. При составлении этого графика использовались данные, которые не были учтены накануне запуска шаттла. В статье говорится о том, что «статистическая наука могла внести ценный вклад в процесс принятия решения о запуске»[28].

Хотели бы вы увидеть такой же график накануне важной презентации?

---

[27] Данные доступны для загрузки из репозитория для машинного обучения Калифорнийского университета в Ирвайне: **archive.ics.uci.edu/ml/datasets/Challenger+USA+Space+Shuttle+O-Ring**

[28] Dalal, S. R., Fowlkes, E. B., & Hoadley, B. (1989). Risk analysis of the space shuttle: pre-Challenger prediction of failure. *Journal of the American Statistical Association, 84*(408), 945–957.

Комментарий Алекса по поводу данных о состоянии «Челленджера»

Внимательные читатели, вероятно, заметили небольшое расхождение между данными, представленными на рис. 4.1, и графиками из отчета комиссии Роджерса на рис. 4.2 и 4.3. На рис. 4.1 температуре 53 °F (12 °C) соответствуют два инцидента, а на рис. 4.2 и 4.3 — три. (Все остальные точки данных совпадают.) Дело в том, что конструкция космического челнока предусматривала шесть основных и шесть второстепенных уплотнительных колец. Третий инцидент при температуре 53 °F (12 °C), отмеченный на рис. 4.2 и 4.3, произошел со второстепенным уплотнительным кольцом и был единственным случаем подобного повреждения, имевшим место в ходе 23 полетов, предшествовавших катастрофе. Приведенный здесь анализ сосредоточен на шести основных уплотнительных кольцах, как и анализ, приведенный в статье в журнале *JASA*.

История «Челленджера» демонстрирует довольно распространенное и пугающее явление. Мы часто сосредоточиваемся на данных, которые, как нам кажется, кодируют нужную нам информацию, отбрасывая при этом те данные, которые мы считаем несущественными. Мы признаем, что далеко не во всех ситуациях последствия могут быть столь же ужасными, как в случае с «Челленджером», когда на карту было поставлено так много.

Мы не утверждаем, что анализ полного набора данных позволил бы принять правильное решение. Никто не может знать это наверняка. Другие факторы тоже, безусловно, сыграли свою роль. Мы просто хотим сказать, что спор с данными часто помогает сделать дополнительные открытия.

И в этом смысле история, рассказанная данными о состоянии «Челленджера», вполне ясна. Однако большинство компаний не спорят со своими данными, развивая вместо этого культуру принятия. Результат этого — систематические провалы проектов по работе с данными, обусловленные неготовностью задавать важные вопросы.

Итак, цель этой главы — научить вас спорить с данными и задавать правильные вопросы.

РАССКАЖИТЕ МНЕ ИСТОРИЮ ПРОИСХОЖДЕНИЯ ДАННЫХ

Все данные берутся из какого-то источника, который нам не следует игнорировать. Итак, мы предлагаем вам спросить: «Каково происхождение этих данных?»

Этот вопрос нравится нам тем, что он является открытым и позволяет быстро оценить согласованность сырых данных с заданным относительно них вопросом. Кроме того, для ответа на него не требуются ни математические, ни статистические знания. Еще важнее то, что сам вопрос создает ощущение открытости и укрепляет доверие к последующим результатам (или заставляет сомневаться в них).

Внимательно проанализируйте ответ на предмет возможных проблем с корректностью и целостностью данных, обусловленных особенностями создавшего их лица или организации.

В частности, постарайтесь получить ответы на следующие вопросы:

— Кто собирал данные?
— Как собирались эти данные? Это данные наблюдений или экспериментальные данные?

Кто собирал данные?

Задавая этот вопрос, мы пытаемся, во-первых, установить, откуда именно были получены данные, а во-вторых, выявить возможные проблемы, связанные с их происхождением, чтобы при необходимости задать дополнительные вопросы.

Многие крупные компании считают, что все их данные берутся из внутреннего источника. Например, компания, использующая данные о рабочей силе (то есть данные, основанные на результатах опросов сотрудников и другой соответствующей информации), на самом деле может использовать данные, собранные третьей стороной и принадлежащие ей. Потребление этих данных может происходить через портал компании. Это может создать иллюзию того, что данные были собраны компанией и принадлежат ей, даже если это не так.

Мы хотим, чтобы вы точно определили того, кто собирал данные. Как главный по данным, вы должны убедиться в том, что полученные извне данные надежны и имеют отношение к поставленной бизнес-задаче. Большую часть данных, полученных из сторонних источников, довольно

трудно использовать в том формате, в котором они предоставляются. Вам или кому-то из вашей команды придется преобразовать данные, полученные от третьей стороны, в нужный формат и придать им необходимую структуру, чтобы привести их в соответствие с уникальными информационными активами вашей компании.

Как собирались эти данные?

Вам также необходимо выяснить, как собирались данные. Этот вопрос поможет вам выявить возможные недопустимые выводы, сделанные об этих данных, а также этические проблемы, связанные с процессом их сбора.

Напомним, что существуют два основных метода сбора данных — наблюдение и эксперимент.

Наблюдение — это пассивный способ сбора данных. Примерами данных наблюдений могут быть количество посетителей веб-сайтов, посещаемость занятий и объем продаж. Экспериментальные данные собираются в условиях эксперимента при участии групп активного воздействия и принятии проверенных временем мер предосторожности, позволяющих обеспечить целостность и избежать искажения результатов из-за смешивающихся переменных. Экспериментальные данные — это золотой стандарт. Благодаря тщательному планированию эксперимента, направленному на обеспечение надежности результатов, эти данные позволяют выявлять причинно-следственные связи. Например, экспериментальные данные могут помочь ответить на следующие вопросы[29]:

— Если мы дадим пациенту новое лекарство, поможет ли это вылечить его?
— Если мы дадим 15%-ную скидку на наш продукт, приведет ли это к росту продаж в следующем квартале?

Однако большая часть бизнес-данных относится к данным наблюдений. Для установления причинно-следственных связей не стоит использовать исключительно данные наблюдений[30]. Поскольку такие данные не были

[29] Обратите внимание на то, что подобные вопросы вам следует задать до начала реализации проекта по работе с данными, как было сказано в главе 1.

[30] Существуют способы использования данных наблюдений для выявления причинно-следственных связей, которые опираются на сильные предположения и продуманную статистику. Они называются методами выявления причинности.

собраны в ходе тщательно продуманного эксперимента, их полезность и основанные на них результаты должны оцениваться в соответствующем контексте. Любые утверждения о причинно-следственной связи, основанные на данных наблюдений, следует воспринимать скептически.

Задав вопрос о способе сбора данных, вы сможете понять, насколько обоснован вывод о наличии причинно-следственной связи. На самом деле некорректное установление причинности — весьма существенная проблема, к которой нам еще не раз предстоит вернуться в следующих главах книги.

Казалось бы, для решения этой проблемы достаточно как можно чаще использовать экспериментальные данные. Однако их сбор не всегда возможен, финансово оправдан и даже этичен. Например, если бы вам поручили изучить влияние «вейпинга» (курения электронных сигарет) на подростков, вы не смогли бы случайным образом разделить испытуемых на экспериментальную и контрольную группы и заставить участников первой группы курить электронные сигареты во имя науки. Это было бы неэтично.

Как главный по данным, вы должны работать с имеющимися у вас данными, одновременно опосредуя их способность влиять на принимаемые бизнес-решения. У некоторых компаний и отделов есть ресурсы, позволяющие проверить многообещающие данные наблюдений с помощью серьезных экспериментов. Однако далеко не все бизнес-проблемы поддаются экспериментальному анализу.

ЯВЛЯЮТСЯ ЛИ ДАННЫЕ РЕПРЕЗЕНТАТИВНЫМИ?

Вы должны убедиться в том, что имеющиеся у вас данные отражают характеристики интересующей вас совокупности. Если вас интересуют покупательские привычки американских подростков, то ваш набор данных должен отражать покупательские привычки всех подростков, живущих в США.

Индуктивная статистика существует именно потому, что у нас редко (если вообще когда-либо) есть все данные, необходимые для решения стоящей перед нами проблемы. Мы вынуждены опираться на выборки[31]. Однако если выборка нерепрезентативна, то выводы, сделанные на ее основе, не будут отражать реальные характеристики генеральной совокупности. Чтобы убедиться в репрезентативности данных, задайте следующие вопросы:

[31] Сбор всех сведений об интересующей совокупности называется переписью.

— Имеет ли место предвзятость выборки?
— Что вы сделали с выбросами?

Имеет ли место предвзятость выборки?

Предвзятость выборки возникает тогда, когда имеющиеся у вас данные систематически отклоняются или отличаются от тех данных, которые вас интересуют. Предвзятость выборки часто обнаруживается по косвенным признакам после принятия множества решений на основе данных, плохо отражающих ту проблему, для решения которой они были собраны. Систематическая неспособность получить предсказанный данными результат заставляет аналитиков вернуться к началу и проверить корректность исходных данных.

Если вы захотите узнать рейтинг одобрения политика на основе опроса избирателей, состоящих в его политической партии, ваша выборка будет предвзятой. Хороший план эксперимента позволяет предотвратить эту проблему.

В своей работе вы можете столкнуться с изначально предвзятыми данными. Данные наблюдений особенно подвержены подобной предвзятости. Вопрос: «Зачем данные были собраны?» поможет вам понять их назначение. При сборе подобных данных редко принимаются меры для обеспечения их непредвзятости.

Вам следует рассматривать все данные наблюдений как изначально предвзятые. Вам не нужно их отбрасывать, но вы всегда должны учитывать их недостатки.

Что вы сделали с выбросами?

Представьте, что в зарплатной ведомости компании вы видите цифру 50 000 000 долларов США рядом с именем нового управляющего. Вы бы посчитали это значение выбросом? Что бы вы с ним сделали?

Выбросы — это точки данных, которые значительно отличаются от всех остальных. Обнаружение выбросов должно спровоцировать дискуссию о том, какие данные следует исключить из анализа. Если кому-то не нравится влияние экстремального значения на результат анализа, это еще не значит, что от этого значения следует избавиться. Для удаления точки данных необходимо иметь хорошее обоснование.

Произвольное присвоение точкам данных статуса выбросов может привести к тому, что ваша выборка станет предвзятой. В случае исключения выброса исходная точка данных и причина ее исключения должны быть задокументированы и доведены до сведения остальных, особенно если это исключение привело к существенному изменению результата.

КАКИЕ ДАННЫЕ Я НЕ ВИЖУ?

Отсутствующие данные — это данные, которые либо не были зафиксированы (не имеют источника), либо вы их просто еще не видели. Рассмотрим следующие примеры:

— Данные о неполной занятости не учитываются при определении уровня безработицы.
— Компания, инвестирующая во взаимные фонды, «списывает» активы с плохой доходностью, в результате чего долгосрочная доходность оставшихся фондов в среднем оказывается выше.
— В истории «Челленджера» не было учтено 16 из 23 точек данных, связанных с полетами этого космического челнока.

Всегда стоит задумываться об информации, которая не была закодирована в рассматриваемых вами данных. Играйте в детектива[32].

Как вы поступили с отсутствующими значениями?

Отсутствующие значения — это буквально дыры в наборе данных. Они представляют собой точки данных, которые не были собраны, или исключенные выбросы (см. предыдущий раздел). Отсутствующие значения представляют проблему, но ее можно решить. Итак, всегда стоит спросить: «Как вы поступили с отсутствующими значениями?»

Предположим, вы работаете в компании, выпускающей кредитные карты, и собираете такие данные заявителей, как имя, адрес, возраст, статус занятости, доход, ежемесячные расходы на жилье и количество имеющихся банковских счетов. Ваша задача — предсказать, не просрочат ли эти заявители

[32] Мы вернемся к этой идее в одной из следующих глав при обсуждении так называемой систематической ошибки выжившего.

платеж в следующем году. Однако несколько заявителей не указывают свои доходы, из-за чего в системе сохраняется пробел — отсутствующее значение.

Вернемся к истории происхождения данных. Эта история начинается с подачи заявки на получение кредитной карты. Возможно, заявитель не указал свой доход, потому что думал, что ему откажут в выдаче кредитной карты, если его доход окажется слишком низким. Это означает, что сам факт отсутствия этого значения может говорить о возможной просрочке платежа в будущем. Такую информацию ни в коем случае не стоит отбрасывать!

Понимая это, дата-сайентист может создать новый категориальный признак под названием «Доход указан?» и ввести значение 1, если человек указал свой доход, и 0, если он этого не сделал. Таким образом, можно закодировать отсутствующие данные с помощью специальной категориальной переменной.

Позволяют ли данные измерить то, что вас интересует?

Мы часто верим в возможность измерить все и вся. Однако при анализе сложных идей, прежде чем что-то измерять, вам необходимо выяснить, позволяют ли предоставленные данные это сделать. Например, подумайте вот о чем:

— Как бы вы измерили лояльность клиента к вашей компании?
— Какие данные вы использовали бы для измерения «капитала бренда» или «репутации»?
— Какие данные могут показать, насколько сильно вы любите своего ребенка? Или домашнего любимца?

Все это очень трудно измерить. Благодаря кодированию информации данные позволяют нам приблизиться к ответам на эти вопросы, но в целом используемые нами данные представляют собой некоторую замену того, что мы пытаемся измерить. И степень, в которой такие данные отражают реальность, варьируется[33].

Поскольку измерение таких сложных показателей, как капитал бренда и репутация, требует косвенных приближений, вы должны быть максимально правдивыми и честными в отношении ваших данных.

[33] Производственным, инженерным и исследовательским организациям также следует позаботиться об определении повторяемости и воспроизводимости данных, измеряемых с помощью технического оборудования.

СОМНЕВАЙТЕСЬ В ДАННЫХ ЛЮБОГО РАЗМЕРА

Может показаться, что сбор большего количества данных позволяет решить проблемы, присущие ограниченным выборкам. Однако не стоит думать, что чем больше выборка, тем надежнее данные. Если данные собраны должным образом, то большая выборка может помочь, однако в случае наличия предвзятости дополнительные данные вас не спасут.

Недолговечная шумиха вокруг больших данных предполагала, что большее количество данных само по себе может обеспечить большую научную строгость. Не думайте, что набор данных слишком большой для того, чтобы с ним спорить. Статистика не предполагает какого-либо порогового значения для размера выборки, превышение которого автоматически избавляет ее от предвзятости. Статистика предполагает поиск компромиссов между тем, что вы хотите узнать, и имеющимися у вас данными[34].

ПОДВЕДЕНИЕ ИТОГОВ

Мы начали эту главу с обсуждения данных о катастрофе шаттла «Челленджер», но перенесли их на пример с автомобилем. Как было сказано в начале этой книги, умные люди и организации нередко допускают ошибки в данных.

Вот почему мы перечислили вопросы, которые вам следует задать, и различные проблемы, которые эти вопросы позволяют выявить. Мы рекомендуем вам использовать эти вопросы, чтобы глубже изучить проблемы, связанные с вашими данными. Вы можете самостоятельно придумать дополнительные вопросы. Мы настоятельно рекомендуем вам поделиться этими вопросами с вашей командой, чтобы согласовать усилия всех ее членов. Постоянно задавая сложные вопросы, главные по данным демонстрируют свою способность анализировать данные и подают хороший пример другим.

[34] При обдумывании подходящего размера выборки специалисты по статистике отталкиваются от величины мощности, о которой мы поговорим в главе 7.

ГЛАВА 5

Исследуйте данные

«Если вы заставляете дата-сайентиста выуживать данные… то заслуживаете тот плохой анализ, который в итоге получаете»[35].

_ *Томас Рэдман, «Доктор данных» и автор статей для журнала Harvard Business Review*

Реализация проектов по работе с данными никогда не оказывается такой простой, какой ее пытаются представить руководству. Как правило, заинтересованные стороны видят отполированную презентацию в PowerPoint, в которой описан четкий путь от вопроса к данным, необходимым для нахождения ответа. Однако здесь отсутствуют все важные решения и допущения, сделанные аналитиками на этом пути. Хорошая команда дата-сайентистов идет не линейным, а извилистым путем, адаптируясь к совершенным по дороге открытиям. По мере продвижения они возвращаются к более ранним этапам и обнаруживают новые пути.

Этот итеративный процесс обнаружения и тщательного изучения данных известен как разведочный анализ данных (EDA, exploratory data analysis). Он был предложен статистиком Джоном Тьюки в 1970-х годах в качестве способа осмысления данных с помощью сводной статистики и визуализации перед применением более сложных методов[36]. Тьюки рассматривал EDA как работу детектива, полагая, что подсказки скрыты в данных, а их правильный анализ может подсказать следующие шаги.

---

[35] Цитата из статьи “Understand Regression Analysis”, Amy Gallo, глава 10 в HBR Guide to Data Analytics Basics for Managers (HBR Guide Series).

[36] Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131–160).

По сути, EDA — это еще один способ «поспорить» с имеющимися у вас данными. Это фундаментальная часть всей работы с данными, которая одновременно задает и меняет направление развития проекта, исходя из сделанных открытий.

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ И ВЫ

Идея разведочного анализа данных может показаться кому-то некомфортной, поскольку она обнажает субъективную природу (искусство?) работы с данными. Если поставить перед двумя командами одну и ту же проблему и предоставить им одни и те же данные, то, используя разные методы анализа, они могут прийти к одинаковым или к разным выводам. На этом пути любые две команды (или два специалиста) вряд ли сделают все одинаково, поскольку для решения поставленной проблемы каждый человек будет использовать свой особый опыт, идеи и инструменты.

Поэтому в этой главе мы описываем разведочный анализ данных как непрерывный процесс, поддержание которого — обязанность главного по данным вне зависимости от того, является ли он рядовым специалистом или руководителем высшего звена. Вы узнаете, какие вопросы следует задавать и на что следует обращать внимание при изучении данных.

Вы менеджер или руководитель?

Если вы — заинтересованное лицо, менеджер или эксперт в предметной области, сделайте так, чтобы аналитики могли при необходимости с вами связаться. Ведите открытый диалог и будьте готовы к повторениям. Работайте с ними над выработкой корректных предположений. Не позволяйте команде заниматься выуживанием данных без понимания бизнес-контекста. В противном случае они могут пойти по пути, который имеет статистический, но не практический смысл. Одно неверное предположение может поставить под угрозу весь дальнейший анализ.

Мы прекрасно понимаем, что менеджеры не могут быть так же сильно погружены в тонкости проекта, как специалисты по работе с данными. Однако возможности для некоторого улучшения есть

всегда. Вам не нужно заниматься микроменеджментом. Просто не игнорируйте эту работу[37].

ОСВОЕНИЕ ИССЛЕДОВАТЕЛЬСКОГО ОБРАЗА МЫШЛЕНИЯ

Существуют десятки инструментов и языков программирования, способных помочь командам аналитиков без особых временных и денежных затрат изучить имеющиеся данные с использованием сводной статистики и визуализаций. Однако EDA следует рассматривать не как набор инструментов или контрольный список вопросов, а скорее как определенный образ мышления, вплетенный в каждый этап работы с данными, который вы можете использовать, даже не будучи профессиональным аналитиком.

Направляющие вопросы

Чтобы освоить исследовательский образ мышления и получить общее представление о процессе EDA, мы предлагаем вам рассмотреть краткий сценарий с использованием популярного набора данных Ames Housing Data (Данные о продаже домов в городе Эймс), созданного в образовательных целях[38].

Хотя единственно верного способа анализа этих данных не существует, для того чтобы помочь своей команде прийти к осмысленному выводу, вы можете задать следующие вопросы:

— Позволяют ли данные ответить на поставленный вопрос?
— Обнаружили ли вы какие-либо взаимосвязи?
— Обнаружили ли вы новые возможности в данных?

Давайте рассмотрим сценарий, а затем разберем каждый из этих трех вопросов, причины поиска ответа на них и проблемы, с которыми вы можете столкнуться.

---

[37] Заинтересованные стороны не должны заниматься микроменеджментом. Между бизнес-лидерами и командами, работающими с данными, должен быть определенный уровень доверия.

[38] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3). Данные можно загрузить с сайта **www.kaggle.com/c/house-prices-advanced-regression-techniques**.

Сценарий

Вы работаете в стартапе, занимающемся недвижимостью, и ваша задача — привлечь трафик на сайт. Однако вам трудно конкурировать с такими технологическими гигантами, как американская компания Zillow, чей знаменитый инструмент оценки стоимости жилья Zestimate®[39] привлекает большое количество людей (и денег) на сайт **Zillow.com**. Чтобы конкурировать с этим, вашему стартапу нужен собственный инструмент прогнозирования. Итак, перед вами поставлена задача построить модель, которая использует в качестве входных данных информацию о доме, а в качестве выходных данных выдает ориентировочную цену продажи.

Начальник присылает вам набор данных, в котором содержится 80 столбцов. Каждый из них описывает те или иные аспекты сотен жилых домов, проданных в городе Эймс, штат Айова, в период с 2006 по 2011 год.

Такое количество данных ошеломит кого угодно. Тем не менее перечисленные выше вопросы могут помочь вам приступить к их анализу.

Давайте разберем каждый их них.

ПОЗВОЛЯЮТ ЛИ ДАННЫЕ ОТВЕТИТЬ НА ПОСТАВЛЕННЫЙ ВОПРОС?

Как бы вам ни хотелось поскорее скормить данные новомодному алгоритму (например, воспользоваться методом глубокого обучения, описанным в главе 12), сначала следует спросить: «Позволяют ли данные ответить на поставленный вопрос?» И для получения ответа на него часто бывает достаточно просто взглянуть на имеющиеся данные.

Определитесь с ожиданиями и руководствуйтесь здравым смыслом

Вы должны иметь довольно хорошее представление о том, какая информация необходима для определения цены продажи дома, например, общая площадь, количество спален, количество ванных комнат, год постройки и так далее. Эти характеристики чаще всего интересуют потенциальных покупателей жилья, заходящих на ваш веб-сайт. Без их учета предсказание цены дома не кажется разумным.

[39] Компания Zillow очень серьезно относится к Zestimate®. В 2019 году она выделила 1 миллион долларов команде дата-сайентистов ради повышения точности прогнозов этого инструмента. **venturebeat.com/2019/01/30/zillow-awards-1-million-to-team-that-reduced-home-valuation-algorithm-error-to-below-4**

Открыв файл, вы видите названия столбцов и типы данных. В нем присутствуют вполне ожидаемые признаки, а также полезные порядковые данные (например, «Общее качество дома, 1–10, где 10 означает «Превосходное»), номинальные данные («Окрестности») и множество других признаков. На первый взгляд, с данными все в порядке.

На следующем этапе вы, вероятно, решите изучить значения, которые принимают переменные. Охватывают ли они те сценарии, которые вы хотите проанализировать? Например, если вы обнаружите, что переменная «Тип здания: тип жилища» принимает только одно значение — «Дом на одну семью», но не включает квартиры, дуплексы или кондоминиумы, то ваша модель будет иметь ограниченный охват по сравнению с моделью компании Zillow. Ее инструмент Zestimate® может предсказать цену продажи кондоминиума — но, если у вас нет исторических данных о них, модель вашей компании не сможет надежно предсказать его цену.

Мораль заключается в следующем: не выуживайте данные, как говорилось в цитате, приведенной в начале главы. Убедитесь в том, что данные позволяют ответить на поставленный вопрос.

Имеют ли данные интуитивный смысл?

Программное обеспечение сгенерирует для вас множество сводных статистических показателей. Ваша задача — поместить эти данные в контекст. Оцените соответствие этой сводной статистики своему интуитивному пониманию проблемы. Еще один ключевой компонент EDA — визуализации. Используйте их для обнаружения аномалий и других странностей в данных.

Визуализация данных

Давайте рассмотрим несколько примеров проведения разведочного анализа данных с использованием гистограмм, диаграмм размаха, столбиковых графиков и диаграмм рассеяния. Если вы уже хорошо знакомы с такими графиками, то можете пропустить этот раздел.

Гистограммы позволяют определить форму распределения непрерывных числовых данных. Рассмотрим гистограмму продажных цен, изображенную на рис. 5.1. На ней мы видим около 125 домов

стоимостью до 200 000 долларов и длинный хвост справа, сформированный самыми дорогими домами. Из-за этого хвоста средняя цена продажи (181 000 долларов) превышает медианную цену (163 000 долларов). То есть из-за небольшого количества дорогих домов среднее значение превышает медианное.

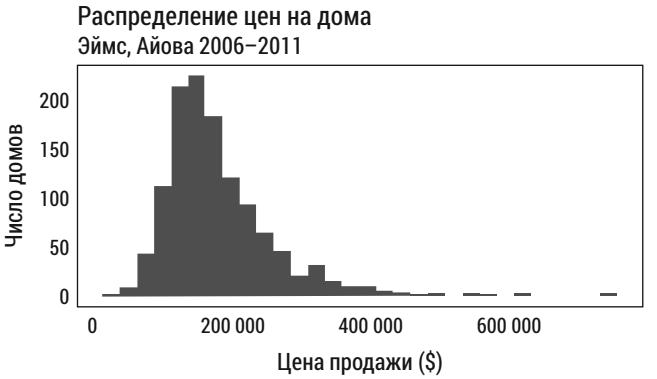


Рис. 5.1. Гистограмма, отражающая форму распределения цен на дома

Гистограммы помогают обнаруживать аномалии. Если бы вы увидели отрицательные значения, говорящие о получении покупателем платы за покупку дома, или неожиданно большие значения у правого края графика на рис. 5.1, что бывает при задании максимального значения (например, когда любое значение, превышающее 500 000 долларов, записывается как 500 000 долларов), вам бы захотелось задать дополнительные вопросы.

Диаграммы размаха[40] можно использовать для сравнения данных, принадлежащих нескольким группам. На рис. 5.2 показана диаграмма размаха для каждого рейтинга качества дома, где 1 означает плохое, а 10 — превосходное.

[40] Диаграммы размаха также называют диаграммами типа «ящик с усами». «Ящик» содержит центральные 50% наблюдений (значения в диапазоне между 25-м и 75-м процентилями), линия в ящике — это медиана, а «усы» показывают диапазон, в котором находятся оставшиеся точки данных. Точки, выходящие за пределы этого диапазона, — потенциальные выбросы.

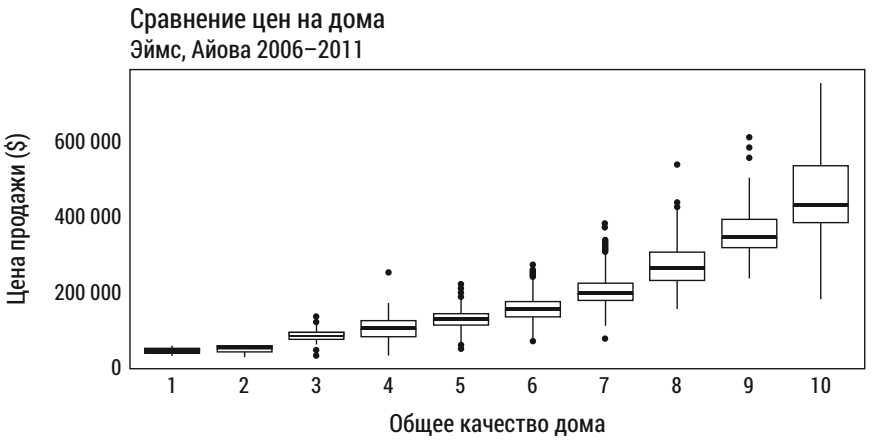


Рис. 5.2. Использование диаграмм размаха для сравнения продажных цен при различных рейтингах качества

В данном случае взаимосвязь между общим качеством дома и его ценой кажется интуитивно понятной. Более качественные дома обычно продаются по более высокой цене. Мы можем обнаружить дом за 200 000 долларов, общее качество которого было оценено на 10 (нижний конец линии). Однако разумно предположить, что он был продан дешевле, чем другие дома с оценкой 10 из-за прочих факторов. Специалистам по работе с данными следует проверять такого рода информацию.

Столбиковые графики (рис. 5.3) отображают распределение категориальных данных.

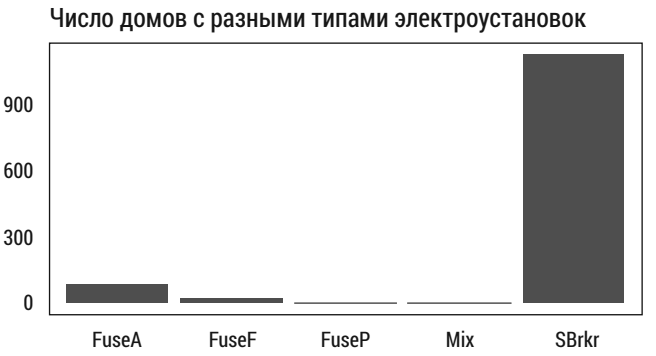


Рис. 5.3. Столбиковый график, показывающий количество домов с разными типами электроустановок

Не все виды визуализаций могут показаться интересными на первый взгляд. Тем не менее ознакомиться с ними все равно стоит — хотя бы для того, чтобы подтвердить (или оспорить) ответ на вопрос: «Имеют ли данные интуитивный смысл?» Согласно графику на рис. 5.3, почти все дома имеют одинаковое значение указанного признака. Однако с точки зрения поставленной перед вами задачи эта информация полезна. Поскольку значение этой переменной одинаковое для большинства домов, она, вероятно, не будет существенно влиять на разницу в их стоимости.

Рис. 5.4. Линейная диаграмма, отражающая количество домов, проданных в разные месяцы

На рис. 5.4 показана линейная диаграмма, отражающая количество домов, проданных в разные месяцы. Явление, при котором продажи домов увеличиваются летом и сокращаются зимой, называется сезонностью. Линейные диаграммы хорошо отражают такие тенденции.

На следующем этапе мы можем изучить диаграмму рассеяния, демонстрирующую зависимость цены дома от его размера (площади первого этажа в квадратных футах).

Зависимость, отображенная на рис. 5.5, интуитивно понятна. Большие дома обычно стоят дороже. Разумеется, из этого правила есть исключения: иногда небольшие дома стоят дороже, чем большие. Вариации есть всегда, но они не отменяют общую тенденцию. И поскольку в конечном итоге мы пытаемся предсказать цену продажи дома, его площадь — весьма полезная информация.

Рис. 5.5. Диаграмма рассеяния, отражающая площадь в квадратных футах и цену продажи

В этом разделе мы лишь в общих чертах обсудили различные способы визуализации данных и то, какую информацию можно быстро получить с их помощью. Если вы хотите глубже изучить методы использования визуализации в процессе исследования данных, мы рекомендуем ознакомиться со следующими книгами:

— *Now You See it: Simple Visualization Techniques for Quantitative Analysis*, Stephen Few (Analytics Press, 2009);
— *The Visual Display of Quantitative Information*, Edward Tufte (Graphics Press, 2011).

Осторожно: выбросы и отсутствующие значения

В каждом наборе данных будут наблюдаться аномалии, выбросы и пропущенные значения. Что с ними можно сделать?

Например, в диаграмме размаха на рис. 5.2 использовалось эмпирическое правило для того, чтобы отметить несколько точек данных в качестве возможных выбросов. Однако вам не следует отключать критическое мышление и автоматически удалять подобные точки как потенциально бесполезные только потому, что на графике они классифицированы как «выбросы». Компания Zillow никогда не удаляет полезную информацию из своих

наборов данных просто потому, что средство визуализация приняло их за выбросы. Учитывайте контекст данных: в мире недвижимости нередко встречаются дома, которые стоят намного больше, чем большинство других домов. Вспомните уроки из предыдущей главы. Для удаления выбросов вы должны иметь хорошее обоснование. Есть ли оно у вас?

А как быть с отсутствующими значениями? Означает ли отсутствие значения в поле «Размер подвала» то, что в доме есть подвал, но нам неизвестна его площадь? Или это значит, что подвала нет, и значение должно быть равно 0?

Мы имеем право забрести в дебри. Специалисты по работе с данными принимают сотни подобных решений в ходе реализации проектов. Однако их суммарный эффект может оказаться весьма значительным. Предоставленные самим себе и лишенные руководства со стороны экспертов в предметной области аналитики могут отбрасывать сложные и нюансированные случаи до тех пор, пока данные не станут слишком оторванными от той реальности, которую они призваны описать. Вот почему всем, включая менеджеров, важно четко понимать, чем занимаются команды дата-сайентистов.

ОБНАРУЖИЛИ ЛИ ВЫ КАКИЕ-ЛИБО ВЗАИМОСВЯЗИ?

К счастью, первые сводные статистические показатели и результаты первой визуализации данных о домах кажутся обнадеживающими, и вы думаете, что эти данные действительно могут быть использованы при построении модели для прогнозирования цены продажи. Поэтому вы переходите к следующему вопросу: «Обнаружили ли вы какие-либо взаимосвязи?»

Визуализация данных показала, что более высокое общее качество дома и его большая площадь связаны с более высокими ценами, и это неудивительно. Это та обратная связь, которую вы хотите получить от данных. Эти взаимосвязи имеют смысл, и выбранные вами переменные будут использоваться при построении модели для прогнозирования стоимости дома. Какие еще переменные могут быть связаны с его ценой продажи?

На данном этапе для обнаружения в данных интересных закономерностей и взаимосвязей имеет смысл использовать сводную статистику, поскольку построение всех возможных диаграмм рассеяния может оказаться нецелесообразным. Вместо этого взаимосвязи, обнаруженные на таких диаграммах, могут быть сведены к статистической корреляции, которая допускает (но не доказывает) существование взаимосвязи между двумя числовыми переменными.

Рис. 5.6. Коэффициент корреляции между площадью дома и ценой продажи составляет 0,62 (определяется степенью близости точек данных к линии тренда)

Корреляция

Корреляция — это мера связанности двух переменных. Наиболее распространенный коэффициент корреляции в сфере бизнеса — коэффициент корреляции Пирсона. Он принимает значения в диапазоне от –1 до 1 и измеряет степень линейной зависимости (простая прямая линия) между парами чисел, отображаемыми на диаграмме рассеяния. Корреляция может быть положительной, когда увеличение одной переменной сопровождается увеличением другой: большие дома продаются за большие деньги. Корреляция также может быть отрицательной: более тяжелые автомобили менее экономичны в плане расхода топлива. Коэффициент корреляции между размером дома и ценой продажи составляет 0,62 (рис. 5.6). Чем ближе точки к линии тренда, тем выше степень корреляции[41].

В данном случае корреляция может помочь двумя способами. Во-первых, нахождение переменных, коррелирующих с ценой продажи, упрощает ее предсказание. Во-вторых, корреляция позволяет уменьшить избыточность данных, поскольку две сильно коррелированные переменные содержат примерно одинаковую информацию. Представьте два столбца с данными, в одном из которых площадь дома указана в квадратных футах, а во втором — в квадратных метрах. Эти значения идеально коррелируют между собой, и для проведения анализа достаточно только одного из них.

[41] Корреляция вовсе не означает, что наклон линии должен быть крутым. Идеальная корреляция между двумя переменными вполне может описываться почти плоской (хоть и не горизонтальной) линией.

Хотя большинство из нас имеет базовое представление о корреляции и часто использует ее, данная метрика может ввести в заблуждение. Давайте разберемся, почему.

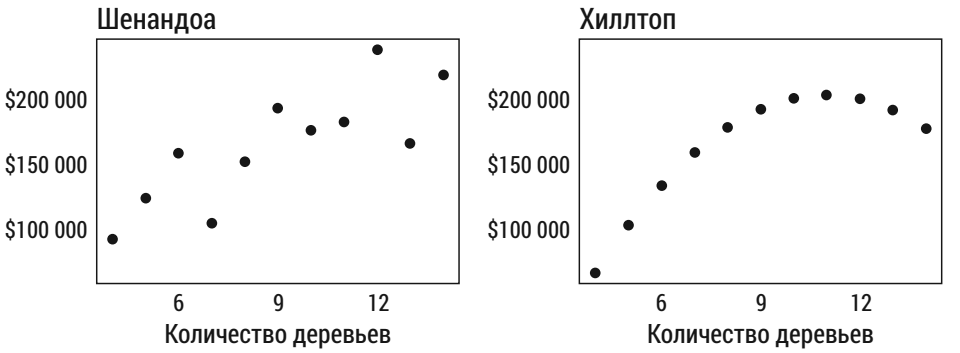


Рис. 5.7. Два набора данных с коэффициентом корреляции 0,8

Осторожно: неверная интерпретация корреляции

Люди часто забывают о том, что корреляция — это мера линейной зависимости, но не все зависимости линейны.

Предположим, что вы анализируете данные по двум районам, в каждом из которых находится по 11 домов. Статистический анализ показывает, что количество деревьев на участке сильно коррелирует с ценой домов в этих районах. Коэффициент корреляции равен 0,8: дома с большим количеством деревьев на участке, как правило, продаются дороже.

Однако визуализация данных показывает нечто неожиданное. На рис. 5.7 слева показана вполне ожидаемая для высокой корреляции картина: линейный тренд с разбросанными вокруг него точками данных. Однако график справа показывает, что количество деревьев положительно коррелирует с ценой дома только до определенной точки (11 деревьев), после которой тенденция меняется на противоположную. В районе Хиллтоп на газонах у некоторых домов деревьев может быть слишком много.

Данные, представленные на рис. 5.7, взяты не из набора данных о недвижимости в Эймсе, с которым мы работали до этого, а из популярного набора данных под названием «Квартет Энскомба»[42]. У него четыре набора числовых

[42] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21. Для получения значений, напоминающих цены на дома, мы умножили зависимую переменную на 22 000.

данных, имеющих идентичные сводные статистические показатели, но разные результаты визуализации. (Здесь мы привели только два и скорректировали данные в соответствии с темой недвижимости.)

Мораль: используйте методы визуализации для проверки заслуживающих внимания корреляций в данных, потому что выявленная линейная зависимость может не рассказать всей истории.

Корреляция отсутствует, но все равно интересно

На рис. 5.8 показаны два графика, которые имеют одинаковый близкий к нулю коэффициент корреляции. Однако это не значит, что на них не происходит ничего интересного. С «датазавром», изображенным на левом графике, вам вряд ли доведется столкнуться, чего нельзя сказать о сценарии на правом графике. На нем на самом деле отображены пять групп линейно коррелированных данных, которые при рассмотрении их как единой группы оказываются линейно некоррелированными. Это явление известно как парадокс Симпсона, и мы поговорим о нем более подробно в главе 13.

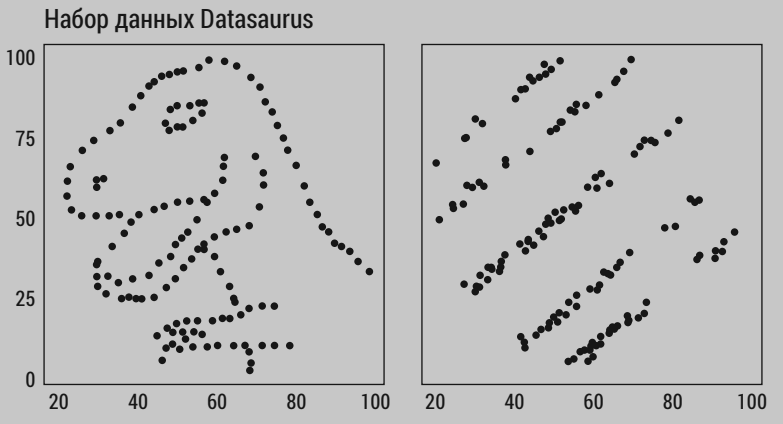


Рис. 5.8. Набор данных Datasaurus можно загрузить бесплатно[43]. Как и в случае с «Квартетом Энскомба», оба представленных здесь набора данных имеют идентичные сводные статистические показатели

[43] Набор данных Datasaurus был создан Альберто Каиро и доступен на GitHub: **github.com/lockedata/datasauRus**

Осторожно: корреляция не означает причинность

Скорее всего, вы уже слышали фразу «корреляция не означает причинность»[44]. Однако повторить ее будет нелишним, учитывая, как часто ее игнорируют и неправильно понимают.

Когда две переменные коррелируют между собой, пусть даже и сильно, это не означает, что одна влияет на другую. Однако многие люди попадают в эту ловушку, пытаясь объяснить корреляцию между двумя переменными наличием причинно-следственной связи между ними. Чтобы показать, что корреляция не подразумевает причинность, статистики используют максимально абсурдные примеры. В частности, продажи мороженого коррелируют с нападениями акул (в обоих случаях пик приходится на летние месяцы). Размер обуви коррелирует с навыками чтения (и то и другое увеличивается с возрастом). Однако предположения о том, что сокращение объема продаж мороженого может снизить риск нападения акул, а покупка обуви большего размера может улучшить навыки чтения, абсурдны. Очевидно, что помимо температуры воздуха на улице в примере с мороженым и возраста в примере с размером обуви есть и другие факторы, играющие роль в формировании этих мнимых взаимосвязей.

Однако в тех случаях, когда в основе корреляции не лежит откровенная шутка, а истинный причинный фактор не известен, о мантре «корреляция не означает причинность» очень часто забывают.

Например, в ходе анализа данных о недвижимости вы обнаруживаете, что показатели школьной успеваемости коррелируют со стоимостью домов. Означает ли это, что близость хорошей школы повышает стоимость дома? Хорошие школы, по-видимому, делают район более привлекательным. А может быть, наоборот: более высокие цены на жилье способствуют повышению школьной успеваемости? Возможно, благодаря увеличению налоговых поступлений школе выделяется больше ресурсов. А может быть, причинно-следственная связь действует в обоих направлениях, создавая петлю обратной связи? В большинстве случаев мы точно этого не знаем. Здесь сочетаются многие факторы, и в имеющемся у нас наборе данных редко можно найти все ответы.

---

[44] Авторы этого руководства поспорили о том, можно ли вообще не упомянуть эту фразу в книге, посвященной науке о данных. О результате этого спора вы можете догадаться сами.

Всегда безопаснее предполагать, что между двумя коррелирующими переменными «нет причинно-следственной связи», если только кто-то не провел эксперимент, доказывающий обратное. Однако не стоит впадать в крайности. Мы по собственному опыту знаем, что иногда компании, академики и СМИ предполагают наличие причинно-следственной связи там, где этого делать не следует, а иногда наоборот — отвергают важную взаимосвязь, приняв ее за ошибку. Пример подобного необоснованного игнорирования взаимосвязи описан в следующей врезке.

Курение и рак легких

Рональд Э. Фишер, один из ведущих статистиков XX века, участвовавший в разработке ряда методов, описанных в этой книге, довольно скептически относился к исследованиям, связывавшим курение табака с заболеваемостью раком.

Больше всего Фишера заботили смешивающиеся переменные. Например, что, если некоторые люди генетически предрасположены к развитию рака легких и курят для того, чтобы облегчить симптомы болезни? По словам Фишера, ранние исследования рисков употребления табака содержали «издавна известную ошибку, выражавшуюся в том, что вывод о причинности делался на основе корреляции»[45].

Однако теперь мы точно знаем, что связь между ними есть. Итак, нам следует проявлять осторожность не только для того, чтобы не увидеть причинность там, где ее нет, но и чтобы не проигнорировать ее там, где она пока еще не доказана.

ОБНАРУЖИЛИ ЛИ ВЫ НОВЫЕ ВОЗМОЖНОСТИ В ДАННЫХ?

Разведочный анализ данных — это не просто процесс, позволяющий лучше разобраться в данных и наметить путь решения стоящих перед нами проблем. Это еще и шанс найти дополнительные возможности в этих данных, которые могут оказаться ценными для вашей организации. Дата-сайентист может обнаружить что-то интересное или странное в наборе данных и сформулировать проблему.

[45] Fisher, R. A. (1958). Cancer and smoking. *Nature*, 182 (4635), 596.

Однако вы не сможете оценить важность найденного вами решения до тех пор, пока не выполните действия, описанные в главе 1 «В чем суть проблемы?»

ПОДВЕДЕНИЕ ИТОГОВ

Чтобы стать главным по данным, вам необходимо постоянно заниматься разведочным анализом данных. Это позволит вам:

— Наметить более четкий путь решения проблемы.
— Уточнить исходную бизнес-задачу с учетом выявленных в данных ограничений.
— Сформулировать новые проблемы, которые можно решить с помощью этих данных.
— Отменить проект. Хотя это не приносит удовлетворения, EDA считается успешным, если он предотвращает трату времени и денег на решение тупиковой проблемы.

Мы провели вас через весь процесс, используя набор данных о ценах на недвижимость (к которому вернемся в главе 9 для построения предсказательной модели), и рассказали о тех препятствиях, с которыми вы можете столкнуться.

Содержание этой главы предполагает ваше участие во всех этапах процесса EDA. Однако иногда это невозможно, особенно для старших руководителей, курирующих множество проектов. Тем не менее пропуск ранних этапов не освобождает главных по данным от обязанности придерживаться исследовательского образа мышления. Подключаясь к проекту на завершающих этапах его реализации, спросите аналитиков, почему они выбрали тот или иной метод анализа данных и с какими проблемами столкнулись. Так вы можете узнать о предположениях, которые сами бы не сделали.

ГЛАВА 6

Изучайте вероятности

«Представления многих людей о вероятности настолько скудны, что они допускают только [одно] из двух ее значений: 50 на 50 и 99%, то есть абсолютную случайность и практически полную уверенность»

_ Джон Аллен Паулос, математик и автор книги «Математическое невежество и его последствия»[46]

Давайте поговорим о вероятности — языке неопределенности — и вернемся к теме, рассмотрение которой мы начали в главе 3 «Готовьтесь мыслить статистически». Напомним, что во всем присутствует вариация. Вариация порождает неопределенность. А теория вероятности и статистика — это инструменты, помогающие нам управлять неопределенностью.

Тот краткий раздел, посвященный вероятности, закончился следующим напутствием: будьте внимательны и помните о том, *что интуиция может сыграть с вами злую шутку.*

Это справедливое утверждение, однако такие темы, как вероятность, заслуживают больше этого предупреждения. Полное ее понимание, если оно вообще возможно, требует прочтения огромного количества учебников, прослушивания длинных лекций и посвящения всей жизни исследованиям и дебатам. И даже это не гарантирует согласия экспертов относительно интерпретации и философии вероятности[47]. У вас, скорее всего, нет времени или желания вникать в подробности этого спора; у нас его тоже нет. Поэтому

[46] Паулос, Дж. А. «Математическое невежество и его последствия» (Издательство: Студия Артемия Лебедева, 2021).

[47] Поищите в Интернете «Интерпретации вероятности», чтобы понять, что мы имеем в виду.

мы избавим вас от них и сосредоточим внимание на том, что поможет вам отточить интуицию и добиться успеха в своей работе.

Итак, цель этой главы — помочь вам углубиться в теорию вероятностей, освоить соответствующий язык и обозначения, а также познакомиться с инструментами и ловушками. К концу этой главы вы сможете думать и говорить о вероятностях на своем рабочем месте, даже если сами не занимаетесь расчетами, а также задавать сложные вопросы о представленных вам вероятностях. Готовность погрузиться в тему вероятности и неопределенности — важный шаг на пути становления главным по данным.

ПОПРОБУЙТЕ УГАДАТЬ

Для начала попробуйте выполнить мысленное упражнение.

Ваша компания, входящая в список Fortune 500, стала жертвой кибератаки: хакеры заразили вирусом 1% всех портативных компьютеров. Доблестная IT-команда быстро разработала способ проверки ноутбука на предмет наличия на нем этого вируса. Это очень хороший, почти идеальный тест. Исследования IT-команды показали, что при наличии в ноутбуке вируса результат теста будет положительным в 99% случаев. А при отсутствии вируса в 99% случаев результат теста будет отрицательным.

При проверке вашего ноутбука на наличие вируса результат оказывается положительным. Какова вероятность того, что на вашем устройстве действительно есть вирус?

Подумайте над этим, прежде чем двигаться дальше.

Правильный ответ — 50%. (Мы докажем это далее в этой главе.)

Удивлены? Это удивляет большинство людей.

Ответ не понятен интуитивно. Даже если вы знаете, что вероятность может сыграть с вами злую шутку, она все равно может вас подловить. Именно это больше всего раздражает в теории вероятности — любая проблема становится настоящей головоломкой. Однако не стоит расстраиваться, если вы не угадали правильный ответ. Настоящий тест заключался в том, задумались ли вы о своей неуверенности в ответе.

Далеко не все это делают. Большинство людей не понимают или не учитывают вероятности. Хотите доказательства? Люди по-прежнему покупают лотерейные билеты, стекаются в Лас-Вегас и приобретают расширенную гарантию на свои телевизоры. Они довольствуются своим прискорбным невежеством в отношении вероятности, особенно когда принимаемые ими

решения связаны с потенциальной выгодой (игровые автоматы) или возможностью избежать проблем в будущем (гарантии на телевизоры). Эта глава даст вам четкое представление о вероятности, правилах ее определения и ошибочных представлениях.

Итак, начнем.

ПРАВИЛА ИГРЫ

Теория вероятностей позволяет количественно оценить возможность наступления того или иного события.

Прежде чем мы погрузимся в математику, стоит отметить, что наш мозг запрограммирован на работу с вероятностями. В повседневной жизни мы постоянно используем вероятностные утверждения. Вы не можете точно знать, произойдет ли то или иное событие в вашей жизни, но вы знаете, что некоторые исходы более вероятны, чем другие. Например, в офисе вы можете услышать фразы наподобие:

— «Вполне вероятно, что они подпишут контракт!»
— «Существует небольшая вероятность того, что мы пропустим крайний срок, назначенный на следующий понедельник».
— «Вряд ли нам удастся достичь квартальных целей».
— «Тревор, как правило, опаздывает на совещания».
— «Согласно прогнозу погоды, сегодня, скорее всего, будет дождь. Давайте перенесем выездную встречу».

У двух людей могут быть разные представления о том, как часто происходит «весьма вероятное» или «вероятное» событие, а значит, обыденный язык здесь не поможет. Нам нужно использовать числа, данные и обозначения для количественной оценки вероятностных утверждений, чтобы наши заявления стали надежнее интуитивных догадок (даже если наша интуиция отличается высокой степенью надежности). Более того, нам нужно соблюдать определенные правила и логику вероятности.

Нотация

Как говорилось ранее, теория вероятностей позволяет количественно оценить возможность наступления того или иного события. Событием может

быть любой исход — от простого (выпадение орла при подбрасывании монеты) до сложного («Дональд Трамп победит на выборах 2016 года»). Даже ребенок может оценить вероятность выпадения орла при подбрасывании монеты как 50 на 50, однако вся индустрия опросов общественного мнения не сумела предсказать результаты выборов 2016 года, несмотря на анализ терабайтов данных.

В этом кратком уроке мы рассмотрим простые случаи.

Вероятность принимает значения в диапазоне от 0 до 1 включительно, где 0 означает невозможность (выпадение 7 при бросании шестигранного кубика с цифрами 1–6), а 1 — абсолютную уверенность (выпадение числа меньшего 7 при бросании шестигранного кубика). Вероятность часто выражается в виде простой дроби (вероятность выпадения орла при подбрасывании монеты составляет 1/2) или в процентах (у вас есть 25%-ный шанс выбрать карту пиковой масти из стандартной колоды игральных карт). Многие люди при описании вероятности используют числа, дроби и проценты взаимозаменяемо.

Для экономии места мы будем использовать сокращение и обозначать вероятность буквой *P*. Описания событий мы также будем сокращать. Например, фразу «Вероятность выпадения орла при подбрасывании честной монеты равна 1/2» можно кратко записать в виде $P(\text{М} == \text{О}) = 1/2$. Или, еще короче, $P(\text{О}) = 1/2$. Фактически весь предыдущий абзац можно показать в виде следующей таблицы.

Табл. 6.1. Сценарии, описанные с помощью сокращенной нотации

Сценарий	Нотация
Вероятность выпадения 7 при бросании шестигранного кубика	$P(\text{К} == 7) = 0$
Вероятность выпадения числа меньшего 7 при бросании шестигранного кубика	$P(\text{К} < 7) = 1$
Вероятность выбора карты пиковой масти из колоды карт	$P(\text{П}) = 0{,}25$

Использование «==» вместо «=»

Если вы уже проходили курс по теории вероятностей или статистике, используемые обозначения вам, скорее всего, знакомы. Однако для большей ясности мы добавили еще кое-что.

Обратите внимание: когда мы проверяем вероятность выпадения орла при подбрасывании монеты, мы пишем $P(М == О)$ вместо $P(М = О)$. Мы делаем это для того, чтобы провести различие между двумя наборами знаков равенства в нашем уравнении. С помощью двойного знака равенства (==) мы фактически проверяем результат подбрасывания монеты М.

С другой стороны, когда мы пишем $P(М == О) = 1/2$, единственный знак равенства в конце записи указывает на то, что результат $P(М == О)$ равен 1/2.

Эта нотация соответствует синтаксису булевой логики, используемому во многих языках программирования.

Выражение $P(К < 7) = 1$ обозначает суммарную вероятность и говорит о том, что «Вероятность выпадения числа меньшего 7 при бросании шестигранного кубика равна 1». Этот результат получается путем сложения $P(К == 1) + P(К == 2) + P(К == 3) + P(К == 4) + P(К == 5) + P(К == 6) = 6 \times 1/6 = 1$ (табл. 6.2). Сумма вероятностей всех исходов должна равняться единице.

Табл. 6.2. Суммарная вероятность выпадения числа меньшего 7 при бросании кубика

Сценарий	Нотация	Вероятность
Выпадение 1	$P(К == 1)$	1/6
Выпадение 2	$P(К == 2)$	1/6
Выпадение 3	$P(К == 3)$	1/6
Выпадение 4	$P(К == 4)$	1/6
Выпадение 5	$P(К == 5)$	1/6
Выпадение 6	$P(К == 6)$	1/6
Выпадения числа меньшего 7	$P(К < 7)$	6/6 = 1 = 100%

Условная вероятность и независимые события

Когда вероятность наступления одного события зависит от наступления другого, это называется условной вероятностью. Условная вероятность

обозначается вертикальной чертой, |, которая читается как «при условии». Вот несколько примеров для большей ясности:

— Вероятность того, что Алекс опоздает на работу, составляет 5%. $P(\text{А}) = 5\%$.
— Вероятность того, что Алекс опоздает на работу при условии, что у него спустит колесо (С), равна 100%. $P(\text{А} \mid \text{С}) = 100\%$.
— Вероятность того, что Алекс опоздает на работу при условии, что на межштатной автомагистрали 75 будет пробка (П), составляет 50%. $P(A \mid \text{П}) = 50\%$.

Как видите, вероятность наступления события сильно зависит от предшествующего ему события или событий.

Когда вероятность наступления одного события не зависит от наступления другого, эти события считаются независимыми. Например, условная вероятность выбора карты пиковой масти из колоды карт при условии выпадения орла при подбрасывании монеты $P(\text{П} \mid \text{О})$ равна вероятности выбора карты пиковой масти самого по себе, $P(\text{П})$. Короче говоря, $P(\text{П} \mid \text{О}) = \text{Р}(\text{П})$, и точно так же $P(\text{О} \mid \text{П}) = P(\text{О})$, потому что между этими событиями нет никакой зависимости. Колоде карт все равно, что произошло с монетой, и наоборот.

Вероятность наступления множества событий

При моделировании вероятности наступления множества событий нотация и правила зависят от того, происходят ли они одновременно (наводнение и отключение электричества) или происходит только одно из них (или наводнение, или отключение электричества).

Одновременное наступление двух событий

Сначала поговорим о двух событиях, наступающих одновременно.

P(выпадения орла при подбрасывании монеты) $= P(\text{О}) = 1/2$.
P(выбора карты пиковой масти из колоды карт) $= P(\text{П}) = 13/52 = 1/4$.

Вероятность того, что произойдет и то и другое, то есть выпадение орла и выбор карты пиковой масти, можно обозначить как $P(\text{О}, \text{П})$. При этом запятая означает «и».

В этом случае события являются независимыми. Одно событие не влияет на другое. Когда события являются независимыми, вероятности их наступления можно перемножить: $P(О, П) = P(О) \times P(П) = 1/2 \times 1/4 = 1/8 = 12{,}5\%$. Тут все довольно просто.

Теперь рассмотрим чуть более сложный пример. Как вы помните, вероятность того, что Алекс опоздает на работу, составляет 5%, $P(A) = 5\%$. А вероятность того, что Джордан опоздает на работу, составляет 10%, $P(Д) = 10\%$. Что вы можете сказать о вероятности того, что мы оба опоздаем на работу, $P(A, Д)$? Уточним, что мы живем в разных штатах, Алекс работает в офисе с 9 до 5, а Джордан — фрилансер[48].

Первое предположение: $P(A, Д) = P(A) \times P(Д) = 5\% \times 10\% = 0{,}5\%$. Вероятность довольно низкая, но действительно ли эти два события независимы друг от друга? Поначалу может показаться, что так и есть, поскольку мы живем и работаем в разных местах. И все же эти события не являются независимыми. В конце концов, мы вместе пишем книгу. Мы оба могли опоздать на работу, потому что накануне вечером допоздна спорили о лучшем способе объяснения концепции вероятности. Таким образом, вероятность опоздания Алекса зависит от опоздания Джордана. Поэтому здесь речь идет об условной вероятности. Предположим, что вероятность опоздания Алекса при условии опоздания Джордана составляет 20%, $P(A \mid Д) = 20\%$.

Это дает нам истинную формулу вероятности одновременного наступления этих двух событий, называемую правилом умножения. Ее можно записать следующим образом: $P(A, Д) = P(Д) \times P(A \mid Д) = 10\% \times 20\% = 2\%$. Это значит, что вероятность одновременного опоздания Алекса и Джордана равна вероятности опоздания Джордана, умноженной на вероятность того, что Алекс опоздает при условии опоздания Джордана.

Итоговая вероятность, 2%, никогда не может превышать наименьшую из отдельных вероятностей, $P(A)$ и $P(Д)$, которая в данном случае составляет 5% для Алекса. Это объясняется тем, что у Алекса есть 5%-ный шанс опоздать во всех возможных сценариях, включая те, в которых опаздывает Джордан.

Это подводит нас к важному правилу теории вероятностей: вероятность одновременного наступления любых двух событий не может превышать вероятность наступления каждого из них в отдельности.

На рис. 6.1 это правило проиллюстрировано с помощью диаграммы Венна. Если представить вероятность в виде области пересечения или перекрытия

[48] Разве можно опоздать на работу, работая на себя? В этом примере — да.

кругов (событий), становится очевидно, что площадь области перекрытия кругов А и Д не может превышать площадь самого маленького круга.

Наступление одного или другого события

Что, если наступает одно или другое событие? Статистика и теория вероятностей учит нас тому, что все зависит от обстоятельств. Начните с предположения и корректируйте его, опираясь на имеющуюся информацию.

Когда два события не могут произойти одновременно, все сводится к простому сложению вероятностей. При бросании кубика не может одновременно выпасть 1 и 2, поэтому вероятность выпадения 1 или 2 равна P(К == 1 или К == 2) = P(К == 1) + P(К == 2) = 1/6 + 1/6 = 2/6 = 1/3.

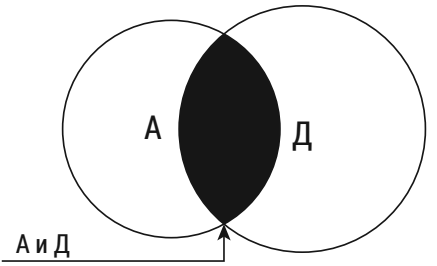


Рис. 6.1. Диаграмма Венна, показывающая то, что вероятность одновременного наступления двух событий не может превышать вероятность наступления каждого из них в отдельности

Рассмотрим чуть более сложный пример с авторами-прогульщиками и вместо определения вероятности того, что на работу опоздают и Алекс, и Джордан, вычислим вероятность опоздания Алекса или Джордана, то есть P(А или Д).

Вам известно, что P(А) = 5%, а P(Д) = 10%. Первым разумным предположением может быть: P(А) + P(Д) = 15%. За 100 дней Алекс опоздает 5 раз, а Джордан — 10. Если мы сложим эти значения, то получим 15 дней, что составляет 15% от 100. Если бы события были взаимоисключающими и никогда не происходили одновременно, это предположение было бы корректным.

Однако помните о том, что мы оба можем опоздать (см. рис. 6.1.). Иногда мы опаздываем на работу друг из-за друга, то есть вероятность того, что опоздают и Алекс, и Джордан, P(А, Д), превышает 0. Мы не можем просто сложить обе вероятности, потому что при этом были бы дважды учтены дни, в которые мы оба опаздываем. Чтобы это компенсировать, мы должны вычесть вероятность того, что мы оба опоздаем на работу после ночного

обсуждения книги, которая составляет $P(А, Д) = 2\%$. В итоге мы имеем вероятность опоздания 5% для Алекса, 10% для Джордана, минус 2%, когда опаздывают оба: $5 + 10 – 2 = 13$ и $13/100 = 13\%$.

Отталкиваясь от этого, мы можем сформулировать правило сложения вероятностей для случая, когда наступает одно или другое событие: $P(А \text{ или } Д) = P(А) + P(Д) – P(А, Д) = 5\% + 10\% – 2\% = 13\%$.

Помните о пересечении

При вычислении вероятности наступления множества событий некоторые испытывают сложности с вычитанием пересекающейся области. Однако делать это необходимо, поскольку вероятность никогда не может превышать 1. Давайте снова обратимся к простому примеру с бросанием кубика. Вероятность выпадения числа, большего 2, равна 4/6. Вероятность выпадения нечетного числа составляет 3/6. Если вы хотите узнать вероятность наступления того или другого события, вы не можете просто сложить 4/6 + 3/6, потому что в итоге у вас получится 7/6 = 1 ⅙, то есть число, превышающее единицу, что нарушает вышеупомянутое правило. Мы должны вычесть пересекающуюся область, включающую все случаи, когда при бросании кубика выпадает число, большее 2 и являющееся нечетным, то есть числа 3 и 5, вероятность выпадения которых составляет 2/6.

Постановка задачи: $P(К > 2 \text{ или } К \text{ нечетное}) =$
Правило сложения: $P(К > 2) + P(К \text{ нечетное}) – P(К > 2, К \text{ нечетное}) =$
Вероятности: $4/6 + 3/6 – 2/6$
Ответ: $5/6$

Выпадение числа 2 — это единственный случай, который не удовлетворяет ни одному из условий.

Вы наверняка уже устали от различных нотаций, игральных костей, монеток и опаздывающих на работу авторов. Чтобы вы могли отдохнуть от всего этого, мы предлагаем вам выполнить следующее мысленное упражнение.

МЫСЛЕННОЕ УПРАЖНЕНИЕ НА ОПРЕДЕЛЕНИЕ ВЕРОЯТНОСТИ

Сэм — замкнутый, но очень способный. Ему 29 лет. Он получил высшее экономическое образование в родной Калифорнии. В студенчестве он был одержим данными, работал волонтером в университетском статистическом консультационном центре и самостоятельно учился программировать на языке Python.

Что из этого более вероятно?

1. Сэм живет в Огайо.
2. Сэм живет в Огайо и работает дата-сайентистом.

Правильный ответ — № 1, хотя в описании нет и намека на то, что Сэм может жить в Огайо, не являясь при этом дата-сайентистом. Это вариация популярной задачи про Линду из книги «Думай медленно… решай быстро»[49], с которой у большинства людей возникают сложности. А какой ответ выбрали вы?

Ответ № 2? Возможно, потому, что мы рассказали вам о том, что Сэм занимался программированием и мог быть дата-сайентистом. Ответ № 2 кажется более вероятным именно потому, что в нем упоминается событие, связанное с прошлым Сэма. Однако он все же менее вероятен, чем ответ № 1. И вот почему.

В данном примере отсутствуют обозначения и числа, но он по-прежнему отражает важный урок из предыдущего раздела. Вероятность одновременного наступления любых двух событий не может превышать вероятность наступления каждого из них в отдельности. Чем больше «и» вы добавляете в то или иное утверждение, тем меньше будет итоговая вероятность. Для того чтобы Сэм был дата-сайентистом и жил в Огайо, он должен для начала просто жить в Огайо. Например, он мог бы жить в Огайо и работать актуарием.

Помните, что вероятность одновременного наступления двух событий определяется правилом умножения. Вероятность того, что Сэм живет в Огайо и работает дата-сайентистом (Д), можно обозначить как $P(О, Д) = P(О) \times P(Д \mid О)$. А поскольку вероятность никогда не превышает единицу, умножение $P(О)$ — вероятности того, что Сэм живет в Огайо — на любую

[49] «Думай медленно… решай быстро», Даниэль Канеман (Издательство: АСТ, 2014).

другую величину вероятности не может увеличить результирующее значение $P(О) \times P(Д \mid О)$. Таким образом, $P(О, Д)$ ни при каких условиях не может превысить $P(О)$, каким бы предпочтительным ни казался ответ № 2.

Все еще сложно? Вы могли прочитать ответ № 2 как условную вероятность: какова вероятность того, что Сэм живет в Огайо при условии, что он работает дата-сайентистом, $P(О \mid Д)$? Вероятность этого может превышать вероятность проживания Сэма в Огайо, $P(О)$. Однако в данном случае разница между «и» и «при условии» имеет большое значение.

Рассмотрим более простой пример. Бейсбольная команда «Нью-Йорк Янкис» имеет преданных поклонников по всему миру. Предположим, что прямо сейчас проходит матч, который смотрят миллионы людей как вживую на стадионе, так и по телевизору. Теперь случайным образом выберите одного жителя планеты. Учитывая, что в мире живут миллиарды людей, крайне маловероятно, что вы выберете фаната «Янкис». Еще менее вероятен выбор фаната «Янкис», смотрящего игру на стадионе, потому что не все фанаты могут там присутствовать. Однако если бы у вас была возможность случайным образом выбрать человека, присутствующего на стадионе, все было бы иначе. Весьма вероятно, что он оказался бы фанатом «Янкис»[50].

Таким образом, вероятность того, что тот или иной человек — фанат «Янкис» и присутствует на игре, сильно отличается от вероятности того, что человек является фанатом «Янкис» при условии, что он присутствует на игре.

Дальнейшие шаги

После выполнения этого мысленного упражнения имеет смысл вспомнить о предупреждении, которое было сделано в начале этой главы: будьте внимательны и помните о том, что ваша интуиция может сыграть с вами злую шутку. Вероятности регулярно будут запутывать и сбивать вас с толку. Возможно, лучшее, что мы можем сделать для борьбы с этой проблемой, — это узнать о самых распространенных ловушках.

Теперь, когда вы познакомились с обозначениями и правилами теории вероятностей, пришло время научиться осознавать и критически осмыслять вероятности, с которыми вам предстоит столкнуться в ходе своей работы. Вот несколько советов, которые помогут вам не сбиться с пути:

---

[50] Эта вероятность не была бы равна 100%, потому что у команды противника тоже есть болельщики.

— Будьте осторожны, делая предположения о независимости событий.
— Знайте, что все вероятности условны.
— Убедитесь в том, что вероятности имеют смысл.

БУДЬТЕ ОСТОРОЖНЫ, ДЕЛАЯ ПРЕДПОЛОЖЕНИЯ О НЕЗАВИСИМОСТИ СОБЫТИЙ

Если события не зависят друг от друга, вы можете перемножить вероятности их наступления. Например, вероятность выпадения двух орлов подряд при подбрасывании честной монеты составляет $P(\text{О}) \times P(\text{О}) = 1/2 \times 1/2 = 1/4$. Однако не все события являются независимыми, поэтому с осторожностью делайте соответствующее предположение при вычислении или анализе вероятностей.

Мы уже упоминали об этом в начале книги в связи с ипотечным кризисом 2008 года. Вероятность того, что человек перестанет платить ипотеку, не является независимой от вероятности того, что его сосед тоже перестанет ее платить, хотя финансисты с Уолл-стрит на протяжении многих лет думали иначе. И то и другое событие неразрывно связано с общим состоянием экономики и мира в целом.

Тем не менее допущение независимости событий, которые таковыми не являются, — весьма распространенная ошибка. Руководство вашей компании может допустить ее при принятии стратегических решений — и, как следствие, сильно недооценить вероятность одновременного наступления нескольких событий.

Представьте заседание совета директоров. Обсуждается вероятность того, что в будущем году компании удастся реализовать три интересных, но рискованных проекта: *A*, *B*, *C*. Осознавая потенциальные риски, руководители компании оценивают вероятность неудачи для каждого проекта как $P(\text{провала } A) = 50\%$, $P(\text{провала } B) = 25\%$, а $P(\text{провала } C) = 10\%$.

Кто-то берет калькулятор и перемножает вероятности: $50\% \times 25\% \times 10\% = 1{,}25\%$. Руководители в восторге: вероятность того, что все три проекта потерпят неудачу, составляет всего 1,25%. В конце концов, ставки высоки, так что всего один успешный проект способен окупить инвестиции, сделанные во все три. А поскольку суммарная вероятность должна быть равна 1, вероятность успеха хотя бы одного проекта составляет 1 минус вероятность провала всех проектов, или $1 - 0{,}0125 = 0{,}9875 = 98{,}75\%$. «Ничего себе, — думают они, — вероятность общего успеха составляет почти 99%!»

Увы, их расчеты неверны. Все три события зависят от общего успеха компании, который может быть подорван такими факторами, как корпоративный скандал, плохие квартальные результаты или какое-то более крупное событие, влияющее на мировую экономику, вроде пандемии COVID-19. События *A*, *B* и *C* зависят от нескольких факторов. Поэтому, когда руководители необоснованно допускают их независимость, они недооценивают вероятность того, что все три проекта потерпят неудачу в будущем году, а значит, переоценивают шансы на то, что по крайней мере один из них окажется успешным.

Если это кажется вам неважным, вспомните финансовый кризис 2008 года и последующую рецессию.

Не допускайте ошибку игрока

С другой стороны, некоторые события являются независимыми, но не воспринимаются таковыми. Это порождает другой вид риска, благодаря которому процветают казино. В данном случае люди переоценивают вероятность наступления того или иного события, основываясь на предшествующих событиях.

Если при подбрасывании честной монеты 10 раз подряд выпадет орел, то вероятность выпадения орла в результате следующего броска все равно будет составлять $P(\text{О}) = 50\%$. В случае с независимыми событиями вероятность наступления одного из них не увеличивается и не уменьшается в зависимости от предыдущих результатов. Однако игроки ошибочно полагают, что величина вероятности меняется — отсюда и название «ошибка игрока»[51].

Каждый последующий бросок кубика не зависит от результата предыдущего броска. То же самое касается игровых автоматов и рулетки. Тем не менее игроки пытаются отыскать закономерности в этих событиях. Они либо думают, что на игровом автомате «должен» выпасть выигрыш, потому что он уже давно не выбрасывал монеты, либо считают, что «горячие» игральные кости позволят им выигрывать и впредь.

Однако каждое последующее событие имеет ту же вероятность выигрыша, что и предыдущее. А поскольку речь идет о казино, то шансы не в вашу пользу. Однако, заметив последовательность редких событий, любители азартных игр делают большие ставки, думая, что настал их счастливый день.

---

[51] Вера в то, что прошлые независимые события могут произойти по прошествии достаточного количества времени, также известна как «закон средних чисел» — наукообразный термин, обозначающий склонность принимать желаемое за действительное.

О, как же они ошибаются. Правда, казино может угостить их «бесплатным» завтраком[52].

ВСЕ ВЕРОЯТНОСТИ ЯВЛЯЮТСЯ УСЛОВНЫМИ

Все вероятности в некотором смысле условны. Вероятность выпадения орла при подбрасывании монеты $P(\text{О})$ равна 50% при условии, что монета является честной. То же самое касается вероятности выпадения единицы при бросании кубика: $P(\text{К} == 1) = 1/6$. Вероятность успеха проекта по работе с данными зависит от коллективного разума группы аналитиков, правильности данных, сложности проблемы, отсутствия вирусов на компьютерах, риска закрытия компании из-за пандемии и так далее.

Также подумайте о том, как компании и люди оценивают успех и компетентность. Обычно это делается исходя из прошлых успехов. Компании нанимают консультанта с успешным послужным списком или адвоката, который выигрывает больше всего дел, а человек обращается к кардиохирургу, чьи пациенты умирают в ходе операции реже всего. Допустим, консультант зарабатывает деньги для своих клиентов в 90% случаев, адвокат выигрывает 80% дел, дошедших до суда, а уровень смертности пациентов кардиохирурга составляет всего 2%.

Однако они могут влиять на эти вероятности. Консультант, юрист и хирург могут решить, браться за дело или нет. Они хорошо представляют свои шансы на успех, и если эти шансы кажутся им слишком небольшими, они могут отказаться. Вероятность успеха каждого из них зависит от выбора проектов с наибольшей вероятностью успеха и избегания тех, которые могут привести к ухудшению их показателей[53].

Вы должны учитывать все факторы, влияющие на степени вероятности, с которыми сталкиваетесь.

Не меняйте зависимости местами

Еще одна ловушка состоит в склонности предполагать то, что $P(A \mid B) = P(B \mid A)$ для двух событий A и B. Обратите внимание на то, как зависимости поменялись местами: в одном случае A зависит от B, в другом — B от A.

[52] Авторы книги ничего не имеют против таких завтраков.

[53] Мы не утверждаем, что консультанты или хирурги так поступают. Так делают только адвокаты.

Вот пример, показывающий разницу между двумя этими случаями. Пусть событие *A* будет «Проживанием в штате Нью-Йорк», а событие *B* — «Проживанием в городе Нью-Йорк». $P(A \mid B)$, то есть вероятность проживания в штате Нью-Йорк при условии, что вы живете в городе Нью-Йорк, сильно отличается от $P(B \mid A)$ — вероятности проживания в городе Нью-Йорк при условии, что вы живете в штате Нью-Йорк. В первом случае вероятность составляет 100%, $P(A \mid B) = 1$, а во втором — нет, поскольку около 60% жителей штата Нью-Йорк живут за пределами города Нью-Йорк.

В таком простом примере все довольно очевидно, однако перестановка зависимостей и предположение о том, что $P(A \mid B) = P(B \mid A)$ — настолько распространенная ошибка, что ей дали название и посвятили целую статью в Википедии — Confusion of the Inverse («ошибка приравнивания двух условных вероятностей»)[54]. Вы наверняка тоже допустили ее в процессе выполнения мысленного упражнения, предложенного в начале этой главы.

Давайте вернемся к сценарию из этого упражнения.

Ваша компания подверглась хакерской атаке, в результате которой 1% ноутбуков оказались заражены вирусом. Положительный результат теста на наличие вируса — это событие +, отрицательный результат — событие –, инфицирование вирусом — событие B. Вам была предоставлена следующая информация: $P(+ \mid \text{B}) = 99\%$, $P(- \mid \text{без B}) = 99\%$ и $P(\text{B}) = 1\%$. Другими словами, вероятность положительного результата теста при наличии вируса на ноутбуке составляет 99%, вероятность отрицательного результата теста при отсутствии вируса на ноутбуке составляет 99%, а вероятность наличия вируса на произвольно выбранном ноутбуке составляет 1%.

Мы хотели определить вероятность того, что компьютер заражен вирусом, при условии положительного результата теста, $P(\text{B} \mid +)$. Именно здесь возникла вышеописанная путаница. Речь шла о $P(\text{B} \mid +)$, а не о $P(+ \mid \text{B})$, однако многие люди при выполнении этого упражнения дают ответ, соответствующий $P(+ \mid \text{B}) = 99\%$.

Вероятности $P(\text{B} \mid +)$ и $P(+ \mid \text{B})$ не одинаковы, однако они связаны между собой теоремой Байеса — одной из самых известных теорем в теории вероятностей и статистике.

---

[54] Confusion of the Inverse: **en.wikipedia.org/wiki/Confusion_of_the_inverse**. Доступ получен 4 июля, 2020.

Теорема Байеса

Теорема Байеса, сформулированная в XVIII веке, — это способ работы с условными вероятностями, который применяется повсюду, начиная с планирования сражений и управления финансами и заканчивая расшифровкой ДНК[55]. Для двух событий *A* и *B* теорема Байеса утверждает следующее:

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

Пусть вас не пугает эта формула. Самое важное — не запомнить ту или иную формулу, а понять, что она делает и почему о ней стоит знать.

Теорема Байеса позволяет связать условную вероятность двух событий. Вероятность наступления события *A* при условии наступления события *B* связана с вероятностью наступления события *B* при условии наступления события *A*. Они не равны, но связаны приведенным выше уравнением.

Это может пригодиться, когда вам известна одна из условных вероятностей и вы хотите определить другую. Например:

— Медицинские исследователи хотят знать вероятность того, что у человека будет положительный результат скринингового теста на рак при условии, что этот человек болен раком, $P(+ \mid \text{Р})$. Тогда они смогут создать более точные тесты, позволяющие немедленно приступить к лечению. Разработчики политики хотят знать обратное — вероятность того, что человек болен раком при условии положительного результата скринингового теста, $P(\text{Р} \mid +)$, потому что они не хотят подвергать людей ненужному лечению на основании ложноположительного результата (когда тест показывает наличие болезни при ее отсутствии).
— Прокуроры хотят знать вероятность того, что подсудимый виновен при условии наличия доказательств, $P(\text{В} \mid \text{Д})$. Это зависит от вероятности обнаружения доказательств при условии, что человек виновен, $P(\text{Д} \mid \text{В})$.
— Ваш поставщик услуг электронной почты хочет знать вероятность того, что электронное письмо — спам при условии, что оно содержит

[55] С подробной историей данной теоремы можно ознакомиться в книге McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy* (American First ed.). Yale University Press.

фразу «Бесплатные деньги!», P(Спам | Деньги). Используя исторические данные, он может рассчитать вероятность того, что электронное письмо содержит фразу «Бесплатные деньги!» при условии, что оно является спамом, P(Деньги | Спам). (Мы более подробно разберем этот пример в главе 11.)

— В вышеописанном мысленном упражнении вы хотите узнать вероятность наличия вируса на вашем компьютере при условии положительного теста, P(В | +). Вам известно обратное — вероятность положительного результата теста при условии наличия вируса в компьютере P(+ | В).

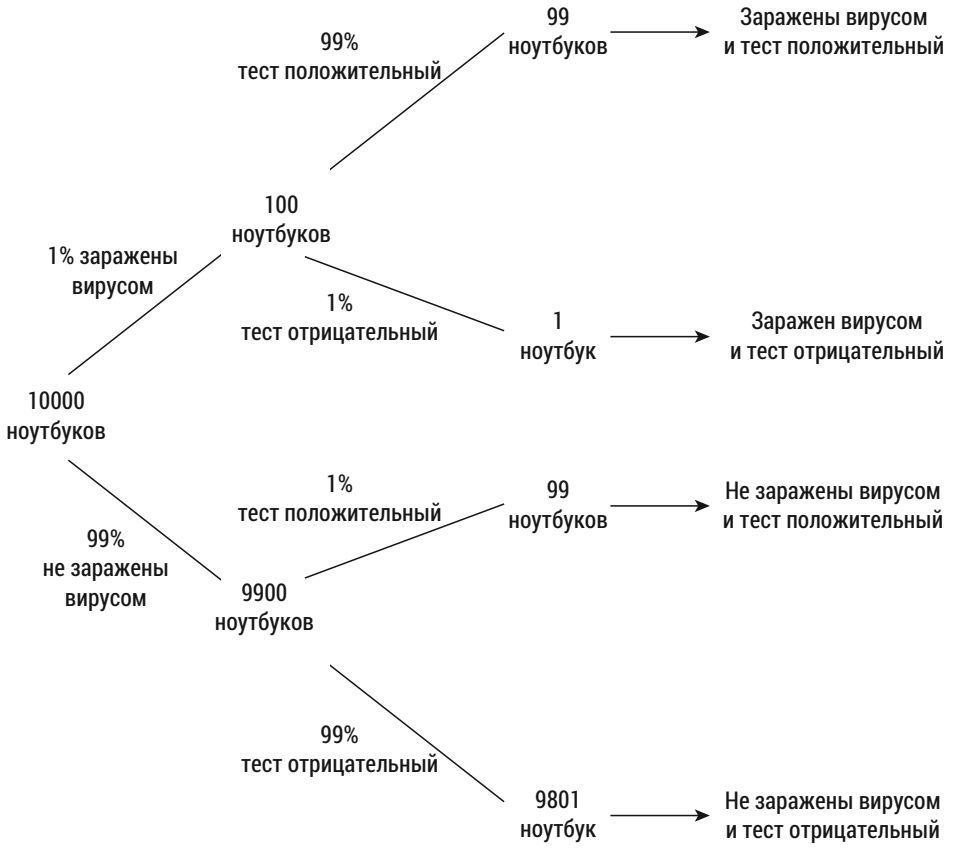


Рис. 6.2. Древовидная диаграмма для сканирования компьютеров в большой компании на наличие вируса

Все условные вероятности в этих примерах связаны теоремой Байеса. Это хорошая новость. Плохая новость — некоторые части этой теоремы трудно рассчитать. Дело в том, что не все вероятности легко выяснить. Например, вероятность того, что человек болен раком при условии положительного результата скринингового теста, может быть легче узнать, чем вероятность наличия этого заболевания у человека с отрицательным результатом теста.

Чтобы определить, достаточно ли у вас информации для применения теоремы Байеса, можно построить древовидную диаграмму (рис. 6.2). В качестве примера мы используем то же самое мысленное упражнение — и наконец покажем, почему правильный ответ составляет 50%. Предположим, что в компании 10 000 ноутбуков. Поскольку вероятность положительного результата теста при наличии вируса на ноутбуке составляет 99%, мы предполагаем, что при тестировании 1% ноутбуков с вирусом мы будем получать отрицательный результат теста, $P(- \mid \text{В}) = 1\%$. Точно так же, учитывая, что вероятность отрицательного результата теста при отсутствии вируса на ноутбуке составляет 99%, мы предполагаем, что при тестировании 1% ноутбуков без вируса мы будем получать положительный результат теста, $P(+ \mid \text{без В}) = 1\%$.

Как видно на рис. 6.2, исходя из имеющейся у нас информации, 10 000 ноутбуков можно разделить на четыре группы: ноутбуки с вирусом, давшие отрицательный или положительный результат при тестировании, и ноутбуки без вируса, давшие отрицательный или положительный результат при тестировании. Давайте разберемся, что это значит. Если вы посмотрите на древовидную диаграмму, то поймете, что нам интересны только две ветви. Первый случай предполагает наличие вируса и положительный результат тестирования — это 99 ноутбуков. Второй случай предполагает отсутствие вируса и опять же положительный результат тестирования — тоже 99 ноутбуков. Такой результат называется ложноположительным.

Дело вот в чем. Мы уже знаем, что результат тестирования компьютера оказался положительным. Это означает, что он может принадлежать только одной из этих двух групп. Вы не знаете, какой именно, но если представить ноутбуки в виде шариков одинакового размера, то при вытаскивании одного из них вслепую из мешка вероятность того, что вы достанете шарик из той или иной группы, будет составлять 50%.

А теперь давайте проверим свое (новое) интуитивное понимание ситуации математикой. Для этого используем теорему Байеса, заменив события A и B событиями В и +: $P(\text{В} \mid +) \times P(+) = P(+ \mid \text{В}) \times P(\text{В})$. Затем подставим известные нам величины вероятности:

$P(+)$ = вероятность положительного результата теста = 198 положительных результатов /10000 = 1,98%

$P(+ \mid \text{В}) = 99/100 = 99\%$

$P(\text{В}) = 100/10000 = 1\%$

Подставив эти значения в выражение $P(\text{В} \mid +) \times P(+) = P(+ \mid \text{В}) \times P(\text{В})$, получаем:

$P(\text{В} \mid +) \times 1{,}98\% = 99\% \times 1\%$,

$P(\text{В} \mid +) = (99\% \times 1\%)/1{,}98\%$

$P(\text{В} \mid +) = 50\%$

Математики многовато, но она позволила нам прийти к правильному ответу: вероятность наличия вируса на вашем ноутбуке при положительном результате теста составляет 50%.

УБЕДИТЕСЬ, ЧТО ВЕРОЯТНОСТИ ИМЕЮТ СМЫСЛ

В этой главе вам пришлось иметь дело с множеством чисел и обозначений, особенно в предыдущем разделе. Теперь давайте сделаем шаг назад и поговорим о способах осмысления и использования вероятностей.

Калибровка

Определяемые вероятности должны иметь смысл.

Например, при условии равных затрат и выгод проект с вероятностью успеха 60% сопряжен с большим риском, чем проект с вероятностью успеха 75%.

Мы знаем, что это кажется очевидным, однако люди часто оценивают события с вероятностью 60% или 75% как весьма вероятные, потому что их вероятность превышает 50%. Но если бы это было так, вероятности не имели бы никакого значения и сводились бы к бинарным решениям типа «да/нет», при которых полностью утрачивается смысл статистического мышления и работы с неопределенностью.

Более того, если вероятность события составляет 75%, оно должно происходить примерно в 75% случаев[56]. Это кажущееся очевидным утверждение,

[56] Мы говорим «примерно», потому что во всем есть вариации. Но в долгосрочной перспективе событие, имеющее вероятность 75%, должно происходить в 75% случаев.

называемое калибровкой, придает вероятности смысл. «Калибровка определяет соответствие фактической частоты наступления тех или иных событий в долгосрочной перспективе вашему прогнозу»[57].

Плохая калибровка делает невозможной точную оценку риска. Если вы самоуверенный юрист, который думает, что выиграет дело с вероятностью 90%, хотя до этого выигрывал только в 60% случаев, вы переоцениваете свои шансы на успех. Это пример плохой калибровки.

Итак, вероятности должны иметь смысл. Помните о том, что редкие события не являются невозможными, а высоковероятные события не обязательно наступают.

Редкие события могут случаться и случаются

Редкое событие может не произойти с вами или с кем-либо из ваших знакомых, но это не значит, что оно не произойдет вообще. Тем не менее у нас часто возникают сложности с пониманием редких событий.

Это правда: вы вряд ли сорвете джекпот в лотерею, однако некоторые люди в нее все-таки выигрывают. Если учесть количество лотерей, проводимых по всему миру каждый день, вероятность того, что столь редкое событие произойдет с кем-то из жителей планеты, пусть даже не с вами, оказывается не такой уж и низкой.

Мы часто забываем об огромном количестве людей, живущих на Земле. При населении в несколько миллиардов человек события типа «1 на миллион» кажутся гораздо более вероятными. На самом деле, участниками таких событий становится гораздо больше людей, чем мы можем себе представить. В мире, где проживает 7,8 миллиарда человек, событие типа «1 на миллион» может происходить ежедневно с участием 7800 человек.

С другой стороны, то или иное событие очень легко представить чрезвычайно редким, чтобы придать ему драматизма (а возможно, даже ввести в заблуждение). Например, в американском футболе очень часто встречаются комментарии, намекающие на редкость происходящего на экране события. «Это первый раз, когда 28-летний новичок пробежал 30 ярдов после двух выездных игр и всего одной игры в предсезонке». Если сформулировать это так, то данное событие действительно может показаться редким.

---

[57] **fivethirtyeight.com/features/when-we-say-70-percent-it-really-means-70-percent**

Не перемножайте вероятности без необходимости

Не перемножайте вероятности прошлых событий без особой необходимости. В противном случае вы можете сделать то или иное событие практически невероятным.

Давайте прикинем вероятность того, что вы читаете именно эту строку на этой странице этой самой книги. Помимо данной строки на этой странице еще примерно 35 строк (1/35), в книге — еще 300 страниц (1/300), а в мире — миллионы книг. Если вы перемножите эти вероятности, то получите бесконечно малое число. Очевидно, мы были созданы друг для друга!

ПОДВЕДЕНИЕ ИТОГОВ

Эта глава должна была научить вас не только основам теории вероятностей, но и смирению. Вероятности — это сложная тема. Однако важное условие успешного изучения новой темы — осознание того, что что-то может пойти не так. Информация, которую вы узнали из этой главы, поможет вам найти дополнительные сведения, прежде чем принимать решения относительно вероятности, особенно в тех случаях, которые на первый взгляд кажутся интуитивно понятными.

В этой главе мы показали, как легко можно ошибиться при определении вероятностей. Иногда ошибка заключается в самой формулировке вопроса, а иногда — в предположениях, основанных на предоставляемой информации. Чтобы избежать недоразумений, пользуйтесь нашими рекомендациями при анализе вероятностей:

— Будьте осторожны, делая предположения о независимости событий.
— Знайте, что все вероятности являются условными.
— Убедитесь в том, что вероятности имеют смысл.

ГЛАВА 7

Бросайте вызов статистике

Кент Брокман: Мистер Симпсон, как вы ответите на обвинения в том, что мелкого вандализма вроде граффити стало меньше на 80%, в то время как количество случаев избиения тяжелыми мешками выросло на шокирующие 900%?
Гомер: О, люди могут придумать любую статистику, чтобы доказать что угодно, Кент. Сорок процентов людей знают это.

_ *Мультсериал «Симпсоны»*

Вы когда-нибудь сталкивались со статистическим утверждением в новостях или на рабочем месте, которое вы хотели бы понять, оценить и, возможно, даже подвергнуть сомнению? Данная глава научит вас именно этому. В ней мы поговорим о статистическом выводе, о том, как пользоваться индуктивной статистикой и оспаривать ее результаты, а также перечислим вопросы, которые вам следует задать для полного понимания сделанных выводов.

КРАТКИЕ УРОКИ ПО СТАТИСТИЧЕСКОМУ ВЫВОДУ

Как было сказано в главе 3 «Готовьтесь мыслить статистически», индуктивная статистика позволяет нам собирать данные о мире, в котором мы живем, и делать на их основании предположения об этом мире.

В данном разделе мы разберем ряд примеров, чтобы показать, насколько интуитивным может быть процесс построения статистического вывода при постепенном введении формальных статистических терминов (часть из них вы узнали ранее в книге, но напоминание никогда не бывает лишним).

Хорошая новость — вы без проблем сможете проследить представленную далее логику статистического вывода вне зависимости от ваших знаний в области статистики.

Обеспечьте себе простор для маневра

Один из самых распространенных и важных примеров применения индуктивной статистики — проведение опросов. Вы не можете опросить всех — только участников выборки, к которым у вас есть доступ. С ее помощью мы пытаемся лучше понять окружающий мир. Иначе говоря, эта выборка помогает нам больше узнать о популяции.

Рассмотрим пример опроса. Случайной выборке, состоящей из 1000 студентов вводных курсов по статистике, проводящихся по всей стране, задают вопрос: «Вам уже надоело то, что статистики используют примеры опросов для объяснения основных статистических концепций?»

Результаты данного опроса таковы: 655 студентов сказали «да». (А как бы проголосовали вы?)

Стали бы вы, основываясь лишь на одной выборке из 1000 студентов, заявлять о том, что истинный процент всех студентов вводных курсов по статистике (популяция), которым надоели примеры с опросами, составляет ровно 65,5? Или вы хотите иметь некоторое пространство для маневра, делая свое предположение?

Скорее всего, второе. Это хорошо, потому что неделю спустя при опросе еще 1000 студентов утвердительный ответ дали 670 человек. Разумеется, 655 и 670 — это весьма близкие значения, и, возможно, вы полагаете, что проведение этих опросов позволило вам приблизиться к истинной доле студентов, готовых утвердительно ответить на задаваемый вопрос. Однако, если бы вы провели этот опрос еще раз, вы получили бы разные ответы вследствие вариации выборки. И с этим ничего нельзя сделать, кроме как представить полученные результаты в контексте. Опросные агентства понимают это и указывают «погрешность» результатов опроса в пределах +/– 3%, которая отражает неопределенность, обусловленную вариацией и случайностью.

В случае с первым опросом значение 65,5% — точечная оценка, и мы могли бы представить результаты как 65,5% +/– 3%, или (62,5%, 68,5%). Интервал (62,5%, 68,5%) называется доверительным и является примером индуктивной статистики. Он позволяет получить некоторые сведения об окружающем мире на основе информации, предоставленной выборкой.

Мы надеемся, что этот доверительный интервал отражает истинный процент всех студентов вводных курсов по статистике, которые устали от примеров с опросами.

Мораль: при использовании выборок наблюдаются вариации, делающие неопределенной вашу оценку количества студентов вводных курсов по статистике, уставших от примеров с опросами. К счастью, доверительные интервалы определяют диапазон правдоподобных значений, в которых может лежать их истинная доля, то есть дают вам некоторое пространство для маневра.

Больше данных — больше доказательств

Если в процессе совершения покупок в Интернете вы видите продукт на сайте Amazon с рейтингом в 1 звезду, основанном на единственном обзоре, вы можете проигнорировать этот обзор — мнение одного человека. Однако если вы увидите продукт с низким рейтингом, основанным на сотнях отзывов (скажем, на 300), ваше мнение будет иным. Существует консенсус, согласно которому данный продукт является некачественным. Поэтому вы выбираете другой продукт — с рейтингом 4,9 звезды, основанном на 200 отзывах[58].

Это говорит о том, что вы уже понимаете, как количество точек данных, лежащих в основе рейтинга товара на Amazon, влияет на ваше доверие к нему. Размер выборки мы будем обозначать буквой N. Вы не испытываете доверия к рейтингу, основанному на одном обзоре ($N = 1$), но вас способны убедить рейтинги, основанные на выборках размером $N = 300$ и $N = 200$. Как вы уже догадались, размер выборки играет огромную роль в статистическом выводе. В самом деле, кажется маловероятным, хотя и не невозможным, чтобы продукт с рейтингом 4,9 звезды и $N = 200$ оказался полным хламом. А как насчет продукта с $N = 1$? Его обзор мог быть написан случайным интернет-троллем.

Мораль: размер выборки имеет значение. Больше данных — больше доказательств. (Мы же говорили, что это интуитивно понятно.)

Бросьте вызов статус-кво

По сути, наука и создание новых знаний предполагает бросание вызова статусу-кво. Когда накапливается достаточно доказательств в пользу того, что

[58] Не забудьте оставить отзыв о нашей книге на сайте Amazon.

прежний образ мышления ошибочен, мы его адаптируем. Это же верно и для статистического вывода.

Простейшая аналогия — презумпция невиновности в американской системе уголовного права. Обвиняемые «невиновны до тех пор, пока их вина не будет доказана» (статус-кво). Подсудимый объявляется «виновным» лишь тогда, когда доказательства вне всяких обоснованных сомнений указывают на ошибочность статус-кво. Бремя доказывания того, что первоначальное предположение о невиновности подсудимого неверно, возлагается на сторону обвинения.

Табл. 7.1. Вопросы, нулевые гипотезы (H_0) и альтернативные гипотезы (H_a)

Вопрос	Нулевая гипотеза (H_0)	Альтернативная гипотеза (H_a)
Изменился ли уровень удовлетворенности клиентов MegaCorp за последний квартал?	Уровень удовлетворенности клиентов MegaCorp не изменился за последний квартал.	Уровень удовлетворенности клиентов MegaCorp изменился за последний квартал.
Способствовала ли реклама MegaBank во время «Суперкубка» увеличению годовой прибыли?	Реклама MegaBank во время «Суперкубка» не изменила годовую прибыль.	Реклама MegaBank во время «Суперкубка» изменила годовую прибыль.
Защищает ли экспериментальная вакцина от вируса COVID-19?	Экспериментальная вакцина не лучше плацебо.	Экспериментальная вакцина лучше плацебо.
Уровень безработицы в США изменился по сравнению с прошлым месяцем?	Уровень безработицы в США не изменился с прошлого месяца.	Уровень безработицы в США изменился с прошлого месяца.

Исследователи, ученые и компании используют эту логику для создания новых знаний, направленных на улучшение общества или бизнеса. Вот как это работает. Они начинают с постановки вопроса[59], подобного тем, которые перечислены в табл. 7.1, и используют его для так называемой проверки гипотезы.

Статус-кво называется нулевой гипотезой, которая обычно обозначается как H_0. Как правило, ее выбирают в надежде впоследствии отбросить в пользу

[59] Как вы помните из главы 1, проект по работе с данными должен начинаться с формулирования четкого вопроса.

нового знания, называемого альтернативной гипотезой, обозначаемой как H_a. Разумеется, нулевая и альтернативная гипотеза зависят от заданного вопроса. В табл. 7.1 показано, как общие вопросы могут быть преобразованы в соответствующие гипотезы. Исследователи стремятся найти доказательства, позволяющие отвергнуть нулевую гипотезу в пользу альтернативной.

Обратите особое внимание на логику проверки гипотез, представленную в табл. 7.1. Какой бы правдоподобной ни казалась гипотеза, изначально вы предполагаете, что она неверна (то есть отталкиваетесь от статус-кво). При наличии достаточного количества доказательств, говорящих о том, что нулевая гипотеза (H_0) очень маловероятна, вы отклоняете ее в пользу альтернативной (H_a).

Мораль: проверка гипотезы — отличительная черта научных экспериментов. Чтобы бросить вызов статус-кво, допустите его истинность в рамках нулевой гипотезы. При наличии достаточного количества доказательств (данных), говорящих о том, что нулевая гипотеза маловероятна, отклоните ее в пользу нового знания, содержащегося в альтернативной гипотезе.

Доказательства обратного

Предположим, вы играете в баскетбол с коллегами, и стажер просится в вашу команду, заявляя о том, что он попадает минимум в 50% случаев. «Потрясающе», — думаете вы. Вашей команде нужен хороший бомбардир[60].

Перед игрой вы мысленно отмечаете (то есть формулируете нулевую гипотезу): процент реализации бросков стажера ≥ 50%.

Игра начинается, и вы передаете ему мяч для выполнения открытого броска. Промах. «Ничего страшного», — думаете вы. Но затем он не попадает снова. Потом промахивается еще раз. И… еще. Четыре промаха подряд. Ну и ну. Это просто ужасно.

Ваша вера в него начинает колебаться. Этот парень действительно умеет играть или просто дурачится? Тем не менее даже у профессионалов бывают неудачные дни, и иногда они промахиваются четыре раза подряд. И вы продолжаете давать ему новые шансы. А он продолжает промахиваться. За всю

[60] Мы понимаем, что 50% — это отличный процент реализации бросков в баскетболе. У Леброна Джеймса, например, этот показатель за всю карьеру составляет 50%. Так что нет, ваш стажер, скорее всего, не играет настолько хорошо, просто значение 50% облегчает расчеты. Однако хорошо, что вы, как главный по данным, задумались о том, не слишком ли это оптимистично.

игру стажер промахнулся 10 раз подряд, и ваша команда проиграла. Вы разочарованы и считаете этого парня лжецом.

Вы возвращаетесь за свой стол и решаете количественно оценить то жалкое выступление, свидетелем которого вы только что стали.

Итак, какова вероятность того, что игрок, реализующий 50% своих бросков, промахнется 10 раз подряд?

Отталкиваясь от базовой вероятности, вы выполняете некоторые расчеты. Вероятность того, что он промахнется один раз, составляет 50%. Вероятность двух промахов подряд составляет 50% × 50% = 25% (при условии, что результаты бросков не зависят друг от друга, как говорилось в предыдущей главе). Продолжая эту логику, вы умножаете показатель 50% сам на себя 10 раз: 0,5 ^ 10 = 0,00098, то есть 0,1%, или примерно 1 из 1000.

Таким образом, вероятность данного конкретного результата, то есть 10 промахов подряд, при условии, что стажер, по его словам, способен реализовать 50% бросков, составляет 1 из 1000.

Эта вероятность, равная 1 из 1000 или 0,001, называется *p*-значением (*p* означает probability — «вероятность»). Теперь вы должны решить, был ли у стажера просто неудачный день или ваша нулевая гипотеза, согласно которой процент реализации бросков стажера составляет 50%, ошибочна?

Десять пропущенных бросков лишь подрывают доверие. Однако то, что вероятность неудачного дня составляет 1 из 1000, довольно убедительно доказывает то, что первоначальное утверждение стажера вряд ли было истинным. Скорее всего, вы отвергли нулевую гипотезу на более ранних этапах игры в пользу альтернативной гипотезы, H_a: процент реализации бросков стажера < 50%.

Остановитесь на мгновение и спросите себя: когда вы начали сомневаться в способностях стажера вместо того, чтобы оправдывать его? Каким было пороговое число промахов, заставившее вас отвергнуть нулевую гипотезу?

Для примера предположим, что это пороговое значение составляло 5 промахов. Если бы стажер промахнулся только 4 раза подряд, вероятность чего составляет 50% × 50% × 50% × 50%[61] = 6,25%, или 1 из 16, вы бы еще могли продолжать верить в то, что он хороший бомбардир. Однако после пятого промаха доказательств обратного стало слишком много. Этот порог в 5 промахов подряд называется уровнем значимости, после превышения которого полученные данные больше не соответствуют исходному утверждению.

---

[61] О'Нил Кэти, Шатт Рэйчел. «*Data Science*. Инсайдерская информация для новичков» (Издательство: Питер, 2019).

Поскольку Вселенная полна вариаций, вы должны смириться с некоторым уровнем случайности (и количеством промахов). Иногда человек может плохо играть без всяких причин. Таким образом, уровень значимости — это некий условный установленный вами предел, до которого вы можете мириться со случайностью и необъяснимыми вариациями, продолжая считать нулевую гипотезу верной. Если *p*-значение меньше уровня значимости, вы отбрасываете нулевую гипотезу и говорите, что результат статистически значим.

Урок: проверка того, что *p*-значение не превышает уровня значимости, с целью отбрасывания нулевой гипотезы — ключевая часть процесса построения статистического вывода. Разумеется, наличие вариаций и произвольный выбор уровня значимости чреваты ошибками при принятии решений.

Сбалансируйте ошибки, допускаемые при принятии решений

Когда вариация приводит к неправильному выводу, это называется ошибкой при принятии решения.

Существуют два типа подобных ошибок, названия которых мало о чем говорят: ошибка первого рода (ложноположительное заключение) и ошибка второго рода (ложноотрицательное заключение). Поскольку описательность названия имеет большое значение, мы предпочитаем называть ошибки первого и второго рода именно ложноположительными и ложноотрицательными заключениями.

Ложноположительное заключение возникает тогда, когда доказательства подтверждают альтернативную гипотезу, которую следовало бы отвергнуть (например, у мужчины оказывается положительный тест на беременность). С другой стороны, ложноотрицательное заключение имеет место тогда, когда вы принимаете ложную нулевую гипотезу (например, у беременной женщины оказывается отрицательный тест на беременность). В табл. 7.2 приведены дополнительные примеры ошибок первого и второго родов.

Вы как лицо, принимающее решения, выбираете вероятность ложноположительного заключения, устанавливая уровень значимости. Со статистической значимостью тесно связано такое понятие, как мощность — вероятность отклонения нулевой гипотезы, когда альтернативная гипотеза верна. Чем выше мощность теста, тем ниже вероятность ложноотрицательного заключения.

Табл. 7.2. Ложноположительные и ложноотрицательные заключения при принятии решения

Вопрос	Нулевая гипотеза	Ложноположительное заключение	Ложноотрицательное заключение
Подсудимый совершил преступление?	Подсудимый невиновен.	Заключение невиновного человека в тюрьму.	Виновный человек остается на свободе.
Вы больны?	Вы не больны.	У вас положительный результат теста, хотя вы не больны.	Вы больны, но тест это не выявил.
Изменился ли показатель удовлетворенности клиентов MegaCorp в прошлом квартале?	Уровень рекомендаций в этом квартале ≤ уровня рекомендаций в прошлом квартале.	Улучшение результатов в этом квартале является случайным.	Результаты в этом квартале действительно улучшились, но тест этого не выявил.

Балансирование ошибок первого и второго родов предполагает компромисс, и, если вы не соберете больше данных, то не сможете уменьшить вероятность одного, не увеличив вероятность другого. Например, вы хотите обеспечить низкий уровень ложноположительных заключений в случае спама. Нулевая гипотеза заключается в том, что «электронное письмо не является спамом». В связи с этим ложноположительное заключение может привести к тому, что электронное письмо от вашей матери окажется в папке со спамом. Обратная сторона этого — большее количество спама в вашем почтовом ящике (больше ложноотрицательных заключений), но вы готовы мириться с этим ради того, чтобы получать большую часть своей личной электронной почты. Однако в случае скрининга заболеваний медицинское сообщество может допустить больше ложноположительных заключений, чтобы уменьшить количество ложноотрицательных (пропущенный диагноз). Если у кого-то есть заболевание, медики хотят его обнаружить.

Мораль: вариации усложняют процесс принятия решений. Иногда вам будет казаться, что ваша альтернативная гипотеза верна, хотя это не так (ложноположительное заключение), а иногда будете ошибочно думать, что верна нулевая гипотеза (ложноотрицательное заключение).

ПРОЦЕСС ПОСТРОЕНИЯ СТАТИСТИЧЕСКОГО ВЫВОДА

В предыдущих пяти кратких уроках мы рассмотрели несколько компонентов процесса статистического вывода. Пришло время понять, как эти компоненты сочетаются друг с другом. Давайте попробуем обобщить их, чтобы вы как главный по данным могли понять и четко объяснить весь процесс построения статистического вывода.

Если вкратце, то в ходе этого процесса вы должны выполнить следующие действия:

1. Задайте осмысленный вопрос.
2. Сформулируйте гипотезы для проверки, используя статус-кво в качестве нулевой гипотезы, а свое предположение — в качестве альтернативной.
3. Задайте уровень значимости. (Чаще всего используется произвольное значение в 5% или 0,05.)
4. Вычислите *p*-значение на основе результата статистического теста.
5. Вычислите соответствующие доверительные интервалы.
6. Отклоните нулевую гипотезу в пользу альтернативной, если *p*-значение оказалось меньше уровня значимости; в противном случае не отклоняйте нулевую гипотезу.

Остановитесь на мгновение и подумайте о перечисленных выше шагах. Если вы можете прочитать и понять все шесть шагов — поздравляем! Вы делаете успехи в изучении языка статистики. Единственное, что мы до этого упускали из виду, — это идея статистического теста, механизма вычисления *p*-значения. Мы использовали его при определении базовой вероятности в примере со стажером-баскетболистом (возведя 50% в 10-ю степень). Однако существуют сотни статистических тестов, используемых для описания, сравнения, оценки рисков и взаимосвязей в данных. Именно этим инструментам уделяется основное внимание в учебниках по статистике. Мы не стали сосредоточиваться на статистических тестах здесь, поскольку вы можете и должны понимать логику, лежащую в основе статистики, независимо от метода расчета *p*-значения.

Возвращаясь к поставленной задаче, мы признаем, что главные по данным чаще всего будут потребителями статистических результатов, а не их создателями. Поэтому в следующем разделе мы перечислим вопросы, которые

вам следует задать, чтобы бросить вызов тем статистическим показателям, с которыми вы сталкиваетесь. Если вы хорошо усвоили материал, изложенный в предыдущих разделах, вы уже должны быть готовы задавать эти вопросы.

ВОПРОСЫ, ПОЗВОЛЯЮЩИЕ БРОСИТЬ ВЫЗОВ СТАТИСТИЧЕСКИМ ПОКАЗАТЕЛЯМ

Мы составили список вопросов, которые вы можете задать своим товарищам по команде с целью критической оценки представленных статистических показателей:

— Каков контекст этой статистики?
— Каков размер выборки?
— Что вы тестируете?
— Какова нулевая гипотеза?
— Каков уровень значимости?
— Сколько тестов вы проводите?
— Каковы доверительные интервалы?
— Имеет ли это практическое значение?
— Предполагаете ли вы наличие причинно-следственной связи?

Давайте рассмотрим каждый из этих вопросов и разберемся в том, почему они важны.

Каков контекст этой статистики?

Контекст статистики не менее важен, чем сами показатели. Услышав фразу: «Продажи выросли на 10%!» — вы должны спросить: «По сравнению с чем?»

Рассмотрим следующий пример. Маркетолог-аналитик сообщает своему начальнику о том, что продажи выросли на 10% по сравнению с прошлым кварталом, но не говорит о том, что объем продаж его крупнейшего конкурента увеличился на 15%. Начальник наверняка предпочел бы знать этот дополнительный контекст. Однако попытки обобщить информацию могут привести к путанице. Главные по данным должны выяснять контекст и базовые показатели для проведения сравнения.

Рассмотрим другой пример. Предположим, новая реклама на YouTube повышает вероятность клика по объявлению на 50%. Без знания контекста это звучит весьма впечатляюще. Однако если рассматривать данный статистический показатель в контексте, становится ясно, что кликабельность рекламы (отношение числа людей, щелкнувших по объявлению, к числу людей, просмотревших рекламу) улучшился с 0,1 до 0,15% (то есть с 10 из 10 000 до 15 из 10 000) или на 0,05% в абсолютных величинах. Данный результат следует преподносить именно так. Указание относительного процентного изменения (0,0015–0,0001)/0,0001 × 100 = 50% создает неверное представление о нем.

Вероятно, в своей работе вы уже сталкивались с подобными примерами, когда вы видите точный, однозначный и впечатляющий статистический показатель, но не знаете, что он на самом деле означает. В таких случаях смело спрашивайте: «Каков контекст этой статистики?»

Каков размер выборки?

К этому моменту вы уже должны понимать важность размера выборки. Небольшое значение N, как правило, сопровождается большим количеством вариаций. Нет проблем: вы просто добавляете дополнительные данные. При достаточном количестве данных результаты будут менее вариативными, верно? В эпоху «больших данных» у вас может возникнуть соблазн просто сделать значение N настолько огромным, чтобы выборка учитывала все вероятности.

Однако в тех случаях, когда значение N очень велико, легко подумать, что N = ВСЕ, то есть в вашем распоряжении имеются все возможные точки данных. Однако подобное допущение не освобождает вас от необходимости задумываться о качестве данных и предвзятости. (Вспомните уроки из главы 4.) Действительно ли ваша выборка охватывает людей, относящихся к интересующей вас категории?

Как отмечается в книге «Data Science. Инсайдерская информация для новичков»:[62]

> Мы утверждаем, что предположение о том, что N=ВСЕ, — одна из самых больших проблем, с которыми мы сталкиваемся в эпоху больших данных. Прежде всего это способ исключения голосов людей, у которых нет

[62] О'Нил Кэти, Шатт Рэйчел. «Data Science. Инсайдерская информация для новичков» (Издательство: Питер, 2019).

времени, энергии или возможностей для участия во всех неформальных (возможно, даже необъявленных) выборах.

Исключение голосов относится не только к выборам. Нуждающиеся могут быть по ошибке лишены права на получение скидок на еду или одежду; на участие в опросах, касающихся государственной политики; или их голоса просто не будут учтены. Может показаться, что достаточно большой набор точно отражает характеристики популяции, однако размер выборки — это еще не все. Хуже того, в «больших данных» можно очень легко обнаружить ложные зависимости. Если препарировать данные определенным образом, в них всегда можно найти что-то интересное.

В тех редких случаях, когда N действительно равно ВСЕЙ популяции (перепись), можете считать, что вам повезло. Вам не придется заниматься построением статистического вывода, потому что в показателях описательной статистики не будет неопределенности при условии корректного сбора данных.

Что вы тестируете?

В основе любого статистического вывода, с которым вы сталкиваетесь на рабочем месте или в новостях, лежит (как мы надеемся) конкретный вопрос, который можно проверить с помощью данных. Не позволяйте специалисту по работе с данными предоставлять статистический показатель, не озвучивая при этом лежащий в его основе вопрос. Убедитесь в том, что ваша команда знает о причинах, по которым та или иная статистика вообще создается. Задайте вопрос: «Что вы тестируете?» — и попросите предоставить на него четкий ответ, сформулированный в нестатистических терминах[63].

Какова нулевая гипотеза?

В этом квартале ваш стажер в MegaCorp тесно сотрудничал с отделом обслуживания клиентов, предлагая идеи для повышения уровня их удовлетворенности. Вы хотите оценить эффективность его идей с помощью простого опроса клиентов MegaCorp, состоящего из единственного вопроса: «Вы бы порекомендовали нас другу?»

[63] Об уточнении самого вопроса мы говорили в главе 1.

Стажер формализует тест и выдвигает нулевую гипотезу: «Уровень рекомендаций в этом квартале не ниже, чем в прошлом». Таким образом:

— H_0: Уровень рекомендаций в этом квартале ≥ Уровню рекомендаций в прошлом квартале.

В случае отвержения нулевой гипотезы будет принята альтернативная гипотеза, которая в данном случае такова: «Уровень рекомендаций в этом квартале ниже, чем в прошлом квартале». Используя статистическую нотацию, альтернативную гипотезу можно записать так:

— H_a: Уровень рекомендаций в этом квартале < Уровня рекомендаций в прошлом квартале.

Остановитесь на мгновение и подумайте о сделанном допущении. Вы не видели никаких данных и статистических показателей, но можете оспорить саму логику подхода вашего стажера. Выдвигая нулевую гипотезу, он изначально настроил себя на победу. Если результаты опросов за два квартала практически не различаются или основаны на небольшой выборке клиентов, то доказательств в пользу отвержения исходного допущения может оказаться недостаточно. Именно поэтому главный по данным должен спросить: «Какова нулевая гипотеза?» Плохо сформулированная нулевая гипотеза может создать обманчивое впечатление истинности некоего утверждения просто в силу отсутствия доказательств обратного.

Помните, что цель науки — бросить вызов существующему положению вещей. Статус-кво соответствует нулевой гипотезе, а альтернативная гипотеза отражает то, во что верите вы. И с помощью собранных данных вы должны доказать, что нулевая гипотеза является маловероятной.

Чтобы доказать эффективность своей работы по повышению уровня удовлетворенности клиентов, ваш стажер должен проверить свою гипотезу следующим образом:

— H_0: Уровень рекомендаций в этом квартале ≤ Уровню рекомендаций в прошлом квартале.
— H_a: Уровень рекомендаций в этом квартале > Уровня рекомендаций в прошлом квартале.

(Мы вернемся к этому примеру чуть позже.)

Допущение эквивалентности

Предположим, вы заменяете ключевой ингредиент в пищевом продукте, чтобы сократить расходы. Ваша команда проводит опрос клиентов, предлагая им оценить вкус по 10-балльной шкале, чтобы выяснить, замечают ли они изменение. При использовании предыдущей рецептуры 18 из 20 человек говорили о своей готовности купить продукт. В ходе нового опроса о готовности купить продукт, приготовленный по новому рецепту, заявили 12 из 20 человек.

При использовании нулевой гипотезы: «Коэффициент покупок нового продукта = Коэффициент покупок прежнего продукта» и уровня значимости 0,05 *p*-значение[64], вычисленное с помощью статистического теста, равно 0,064. Поскольку *p*-значение превышает 0,05, нулевая гипотеза не отклоняется. Ваш начальник Джордж воспринимает это так: «Моя команда аналитиков показала, что между старым и новым более дешевым рецептом нет никакой статистически значимой разницы. Можно сократить расходы».

Джордж считает старый и новый рецепты эквивалентными, но у него просто может не быть достаточного количества данных, доказывающих обратное. Мораль здесь такова: не суметь опровергнуть статус-кво — это не то же самое, что подтвердить его[65].

Каков уровень значимости?

Как вы помните, уровень значимости — это пороговое значение, до достижения которого мы готовы мириться с тем, что данные не согласуются с нулевой гипотезой, продолжая при этом считать ее верной.

По традиции уровень значимости задается в 5% или 0,05. В некоторых отраслях может использоваться 1% или 0,01. Некоторые исследователи используют еще более низкое значение. Например, сотрудники Европейской организации по ядерным исследованиям (ЦЕРН) применяли невероятно низкий уровень значимости в процессе поиска крошечной физической частицы, известной как бозон Хиггса[66]. Чем меньше уровень значимости, тем меньше вероятность ложноположительного заключения.

---

[64] Мы использовали двусторонний точный тест Фишера.

[65] В этом примере требуется выполнение так называемой проверки эквивалентности, обсуждение которой выходит за рамки данной главы. Однако имейте ее в виду, расскажите о ней своей команде и применяйте ее. Если вам понятна логика этой главы, у вас не возникнет сложностей с пониманием данной концепции.

[66] “5 Sigma What’s That?” **blogs.scientificamerican.com/observations/five-sigmawhats-that**

Скорее всего, вы начнете с уровня значимости в 5%, однако имейте в виду, что при таком значении вы можете ошибочно отклонять нулевую гипотезу (то есть делать ложноположительное заключение) в 1 случае из 20. Это приемлемо для вас?

Очень легко выбрать уровень значимости, при котором ваши результаты всегда будут статистически значимыми. Во многих инструментах по умолчанию задано значение в 5%. Однако этот уровень может не соответствовать особенностям вашей отрасли. Кроме того, этот уровень может быть установлен вашим специалистом по работе с данными, который умолчал об этом изменении, сообщив вам лишь о том, что результат оказался статистически значимым. В худшем случае кто-то может провести тест и выбрать уровень значимости задним числом, — это все равно что бросить дротик, а затем передвинуть в нужное место мишень. Например, кто-то может провести статистический тест, получить *p*-значение 0,11, а затем задать уровень значимости 0,15, чтобы результат оказался статистически значимым.

Вот почему всегда важно спрашивать: «Каков уровень значимости?»

С практической точки зрения понижение уровня значимости, скажем, с 5 до 1% сокращает количество ложноположительных заключений. Это задает более высокую планку для отклонения нулевой гипотезы. В этом случае данные должны быть более экстремальными (или, по крайней мере, убедительными), чтобы вы отвергли нулевую гипотезу. Звучит не так уж и плохо, правда? Однако обратная сторона этого — увеличение числа ложноотрицательных заключений. Достичь компромисса в данном случае непросто, и какой-то универсальной рекомендации дать нельзя. Достижение правильного баланса зависит от конкретной проблемы и вашей способности справляться с последствиями ошибок, связанных с ложноотрицательными и ложноположительными заключениями.

Сколько тестов вы проводите?

После выяснения уровня значимости спросите своих специалистов по работе с данными, сколько тестов они проводят. Поскольку они смотрят на данные по-разному, они могут провести десятки, а то и сотни неформальных статистических тестов с уровнем значимости в 5%. Например, предположим, что исследователь тестирует большой набор данных о больных раком и типах пищевых продуктов, которые они едят, пытаясь выявить те продукты, которые могут быть связаны с более высокими показателями выживаемости.

При наличии в базе данных 100 различных видов продуктов питания и использовании уровня значимости в 5%, 5 продуктов покажутся статистически значимыми в борьбе с раком, даже если ни один из них не оказывает реального эффекта[67].

Каковы доверительные интервалы?

Ранее мы уже немного поговорили о доверительных интервалах и некоторых их компонентах. Пришло время собрать все фрагменты вместе.

Что мы подразумеваем под словом «доверие»? Как и в случае с понятием «значимость», смысл этого слова в статистике несколько отличается от повседневного. В статистике значимость и доверие неразрывно связаны. На самом деле между уровнем значимости и уровнем доверия существует симметрия — уровень значимости в 5% соответствует уровню доверия в 95%. Если более формально, то уровень доверия = 1 — уровень значимости. Поэтому вместо фразы «Мы отвергли нулевую гипотезу на уровне значимости 5%» вы можете услышать фразу: «Мы отвергли нулевую гипотезу на уровне доверия 95%».

Теперь давайте разберемся, почему человеку, анализирующему статистические результаты, следует запрашивать доверительные интервалы. Как говорилось ранее, доверительный интервал должен содержат истинное значение интересующего вас параметра популяции. В примере с опросом, который рассматривался ранее в главе, 95% доверительный интервал при размере выборки N = 1000 составлял (62,5%, 68,5%). Предположим, что вместо 1000 студентов нам удалось опросить только 100, и 65% из них сказали «да». В данном случае 95% доверительный интервал составляет (54,8%, 74,2%). Данный интервал намного шире исходного из-за гораздо меньшего размера выборки. В связи с этим мы допускаем больший диапазон значений, которому, по нашему мнению, должна принадлежать интересующая нас доля популяции. Однако по мере увеличения размера выборки N доверительный интервал сокращается. Больше данных — больше доказательств и меньше неопределенности. Логично, не правда ли? Если вам удастся собрать данные обо всей популяции, то необходимость в доверительном интервале отпадет: вы найдете истинное значение интересующего вас параметра популяции.

---

[67] Это можно исправить с помощью так называемой поправки на множественную проверку гипотез.

Доверительные интервалы также позволяют оценить размер эффекта в статистическом тесте[68]. Предположим, вы хотите узнать, совпадает ли рост у баскетболисток из США и Европы. Первым делом вы формулируете нулевую и альтернативную гипотезы:

— H_0: Средний рост американских баскетболисток = Среднему росту европейских баскетболисток.
— H_a: Средний рост американских баскетболисток ≠ Среднему росту европейских баскетболисток.

Теперь представьте, что ваш аналитик собирает данные и вычисляет *p*-значение для сравнения с уровнем значимости в 5%. Согласно результатам этого сравнения *p*-значение меньше уровня значимости. У баскетболисток из США и Европы разный рост, и результаты являются статистически значимыми[69].

Однако не кажется ли вам, что вы что-то упускаете? Иногда мы рассматриваем статистическую значимость как некое подтверждение. О, ваши результаты статистически значимы? Это означает, что они на 100% верны. Однако статистические тесты проводятся для обнаружения любой разницы, независимо от степени ее важности. Вот почему вам никогда не стоит довольствоваться *p*-значениями. Вернемся к примеру с баскетболистками и предположим, что средний рост игроков из США и Европы составляет 72 дюйма (183 см) и 71,5 дюйм (182 см) соответственно, а 95% доверительный интервал для этой разницы составляет 0,5 +/– 0,4 дюйма (1 см).

Имеет ли размер эффекта в полдюйма (1 см) практическое значение и представляет ли он вообще какой-либо интерес?

Имеет ли это практическое значение?

Крайне небольшие эффекты могут быть обнаружены при исследовании большой выборки. Если вы видите только *p*-значения, а не доверительные интервалы, то можете подумать, что обнаружили большой эффект, хотя на самом деле выявили лишь незначительное различие, не имеющее практической

[68] В статистике понятие «размер эффекта» может иметь множество значений. Здесь мы говорим о размере эффекта просто как о разнице между двумя числами.

[69] Нет, на самом деле мы не собирали данные и не проводили подобное исследование.

ценности. Итак, глядя на доверительные интервалы, спросите себя, является ли то, что вы видите, практически значимым эффектом.

Предполагаете ли вы наличие причинно-следственной связи?

Вы уже почти забыли о стажере. Вам интересно, привела ли его работа к повышению уровня удовлетворенности клиентов в этом квартале по сравнению с предыдущим. Чтобы представить вам доказательства улучшения, стажер сформулировал нулевую и альтернативную гипотезы следующим образом:

— H_0: Уровень рекомендаций в этом квартале ≤ Уровню рекомендаций в прошлом квартале.
— H_a: Уровень рекомендаций в этом квартале > Уровня рекомендаций в прошлом квартале.

В конце каждого квартала проводился опрос с использованием выборки, состоящей из 100 клиентов. В предыдущем квартале о своей готовности рекомендовать компанию сообщили 50/100 клиентов, а в этом квартале — 65/100. Являются ли результаты статистически значимыми при уровне 5%?

С помощью статистического теста[70] стажер вычисляет *p*-значение. Оно равно 0,02, то есть меньше 0,05, что позволяет вам отклонить нулевую гипотезу и признать то, что разница в результатах двух кварталов является статистически значимой. Стажер очень радуется и чувствует, что ему удалось компенсировать свое плохое выступление на баскетбольной площадке. «Похоже, мне удалось повысить уровень удовлетворенности клиентов».

Но так ли это? Корреляция не доказывает наличие причинно-следственной связи. Уровень удовлетворенности клиентов мог повыситься благодаря целому ряду факторов, и если только не был проведен спланированный эксперимент и не были тщательно измерены различия между старым подходом и идеями стажера, то у вас нет оснований предполагать наличие причинно-следственной связи.

---

[70] Тест проводился с помощью языка программирования R для статистической обработки данных: 'prop.test(c(65, 50), c(100, 100), alternative = «greater»)'

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе вы узнали о статистическом выводе и о том, как можно оспаривать предоставляемые вам статистические данные. В частности, вы познакомились с вопросами, которые стоит задавать по поводу тех или иных статистических утверждений, а также узнали, почему это важно. Вот эти вопросы:

— Каков контекст этой статистики?
— Каков размер выборки?
— Что вы тестируете?
— Какова нулевая гипотеза?
— Каков уровень значимости?
— Сколько тестов вы проводите?
— Каковы доверительные интервалы?
— Имеет ли это практическое значение?
— Предполагаете ли вы наличие причинно-следственной связи?

Вооружившись этим списком, вы сможете эффективно оспаривать, понимать и оценивать статистические показатели, с которыми сталкиваетесь.

ЧАСТЬ

III

Освойте набор инструментов дата-сайентиста

Скорее всего, взять в руки эту книгу вас побудили такие термины, как машинное обучение, искусственный интеллект и глубокое обучение. В этой части мы собираемся лишить их ореола таинственности.

Сфера данных, как бы мы ее ни назвали, постоянно изменяется. Однако фундаментальные концепции и инструменты существуют на протяжении десятилетий и лежат в основе самых актуальных тенденций, включая анализ текста и изображений. В части III вы найдете высокоуровневое описание этих концепций и методов.

Эта часть состоит из следующих глав:

Глава 8. Ищите скрытые группы.
Глава 9. Освойте модели регрессии.
Глава 10. Освойте модели классификации.
Глава 11. Освойте текстовую аналитику.
Глава 12. Концептуализируйте глубокое обучение.

Вы также узнаете о распространенных ошибках и ловушках, в которые попадают даже опытные аналитики.

ГЛАВА 8

Ищите скрытые группы

«Если вы проанализируете данные достаточно тщательно, то сможете отыскать послания Бога»

_ *Дилберт*[71]

Представьте, что вам звонит друг и просит помочь категоризовать его музыкальную коллекцию, представляющую собой набор винтажных виниловых пластинок. Вы соглашаетесь.

По дороге вы задумываетесь о способе организации такой коллекции. Начать можно с очевидных категорий, например, с музыкальных жанров и поджанров. Также можно сгруппировать музыкальные композиции по периодам, в которые они были выпущены. Эту информацию легко найти на обложке альбома.

Однако, когда вы приезжаете к своему другу, он вручает вам высокую стопку черных виниловых пластинок без обложек.

Ваш друг говорит, что купил эти пластинки на гаражной распродаже и понятия не имеет о жанрах, исполнителях или периодах выхода записанных на них композиций. Вы вынуждены отказаться от своих предвзятых представлений о способах классификации записей, поскольку у вас нет обложек альбомов, на которые вы могли бы опереться при их группировке. Задача категоризации пластинок внезапно оказывается намного сложнее, чем вы предполагали.

---

[71] Адамс, Скотт. Мультсериал «Дилберт». 3 января 2000 года.

Набравшись смелости, вы с другом достаете проигрыватель, прослушиваете альбомы и начинаете группировать их по категориям в зависимости от того, насколько они похожи. По мере прослушивания пластинок вы создаете новые группы, объединяете небольшие группы в одну и иногда переносите пластинку из одной группы в другую после ожесточенных споров о том, к какой группе она «ближе».

В конце концов у вас формируется 10 категорий, каждой из которых вы присваиваете описательное название.

То, что вы с другом только что сделали, называется обучением без учителя или неконтролируемым обучением. Вместо того чтобы опираться на предвзятые представления о данных, вы позволили данным организоваться самостоятельно[72].

Эта глава посвящена обучению без учителя — набору инструментов, предназначенных для обнаружения скрытых закономерностей и групп в наборах данных при отсутствии заранее определенных групп. Эта мощная техника используется в самых разных областях, начиная с распределения клиентов по разным маркетинговым категориям и заканчивая организацией музыкальных композиций на платформах Spotify или Pandora и упорядочиванием фотографий в телефоне.

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

В основе обучения без учителя или неконтролируемого обучения лежит идея о существовании скрытых групп в совокупности данных. Есть много способов, позволяющих выявить эти интересные закономерности и группы, если таковые действительно существуют. Как главный по данным, вы должны уметь ориентироваться в многочисленных методах обучения без учителя при поиске скрытых групп данных.

Но с чего начать, учитывая пугающе большое количество доступных методов неконтролируемого обучения? К счастью, для применения этих методов вам достаточно базового понимания связанных с ними основных действий. В данном случае речь идет:

— о снижении размерности с помощью анализа главных компонент;

— кластеризации методом *k*-средних.

[72] Ну вроде того. На самом деле все не так просто.

В этой главе мы рассмотрим данные методы и разберемся в том, что они означают и как именно позволяют достичь целей по снижению размерности и кластеризации соответственно.

СНИЖЕНИЕ РАЗМЕРНОСТИ

Снижение размерности — это процесс, с которым вы уже знакомы. Его примером может служить фотография, которая сводит трехмерный мир к плоскому двухмерному изображению, которое можно носить в кармане.

В случае с наборами данных мы работаем со строками и столбцами — наблюдениями и признаками. Количество столбцов (признаков) в наборе данных называется размерностью данных, а процесс объединения множества признаков в меньшее количество новых категорий при сохранении информации о наборе данных — снижением размерности. Проще говоря, мы ищем скрытые группы в столбцах набора данных, чтобы объединить несколько столбцов в один.

Давайте разберемся, почему это важно. С практической точки зрения в наборах данных с множеством признаков очень сложно разобраться. Их загрузка в компьютер может занимать много времени, и с ними тяжело работать. Из-за этого процесс разведочного анализа данных становится крайне утомительным, а в некоторых случаях — фактически нереализуемым. Например, в биоинформатике размерность набора данных может быть огромной. Каждое наблюдение исследователей может включать экспрессии тысяч генов, многие из которых сильно коррелируют друг с другом (а, следовательно, являются потенциально избыточными).

Снижение размерности данных позволяет сократить время вычислений, устранить избыточность и улучшить визуализацию результатов. Но как именно это можно сделать?

Создание составных признаков

Один из способов снизить размерность набора данных — объединение нескольких столбцов в составной признак. Давайте посмотрим, как это делается, на примере реальных данных о результатах сравнительных тестов 32 автомобилей, опубликованных в журнале *Motor Trend* за 1974 год. Сравнение этих автомобилей проводилось по 11 признакам, таким как расход топлива в милях на галлон, мощность двигателя в лошадиных силах, вес и другие

характеристики автомобиля[73]. Наша задача — создать метрику «эффективности» для ранжирования автомобилей от наиболее до наименее эффективных.

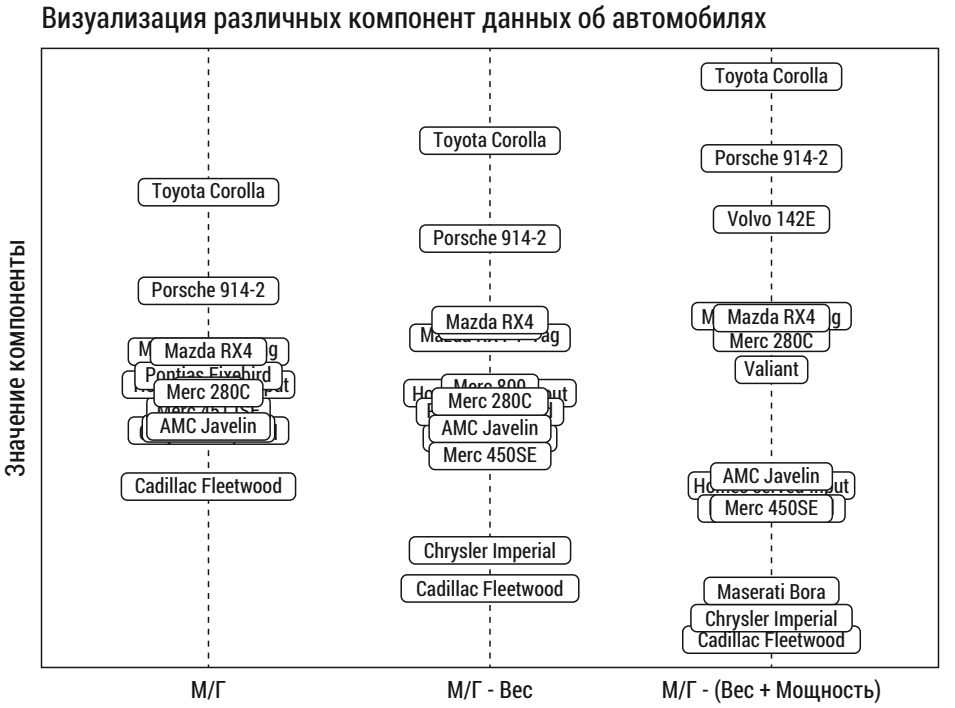


Рис. 8.1. Ранжирование автомобилей на основе различных составных признаков. Обратите внимание на увеличение дисперсии, то есть на то, как автомобили отдаляются друг от друга по мере объединения все большего количества признаков в единое измерение под названием «эффективность»

Очевидная точка отсчета для данного исследования — расход топлива в милях на галлон (М/Г). Результат ранжирования автомобилей по данному показателю находится в левой части диаграммы на рис. 8.1. Как видите, если не считать лучшего и худшего автомобиля с точки зрения расхода топлива, большинство автомобилей оказалось сгруппировано в центре. Можем ли мы добавить дополнительную информацию для дальнейшего разделения данных? Перейдем к средней части диаграммы. Здесь мы создали составной

[73] Речь идет о наборе данных mtcars, входящем в состав программы R. **http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html**. Для упрощения восприятия визуализации мы отображаем только 15 автомобилей из 32.

признак: расход топлива минус вес автомобиля[74], М/Г — вес. Обратите внимание: простое объединение двух признаков в один составной обеспечивает гораздо больший разброс.

Теперь давайте сделаем еще один шаг и создадим третью метрику эффективности. Эффективность = М/Г — (Вес + Мощность). (Подобное выражение называется линейной комбинацией.) Эта комбинация столбцов позволила разделить данные сильнее, чем другие признаки. Благодаря большему разбросу она дает нам больше информации об автомобилях, открывая нечто интересное. Как видите, тяжелые, пожирающие бензин автомобили находятся внизу, а легкие и экономичные — вверху. По сути, объединив исходные измерения, мы создали новое измерение данных (эффективность), которое позволяет нам игнорировать три исходных измерения. В этом и заключается суть снижения размерности.

В этом примере мы заранее знали, что объединение расхода топлива, веса и мощности двигателя автомобиля в новую составную переменную позволит обнаружить кое-что интересное в совокупности имеющихся данных. Но что, если вы не знаете, как и какие признаки следует комбинировать? Именно так обстоят дела в случае неконтролируемого обучения, и здесь в игру вступает анализ главных компонент.

АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ

Анализ главных компонент (АГК) — это метод снижения размерности, изобретенный в 1901 году[75] задолго до того, как термины «дата-сайентист» и «машинное обучение» стали частью бизнес-терминологии. АГК до сих пор остается популярной, хотя и зачастую неправильно понимаемой техникой. Мы постараемся устранить эту путаницу и разъяснить, в чем суть данного метода и его полезность.

В отличие от подхода, использованного в примере с автомобилями, алгоритм АГК заранее не знает, какие группы признаков следует объединять в составные признаки, и потому рассматривает все возможности. С помощью особых математических приемов он создает разные конфигурации

[74] Поскольку признаки имеют разный размах, перед объединением их необходимо привести к одной числовой шкале.

[75] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559–572.

измерений в поисках линейных комбинаций признаков, обеспечивающих максимальный разброс данных. Лучшие из этих составных признаков называются главными компонентами. Более того, главные компоненты — это новые измерения данных, которые не коррелируют друг с другом. Если бы мы применили данный подход в примере с автомобилями, то помимо измерения «эффективность» мы могли бы выявить и такое измерение, как «производительность».

Вероятно, вы хотите узнать, как можно определить, позволит ли АГК объединить признаки набора данных в значимые группы, образующие главные компоненты. Что именно мы ищем с помощью АГК?

Попробуем ответить на эти вопросы с помощью следующего мысленного эксперимента. Поскольку авторы данной книги не разбираются в автомобилях, в следующем примере будет использоваться другой (гипотетический) набор данных.

Главные компоненты спортивных способностей

Представьте, что вы работаете в спортивном лагере. У вас есть электронная таблица с сотнями строк и 30 столбцами. Каждая строка содержит информацию о физической подготовке спортсмена: количество отжиманий, приседаний и повторений становой тяги, которые он может сделать за одну минуту; за сколько времени он пробегает 40 метров, 100 метров, 1600 метров; показатели жизненно важных функций, такие как пульс в состоянии покоя и артериальное давление; и еще несколько показателей физической подготовки и состояния здоровья. Ваш начальник дал вам расплывчатое задание «обобщить данные», но вы увязли в большом количестве столбцов электронной таблицы. Без сомнения, в ней содержится огромный объем информации. Но нельзя ли свести эти 30 признаков к более разумному количеству, которое можно было бы использовать для обобщения и визуализации имеющихся данных?

Первым делом вы замечаете несколько очевидных закономерностей. Спортсмены, которые могут сделать больше всего отжиманий, как правило, делают больше повторений становой тяги, а те, кто медленно пробегает стометровку, имеют ту же проблему в случае со спринтом на 40 метров. Похоже, что многие признаки коррелируют друг с другом, поскольку измеряют связанные между собой показатели физической подготовки. Эта корреляция заставляет вас задуматься о возможном способе их сжатия в меньшее число

не коррелирующих друг с другом измерений, содержащих максимально возможное количество информации из исходных данных. В этом и заключается суть АГК*!*

Рис. 8.2. Анализ главных компонент позволяет сгруппировать и объединить столбцы набора данных в новые не коррелирующие между собой измерения

На рис. 8.2 в общих чертах представлено то, чего вы пытаетесь достичь. Однако вы не можете с легкостью исследовать корреляции между 30 переменными даже с помощью компьютера. (Для изучения каждой пары признаков потребовалось бы 435 отдельных диаграмм рассеяния[76].) Итак, вы прогоняете данные через алгоритм АГК, чтобы воспользоваться существующими в наборе данных корреляциями, и на выходе получаете два набора данных[77].

На рис. 8.3 показан первый набор. Строки таблицы содержат признаки спортсменов, а столбцы — веса этих признаков и их вклад в главную компоненту. Эти веса отражают важный шаг в процессе АГК, связанный с созданием новых измерений в данных. (Обратите внимание на то, что здесь слово «вес» — специальный термин, который имеет отношение к АГК, а не к поднятию тяжестей.) Мы визуализировали веса, однако имейте в виду, что в числовом выражении они представляют собой меру корреляции, значение которой находится в диапазоне от –1 до 1. Чем ближе ее значение к любому из экстремумов, тем сильнее корреляция и тем выше вклад исходного признака. Итак, вам нужно отыскать интересные закономерности в весах главных компонент (обозначенных на следующем рисунке аббревиатурой ГК). Веса, которые отклоняются далеко от вертикальной нулевой линии, могут рассказать много интересного.

---

[76] Количество сочетаний из 30 по 2 = 30!/((30 – 2)! 2!) = 435.

[77] Ни одна из программ не возвращает результаты АГК, показанные здесь. Чтобы обойтись без множества уравнений и чисел, мы решили сосредоточиться на визуализации.

В первом столбце «Веса для ГК 1» вы видите большие веса для таких признаков, как отжимания, приседания и становая тяга; они, как отмечалось ранее, положительно коррелируют друг с другом. АГК обнаружил это автоматически. Эту комбинацию признаков можно назвать «Сила». Глядя на веса для ГК 2, вы замечаете связь отрицательных столбиков с показателями «Скорости» (низкий пульс в состоянии покоя, медленный бег на 40 метров, медленный бег на 100 метров). Аналогичным образом вы можете дать ГК 3 название «Выносливость», а ГК 4 — «Здоровье».

Раньше у вас было несколько коррелирующих между собой измерений. Однако четыре новых измерения представляют собой четыре составных признака, которые не коррелируют друг с другом. А отсутствие корреляции означает, что каждое новое измерение предоставляет новую, непересекающуюся информацию. По сути, мы разбиваем содержащуюся в наборе данных информацию на отдельные измерения, как указано в строке «% информации для каждой компоненты». Используя всего лишь четыре новых признака, мы можем сохранить 91% информации, содержащейся в исходном наборе данных.

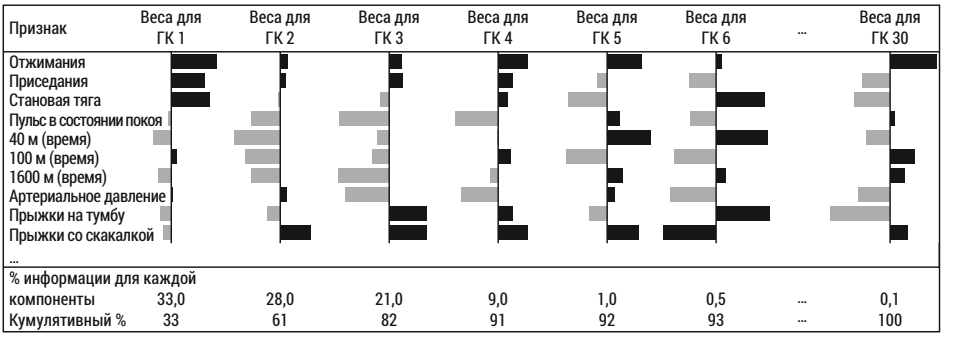


Рис. 8.3. АГК находит оптимальные веса для создания составных признаков, которые представляют собой линейные комбинации других признаков. Иногда вы можете присвоить новому составному признаку описательное название

С помощью весов, указанных на рис. 8.3, 30 исходных показателей физической подготовки каждого спортсмена можно преобразовать в такие главные компоненты, как «Сила», «Скорость», «Выносливость» и «Здоровье», используя линейные комбинации. Например, сила спортсмена рассчитывается по следующей формуле:

Сила = 0,6*(количество отжиманий) + 0,5*(количество повторений становой тяги) + 0,4*(количество приседаний) + (незначительный вклад остальных признаков)

Значения (веса) 0,6, 0,5 и 0,4 — результат АГК. Мы просто решили их визуализировать.

Выполнение этой серии вычислений для всех спортсменов дает нам второй результат применения алгоритма АГК, показанный на рис. 8.4. Это новый набор данных того же размера, что и исходный, только на этот раз максимально возможное количество информации было сосредоточено в первой группе некоррелированных главных компонент (также известных как составные признаки). Обратите внимание на резкое сокращение величины вклада главных компонент, начиная с пятой.

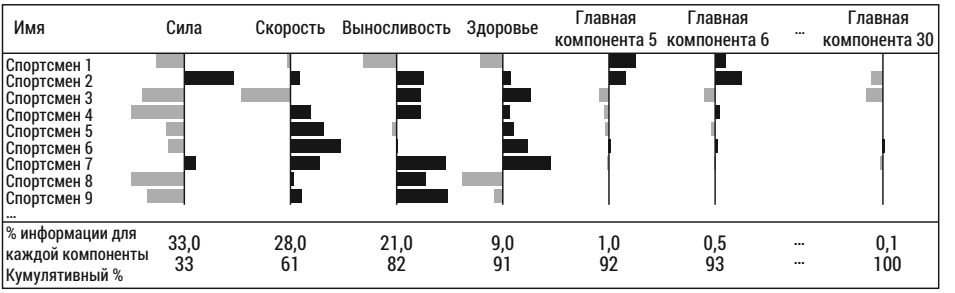


Рис. 8.4. Алгоритм АГК создает новый набор данных того же размера, что и исходный, где столбцы представляют собой составные признаки, называемые главными компонентами

Таким образом, вместо использования 30 переменных для объяснения 100% информации, содержащейся в исходном наборе данных, набор данных, показанный на рис. 8.4, может объяснить 91% этой информации с помощью всего лишь четырех признаков. Это позволяет нам проигнорировать 26 столбцов. Вот это понижение размерности! Вооружившись этим набором данных, вы можете выяснить, кто из спортсменов самый сильный, самый быстрый или обладает любой комбинацией этих признаков. Визуализировать и интерпретировать данные стало намного проще.

Анализ главных компонент. Резюме

Давайте сделаем шаг назад, чтобы кое-что прояснить.

Во-первых, когда речь идет о столбце в наборе данных, хорошим синонимом информации является дисперсия (мера разброса). Подумайте об этом так. Предположим, что мы добавили новый столбец в набор данных о спортсменах, показанный на рис. 8.2, под названием «Любимая марка обуви», и каждый спортсмен ответил: «Nike». В таком случае в этом столбце не было бы никаких вариаций, позволяющих отличить одного спортсмена от другого. Нет вариации = нет информации.

Основополагающая идея АГК — взять всю содержащуюся в наборе данных информацию (множество столбцов) и сжать как можно больше этой информации в как можно меньшее количество отдельных измерений (меньшее количество столбцов). Для этого алгоритм определяет, как именно каждое из исходных измерений коррелирует с другими. Корреляция, существующая между многими измерениями, объясняется тем, что они измеряют одну и ту же основополагающую вещь. В этом смысле у нас есть лишь несколько истинных измерений данных, охватывающих большую часть информации, содержащейся в наборе данных. Математика, лежащая в основе АГК, по сути «вращает» измерения, сводя их к меньшему количеству главных компонент и позволяя нам рассматривать их без потери большого количества информации.

Это напоминает процесс фотографирования. Например, вы можете сфотографировать великие пирамиды Египта с бесчисленного количества ракурсов, однако некоторые ракурсы оказываются более информативными, чем другие. Если вы сделаете снимок с помощью дрона сверху, то пирамиды будут выглядеть как квадраты. Если вы сфотографируете их, стоя точно напротив одной из граней, они будут выглядеть как треугольники. На какой угол необходимо повернуть камеру, чтобы зафиксировать максимальное количество информации при сведении трехмерного мира Гизы в двухмерную фотографию, способную произвести впечатление на друзей? Оптимальный ракурс можно найти с помощью АГК.

Потенциальные ловушки

Теперь, когда вы познакомились с основами АГК, мы должны признать, что в реальном мире наборы данных никогда не удается свести к столь же четко различимым главным компонентам, как в примере со спортсменами.

Из-за неупорядоченности данных результирующие главные компоненты зачастую бывают лишены ясного значения и описательных названий. Мы по опыту знаем, что в погоне за броским названием для главной компоненты люди зачастую создают описание несуществующих данных. Как главному по данным, вам не следует принимать уже готовые определения главных компонент. Когда кто-то представляет вам уже названные компоненты, постарайтесь оспорить их определения, выяснив, какие именно уравнения лежат в основе той или иной группировки.

Более того, АГК не сводится к исключению неважных или неинтересных переменных. Мы часто видим, как люди совершают эту ошибку. Главные компоненты генерируются на основе всех исходных признаков. Для этого ничего не удаляется. В примере со спортсменами каждый исходный признак может быть сгруппирован с несколькими другими для получения четырех главных компонент: Сила, Скорость, Выносливость и Здоровье. Помните о том, что набор данных, полученный в результате применения алгоритма АГК, по размеру аналогичен исходному. Аналитик должен сам решить, когда отбрасывать неинформативные компоненты, поскольку одного правильного способа сделать это просто не существует. Это означает, что, когда вам представляют результаты АГК, вам следует выяснить, как именно те, кто его проводил, решили, сколько компонентов стоит оставить.

Наконец, АГК основывается на предположении о том, что высокая дисперсия свидетельствует о присутствии в переменных чего-то интересного или важного. В некоторых случаях это предположение оказывается оправданным — но не всегда. Например, признак может иметь высокую дисперсию и при этом не иметь особого практического значения. Представьте, что мы добавили к данным о спортсменах такой признак, как количество жителей в родном городе каждого из них. Несмотря на большие различия, этот признак никак не связан с данными об их спортивных результатах. Поскольку алгоритм АГК стремится отыскать существенные вариации, он может ошибочно принять этот признак за нечто важное, хотя на самом деле это не так.

КЛАСТЕРИЗАЦИЯ

Группы признаков (столбцы) могут рассказать одну историю, как в случае с АГК, а группы наблюдений (строки) — другую. Именно здесь в игру вступает кластеризация[78].

По нашему опыту, кластеризация — самая интуитивно понятная техника работы с данными, потому что ее название точно отражает ее суть (в отличие от названия «Анализ главных компонент»). Если бы ваш начальник поручил вам разделить спортсменов на группы, вы бы поняли задачу. При анализе данных, представленных на рис. 8.5, у вас возник бы ряд вопросов — например, относительно возможного количества групп и способов их категоризации. Тем не менее у вас было бы от чего оттолкнуться. Например, вы могли бы сформировать одну группу из наиболее сильных и медленно бегающих спортсменов, а другую — из самых слабых и быстрых. Вы могли бы назвать эти группы «Бодибилдеры» и «Бегуны на длинные дистанции».

Потратьте минутку на размышления о том, как бы вы подошли к группировке этих данных и какие решения вам пришлось бы принять в ходе этого процесса. Если бы вам было лень это делать, вы могли бы сказать: «Каждый человек в этой таблице — спортсмен, поэтому есть только одна группа — спортсмены». Или: «Каждый человек образует отдельную группу. Всего есть N групп». И то и другое утверждение абсолютно бесполезно. Однако они позволили вам понять очевидное: количество групп должно быть больше 1, но меньше N.

Еще одно решение, которое вам придется принять самостоятельно, связано с определением степени «похожести» одного спортсмена на другого. Рассмотрим подмножество данных в табл. 8.1. Какие два из этих спортсменов сильнее всего похожи друг на друга?

Вы можете привести аргумент в пользу любой пары. Все зависит от того, по каким критериям вы оцениваете их «схожесть». Спортсмены A и B похожи по количеству отжиманий и пульсу. Спортсмены A и C демонстрируют самые лучшие результаты в беге на 1600 м и количестве отжиманий соответственно. А спортсмены B и C похожи тем, что бегают медленнее всех остальных. Здесь вы можете увидеть то, что хотите. Все зависит от того, какие признаки для вас наиболее важны, а также от того, что вы подразумеваете под

[78] АГК и кластеризация никак не связаны между собой, так что их можно использовать независимо друг от друга.

понятием «сходство». Процесс неконтролируемого обучения, разумеется, ничего об этом не знает.

Рис. 8.5. Кластеризация — это способ группировки строк набора данных, тогда как АГК позволяет группировать столбцы

Табл. 8.1. Какие из этих двух атлетов больше всего похожи друг на друга?

Спортсмен	Отжимания	Пульс в состоянии покоя	1600 м (время)
A	40	50	4:30 мин.
B	30	55	8:00 мин.
C	100	65	9:00 мин.

Этот пример демонстрирует основные проблемы кластеризации: сколько кластеров у нас должно быть? По каким критериям любые два наблюдения могут считаться «похожими»? И как лучше всего группировать такие наблюдения?

Начать можно с кластеризации методом *k*-средних[79].

КЛАСТЕРИЗАЦИЯ МЕТОДОМ К-СРЕДНИХ

Метод *k*-средних весьма популярен среди дата-сайентистов. С его помощью вы сообщаете алгоритму необходимое количество кластеров (*k*), после чего он группирует ваши N строк с данными в *k*-кластеров. Точки данных внутри кластера находятся «поблизости», в то время как сами кластеры максимально удалены друг от друга.

Запутались? Давайте рассмотрим пример.

[79] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory,* 28(2), 129–137.

Кластеризация точек продаж

Компания хочет распределить свои 200 точек продаж, показанных на рис. 8.6, по шести регионам континентальной части США. Их можно было бы распределить по стандартным географическим регионам (например, Средний Запад, Юг, Северо-Восток и так далее), однако местонахождение магазинов компании вряд ли удалось бы согласовать с этими предопределенными границами. Вместо этого компания попыталась сгруппировать данные с помощью метода *k*-средних. Набор данных состоит из 200 строк и двух столбцов, в которых указаны значения широты и долготы[80].

Рис. 8.6. 200 торговых точек компании до кластеризации

Цель состоит в нахождении на карте шести новых местоположений, каждое из которых является «центром» кластера. В числовом выражении эта центральная точка, по сути, представляет собой среднее значение всех членов группы (отсюда и название метода *k*-средних). В данном примере центры

[80] В этом примере мы делаем множество упрощающих допущений. С технической точки зрения этот метод не подходит для группировки точек на сфере, поскольку координаты широты и долготы не находятся в евклидовом пространстве. Используемая нами метрика расстояния не учитывает кривизну Земли, а также практические ограничения, вроде доступа к автомагистралям.

кластеров могут быть возможными локациями региональных офисов, и каждый из 200 магазинов может быть привязан к ближайшему офису.

Вот как это работает. Сначала алгоритм выбирает шесть случайных местоположений в качестве потенциальных региональных офисов. Почему случайных? Потому что нужно с чего-то начать. Затем, используя расстояние между точками на нашей карте (что называется «по прямой»), каждый из 200 магазинов назначается тому или иному из шести кластеров в зависимости от того, к какому из стартовых местоположений он ближе всего. Результат показан в левом верхнем углу на рис. 8.7 («Раунд 1»).

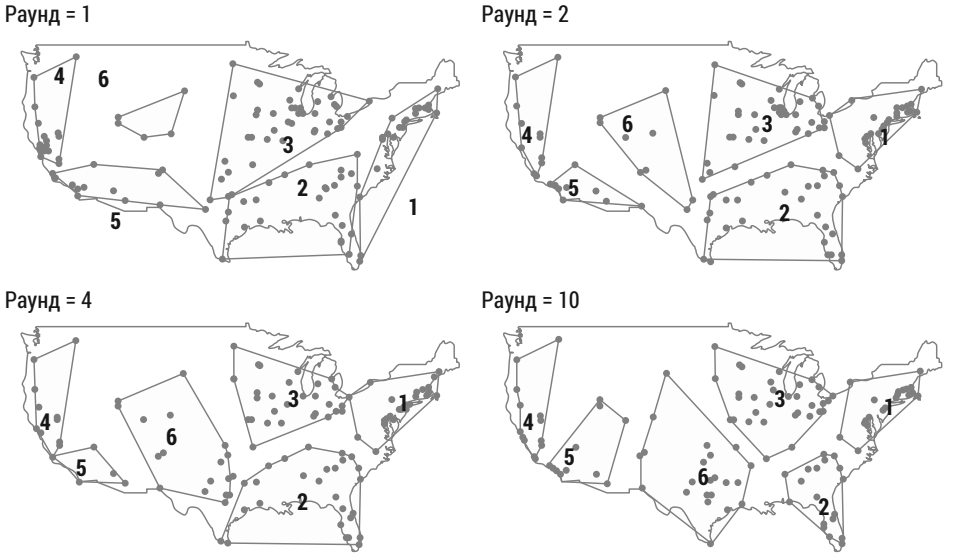


Рис. 8.7. Применение метода *k*-средних для кластеризации розничных магазинов

Каждое число обозначает начальное местоположение и связано с ограничивающим кластер многоугольником. Обратите внимание на то, что в «Раунде 1» местоположение «6» находится далеко от своего кластера, по крайней мере, в этой первой итерации. Также обратите внимание на то, что некоторые выбранные местоположения оказались в океане.

В каждом раунде алгоритма все точки в кластере усредняются для получения центральной точки (называемой «центроидом»), в которую перемещается число. В результате каждый из 200 магазинов может оказаться ближе к другому региональному офису, которому он и переназначается. Процесс продолжается до тех пор, пока точки не перестанут переходить из кластера

в кластер. На рис. 8.7 показаны результаты последовательных раундов кластеризации методом *k*-средних.

Таким образом, компания объединила 200 своих магазинов в шесть кластеров и нашла в каждом из них потенциальное место для расположения регионального офиса.

Итак, алгоритм *k*-средних пытается выявить в данных естественные кластеры и постепенно стягивает *k*-случайных начальных точек к центрам этих кластеров.

Потенциальные ловушки

В предыдущем примере мы использовали расстояние «по прямой», однако есть несколько типов формул расстояния, которые можно применить при кластеризации набора негеографических данных. Их обсуждение выходит за рамки этой книги. Кроме того, ни одна из этих формул не является правильной. Тем не менее не стоит полагать, что ваша команда аналитиков использовала самую подходящую, а не самую простую формулу расстояния. Обязательно спросите, какую именно формулу они использовали и почему.

Вам также необходимо учитывать масштаб своих данных. Не следует слепо доверять результатам, потому что математика может сгруппировать по степени «близости» два доминирующих значения. Например, возьмем трех сотрудников, данные о которых приведены в табл. 8.2. Какие два кажутся вам максимально «близкими» друг к другу?

Табл. 8.2. Немасштабированные данные могут сбить алгоритмы кластеризации с толку

Сотрудник	Возраст	Количество детей	Доход, $
A	36	3	100 000
B	37	2	80 000
C	22	0	101 000

При отсутствии должного масштабирования данных значение дохода будет доминировать в большинстве формул расстояния, поскольку разница в его абсолютном значении между любыми двумя точками данных — самая существенная. Это означает, что «расстояние» между людьми A и C будет «меньше», чем между A и B, если судить по уровню дохода. И это несмотря на то, что сотрудники A и B могли бы образовать более предпочтительную

группу, состоящую из двух работающих родителей в возрасте более 30 лет, в то время как человек С — новичок, который только что окончил колледж и получил высокооплачиваемую должность в фирме.

Наконец, помните о том, что при создании групп мы прибегаем к помощи компьютера, а это означает, что правильного ответа не существует. Все модели ошибочны. Однако при правильном подходе метод *k*-средних может оказаться полезным.

Иерархическая кластеризация

Прежде чем завершить этот раздел, стоит упомянуть еще об одном популярном алгоритме кластеризации под названием «иерархическая кластеризация». При использовании этого алгоритма количество кластеров не определяется заранее, как в случае с методом *k*-средних.

Вспомните пример из начала этой главы, в котором вам с другом нужно было упорядочить музыкальные записи при отсутствии обложек альбомов. Вы не знали, сколько существует кластеров. По сути, вы начали с N-групп, каждая из которых состояла из одной записи. Однако в процессе прослушивания пластинок группы начали формироваться естественным образом. Возможно, вы объединили две записи в категорию «современный джаз». Если у вас также была группа из трех записей в жанре «классический джаз», вы могли счесть такую детализацию излишней и объединить две группы в одну под общим названием «джаз».

Подобный способ создания групп «снизу вверх» позволяет произвести иерархическое упорядочение ваших данных. При этом вы сами решаете, на каком уровне иерархии должны находиться конечные группы.

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе вы узнали об обучении без учителя, которое часто описывается как способ, позволяющий данным организоваться в группы самостоятельно. Однако, как отмечалось в сноске в начале главы, все не так просто. Способность обнаруживать группы в наборе данных — это большая сила, а, как мы знаем, чем больше сила, тем больше ответственность. Мы надеемся, что вы уловили эту мысль.

Табл. 8.3. Обучение без учителя. Резюме

Обучение без учителя	Снижение размерности	Кластеризация
Пример	Анализ главных компонент	Метод *k*-средних
Что это?	Группировка и объединение столбцов (признаков)	Группировка строк (наблюдений)
Что делает?	Находит меньший набор новых некоррелированных признаков, содержащий большую часть информации в наборе данных.	Группирует похожие наблюдения, создавая *k*-значимых «кластеров» в данных.
Зачем?	Это позволяет вам визуализировать и исследовать данные или уменьшить размер набора данных для ускорения процесса вычислений. Как правило, АГК является промежуточным этапом анализа.	Данный метод позволяет выявить закономерности и структуру данных и дает возможность по-разному воздействовать на кластеры (например, запускать разные маркетинговые кампании для разных сегментов рынка).
Необходимый контроль	Пользователь должен решить, как масштабировать данные, сколько главных компонент оставить и как их интерпретировать.	Пользователь должен решить, как масштабировать данные, а также выбрать подходящую метрику «расстояния» и необходимое количество кластеров.

Возможность какой-либо группировки данных зависит от выбранного алгоритма, его реализации, качества исходных данных и существующей в них вариации. Это означает, что принятие разных решений может приводить к созданию разных групп. Проще говоря, обучение без учителя требует контроля. Вы не можете просто нажать кнопку на компьютере и позволить данным организоваться самостоятельно. Вам необходимо принять определенные решения, которые мы обобщили (наряду с описанными в этой главе алгоритмами) в табл. 8.3.

В завершение следует еще раз сказать о том, что при обучении без учителя не бывает ни правильных группировок, ни правильных ответов. На самом деле вы можете считать подобные упражнения продолжением своего исследовательского путешествия по области анализа данных, описанного в главе 5, позволяющие вам взглянуть на данные под другим углом.

ГЛАВА 9

Освойте модели регрессии

«Регрессионный анализ похож на один из тех изощренных мощных инструментов, который относительно легко использовать, но сложно делать это правильно. А его неправильное использование потенциально опасно»

_ *Чарльз Уилан, цитата из книги «Голая статистика»*[81]

ОБУЧЕНИЕ С УЧИТЕЛЕМ

Предыдущая глава была посвящена обучению без учителя — способу обнаружения закономерностей или кластеров в наборе данных без использования заранее определенных групп. Помните, что к неконтролируемому обучению мы подходим без каких-либо предвзятых представлений. Вместо этого мы опираемся на основополагающие аспекты данных, задаем некоторые границы и позволяем данным организоваться самим.

Однако во многих случаях о наборе данных что-то известно. Тогда вы можете использовать обучение с учителем или контролируемое обучение для выявления в нем взаимосвязей с помощью входных и известных выходных данных. В данном случае у вас есть правильные ответы, на которых вы можете «учиться». Затем вы можете оценить надежность модели, сравнив ее результаты с тем, что вам известно о реальном мире. Хорошая модель позволит вам делать точные прогнозы и объяснять некоторые основополагающие взаимосвязи между входными и выходными данными.

---

[81] «Голая статистика. Самая интересная книга о самой скучной науке», Чарльз Уилан (Издательство: Манн, Иванов и Фербер, 2022).

Как вы, вероятно, помните, обучение с учителем уже упоминалось во введении — в самом начале вашего пути становления главным по данным. Тогда мы попросили вас спрогнозировать, будет ли новый ресторан сетевым или независимым. Чтобы сделать соответствующее предположение, вы сначала изучили местоположения существующих ресторанов (входные данные) и известные метки «сетевой» или «независимый» (выходные данные). Вы обнаружили взаимосвязи между входными и выходными данными и создали «модель» в своей голове, которую использовали для того, чтобы обоснованно спрогнозировать метку для нового местоположения.

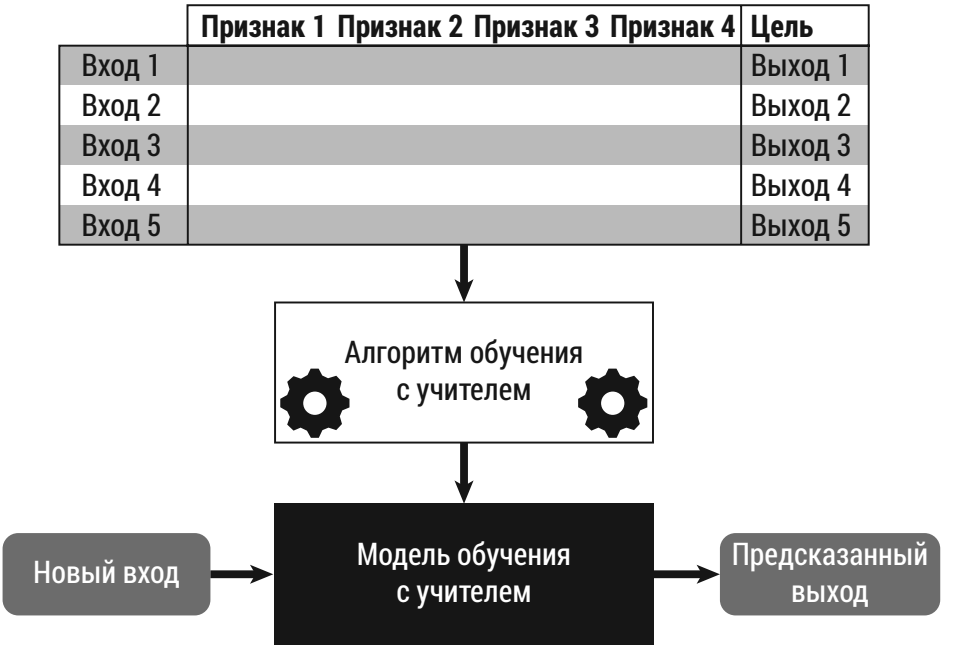


Рис. 9.1. Базовая парадигма обучения с учителем: сопоставление входных данных с выходными

Может быть, вас это удивит, но все задачи контролируемого обучения подчиняются одной и той же парадигме. Она представлена на рис. 9.1. Входные и выходные данные, называемые обучающими, подаются на вход алгоритма, который использует взаимосвязи между входными и выходными данными для создания прогностической модели (уравнения). Эта модель может принимать новые входные данные и сопоставлять их с прогнозируемыми выходными данными. Когда выходные данные представляют собой числа,

модель контролируемого обучения называется регрессионной. Когда выходными данными являются метки (категориальные переменные), модель называется классификационной.

О регрессионных моделях мы поговорим в этой главе, а о классификационных — в следующей.

Эта парадигма охватывает множество интересных и ценных с практической точки зрения задач контролируемого обучения, применяемых как в старых, так и в новых технологиях. Детектор спама в вашей электронной почте, оценка стоимости вашего дома или квартиры, перевод речи, приложения для распознавания лиц и беспилотные автомобили — все это результат контролируемого обучения. В табл. 9.1 указаны входные и выходные данные, а также типы моделей, используемые в вышеперечисленных сферах.

Табл. 9.1. Области применения контролируемого обучения

Приложение	Входные данные	Выходные данные	Тип модели
Детектор спама	Текст электронного письма	Спам или Не спам	Классификационная
Оценка недвижимости	Характеристики и местоположение объекта	Примерная цена продажи	Регрессионная
Перевод речи	Текст на английском языке	Текст на китайском языке	Классификационная (каждое слово является меткой)
Распознавание лиц	Изображение	Лицо обнаружено или Лицо не обнаружено	Классификационная
Умные колонки	Аудио	Среагировало ли устройство на имя «Алекса»?	Классификационная

По мере развития областей применения контролируемого обучения становится все легче упустить из виду тот факт, что в их основе лежит классический метод линейной регрессии, разработанный примерно в 1800 году. Линейная регрессия, в частности, метод наименьших квадратов[82], — рабочая

[82] Когда вы слышите словосочетание «линейная регрессия», чаще всего речь идет именно о регрессии методом наименьших квадратов. Существуют и другие типы линейной регрессии, но метод наименьших квадратов наиболее популярен.

лошадка контролируемого обучения; она часто применяется в первую очередь при прогнозировании чего-либо. Этот мощный метод используется повсеместно, и им нередко злоупотребляют.

ЛИНЕЙНАЯ РЕГРЕССИЯ: ЧТО ОНА ДЕЛАЕТ

Предположим, вы продаете лимонад в торговом центре и предполагаете, что температура влияет на объем продаж. То есть чем жарче на улице, тем больше лимонада вы продаете. Эта закономерность, если она верна, может помочь вам планировать закупки и прогнозировать уровень продаж в те или иные дни.

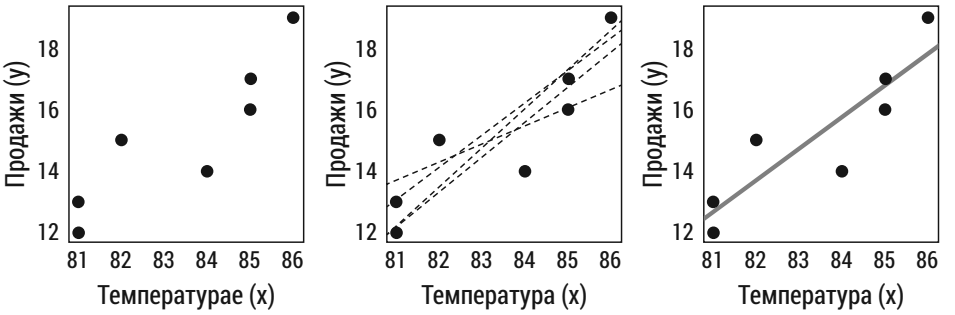


Рис. 9.2. Этим данным достаточно хорошо соответствует множество линий, но какая из них лучше? Определить это нам поможет линейная регрессия

Вы наносите на график исторические данные (левый график на рис. 9.2) и замечаете нечто, напоминающее линейный тренд. Чтобы провести линию через эти точки данных, вы можете использовать уравнение[83] Продажи = m(Температура) + b. Простое уравнение, подобное этому, является своего рода моделью[84]. Но как выбрать числа m (наклон линии) и b (точка пересечения с осью) для построения этой модели?

---

[83] При изучении алгебры вы познакомились с уравнением прямой линии: $y = mx + b$. Для любого входа x вы можете получить выход y, умножив x на m и прибавив b. Если $y = 2x + 5$, то вход $x = 7$ дает выход $y = 2\times7 + 5 = 19$.

[84] Краткое напоминание по поводу терминологии: выход y называется переменной отклика, целевой или зависимой переменной. Вход x называется признаком, предиктором или независимой переменной. Вы можете столкнуться со всеми этими терминами в своей работе.

Можно попробовать угадать. На среднем графике на рис. 9.2 показаны четыре возможные линии, и все эти предположения кажутся вполне разумными. Но это всего лишь предположения: как бы они ни были к этому близки, они не объясняют основополагающие взаимосвязи в данных.

Линейная регрессия предполагает выполнение вычислений для получения линии наилучшего соответствия. Под наилучшим соответствием мы подразумеваем то, что данная линия оптимальным образом объясняет имеющийся линейный тренд и разброс данных. Она представляет собой оптимальное решение, насколько это возможно с математической точки зрения с учетом предоставленных данных. На правом графике на рис. 9.2 показан результат применения линейной регрессии в виде уравнения: Продажи = 1,03 (Температура) — 71,07.

Давайте посмотрим, как это работает.

Регрессия методом наименьших квадратов: больше, чем умное название

Давайте на мгновение сосредоточимся на нашей выходной переменной — уровне продаж лимонада. Если бы мы хотели спрогнозировать объем будущих продаж, разумно было бы вычислить среднее значение для прошлых показателей (12 + 13 + 15 + 14 + 17 + 16 + 19) / 7 = 15,14 доллара. Итак, мы получили простую линейную модель, согласно которой Продажи = 15,14 доллара.

Обратите внимание на то, что это по-прежнему линейное уравнение, только без учета температуры. Это означает, что вне зависимости от температуры мы прогнозируем продажи на уровне 15,14 доллара. Мы знаем, что это весьма наивное предположение, но оно соответствует нашему определению модели, сопоставляющей входные данные с выходными, даже если все выходные данные являются одинаковыми.

Итак, насколько хороша наша простая модель? Чтобы оценить ее производительность, давайте подсчитаем, насколько далеко прогнозируемый уровень продаж находится от каждого из фактических показателей. При температуре 86 °F (30 °C) вы продали лимонада на 19 долларов, а модель предсказывала 15,14 доллара. При температуре 81 °F (27 °C) вы продали лимонада на 12 долларов, а прогноз модели опять же составлял 15,14 доллара. В первом случае прогноз оказался примерно на 4 доллара ниже фактического значения, а во втором — примерно на 3 доллара выше. На данный момент наша модель далека от идеала. И чтобы как следует оценить ее предсказательную

способность, нам нужно понять разницу между прогнозом модели и тем, что произошло на самом деле. Эта разница называется погрешностью и показывает, насколько сильно прогнозное значение отклоняется от фактического.

Как же можно измерить эту погрешность, чтобы понять, насколько хорошо работает наша модель? Для этого мы могли бы взять каждую фактическую цену продажи и вычесть ее из среднего прогнозного значения, которое составляет 15,14 доллара. Однако в этом случае результат суммирования погрешностей всегда будет равен нулю, потому что среднее значение, которое мы используем в качестве предиктора, представляет собой арифметический центр этих точек. Итоговая разность между всеми этими точками и центральным значением всегда равна нулю.

Однако нам все-таки нужен какой-то способ агрегировать эти погрешности, поскольку модель явно далека от идеала. Наиболее распространен так называемый метод наименьших квадратов, который предполагает возведение всех значений разности в квадрат для того, чтобы сделать их положительными[85]. При суммировании этих чисел результат не будет равен нулю (такое возможно, только если в наших данных вообще нет погрешностей). Полученный результат мы называем суммой квадратов.

Рассмотрим рис. 9.3(а). На нем вы видите исходную диаграмму рассеяния, где x = Температура, а y = Продажи. Обратите внимание: мы применили здесь нашу наивную модель, прогнозное значение которой не зависит от температуры и всегда равно 15,14, что дает нам горизонтальную линию Продажи = 15,14. Другими словами: точка данных с температурой = 86 (30 °C) и фактическим значением продаж = 19 соответствует прогнозному значению продаж = 15,14. Сплошная вертикальная линия показывает разницу между фактическим и прогнозным значением для этой точки (и всех остальных точек). При использовании регрессии мы возводим длину этой линии в квадрат, чтобы получить квадрат с площадью 14,9.

Фактическому уровню продаж в 15 долларов так же соответствует прогнозное значение 15,14 доллара, поэтому соответствующий квадрат имеет площадь $(15{,}14–15)^2 = 0{,}02$. Сложив площади квадратов для всех точек, мы получим суммарную погрешность нашей простой модели. В правой части

[85] Использование абсолютных значений также позволило бы сделать отклонения положительными перед агрегированием. Однако возведение в квадрат более предпочтительно с математической точки зрения, поскольку оно имеет свойство дифференцируемости, что было жизненно важно на ранних этапах применения метода линейной регрессии, когда все расчеты приходилось делать вручную.

рис. 9.3(a) процесс вычисления суммы квадратов ошибок представлен наглядно. Чем больше сумма квадратов, тем хуже модель соответствует данным, и наоборот: чем меньше сумма квадратов, тем лучше соответствие.

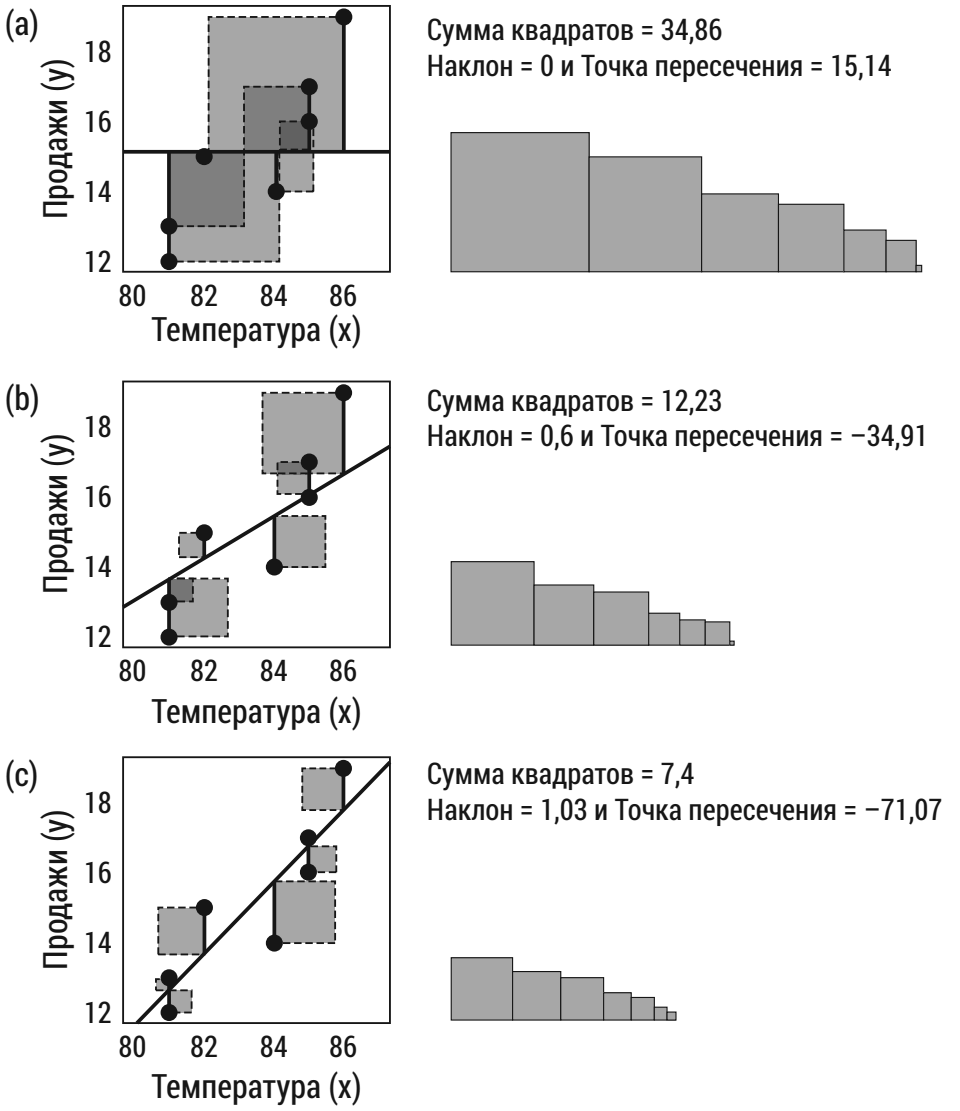


Рис. 9.3. Регрессия методом наименьших квадратов предполагает нахождение такой линии, при которой сумма площадей квадратов отклонений прогнозируемых значений от фактических является минимальной

Возникает вопрос: можем ли мы найти такое значение наклона и точки пересечения, которые позволяют оптимизировать (то есть минимизировать) сумму квадратов отклонений? В настоящее время наша наивная модель не имеет наклона, но имеет точку пересечения на отметке 15,14 доллара.

Очевидно, что модель, представленная на рис. 9.3 (a), не оптимальна. Чтобы приблизиться к оптимуму, давайте добавим наклон m, введя в уравнение переменную температуры. На рис. 9.3(b) мы предполагаем, что разумными значениями для наклона и точки пересечения могут быть 0,6 и –34,91 соответственно. Это превращает нашу модель с плоской линией, представленную на рис. 9.3(a), в наклонную линию, которая отражает некоторый восходящий тренд. Кроме того, мы сразу замечаем уменьшение общей площади квадратов.

Благодаря включению в модель переменной температуры значение отклонения существенно уменьшилось. Прогнозное значение для точки с температурой = 86 (30 °C) изменилось с 15,14 в случае простой модели на Продажи = 0,6(86) – 34,91 = 16,69, что привело к уменьшению вклада соответствующего наблюдения в сумму квадратов с 14,9 до $(16{,}69–19)^2 = 5{,}34$.

Вы могли бы подставлять значения наклона и точки пересечения вплоть до получения единственной комбинации, минимизирующей сумму квадратов. Однако линейная регрессия позволяет сделать это математически. На рис. 9.3(c) показана наименьшая сумма квадратов ошибок для имеющихся данных. Малейшее отклонение от указанных значений наклона и точки пересечения привело бы к увеличению площадей этих квадратов.

Вы можете использовать эту информацию, чтобы оценить, насколько хорошо итоговая модель соответствует данным. Результат линейной регрессии, показанный на рис. 9.3(c), все еще не идеален, однако он явно превосходит модель, представленную на рис. 9.3 (a), прогнозное значение которой каждый раз составляло 15,14 доллара.

Насколько превосходит? Мы начали с площади (суммы квадратов) в 34,86, а при использовании итоговой модели эта площадь уменьшилась до 7,4. Это значит, что мы уменьшили общую площадь на (34,86–7,4) = 27,46, то есть на 27,46/34,86 = 78,8%. В таких случаях часто говорят, что модель «объяснила», «описала» или «предсказала» 78,8% (или 0,788) дисперсии данных. Это число называется «R-квадратом» или R^2.

Если модель идеально соответствует данным, $R^2 = 1$. Однако не стоит рассчитывать на то, что в работе вам будут часто встречаться модели с высоким

значением R^2 [86]. Когда такое происходит, это говорит о том, что, скорее всего, произошла ошибка, и вам следует пересмотреть процессы сбора данных. Как вы помните из главы 3, вариации присутствуют во всем, и их невозможно объяснить полностью. Так уж устроена Вселенная.

ЛИНЕЙНАЯ РЕГРЕССИЯ: ЧТО ОНА ДАЕТ

Давайте быстро повторим то, что мы обсуждали ранее, в контексте парадигмы контролируемого обучения, представленной на рис. 9.1. У нас был набор данных, состоящий из столбца с входными значениями и столбца с выходными значениями, который мы подали на вход алгоритма линейной регрессии. Этот алгоритм извлек из данных оптимальные коэффициенты для подстановки в линейное уравнение Продажи = m(Температура) + b, создав модель Продажи = 1,03(Температура) — 71,07, которую можно использовать для прогнозирования прибыли от продажи лимонада.

Модели линейной регрессии пользуются популярностью во многих отраслях, потому что они не только делают прогнозы, но и объясняют то, как входные признаки соотносятся с выходными данными. (Кроме того, их совсем не трудно вычислить.) Коэффициент наклона, равный 1,03, говорит о том, что при повышении температуры на один градус можно ожидать увеличения продаж на 1,03 доллара. Это значение сообщает нам как величину, так и направление влияния входных данных на выходные.

Учитывая то, что в мире и в собираемых данных присутствует случайность и изменчивость, можно предположить наличие встроенной изменчивости и в значениях коэффициентов линейной регрессии. Если бы вы собрали новый набор данных о продажах своего лимонада, вы могли бы обнаружить, что при росте температуры на 1 градус ваша выручка увеличивается не на 1,03, а на 1,25 доллара. Данные, подаваемые на вход алгоритма, являются выборкой, поэтому вам следует думать о полученных результатах в терминах статистики. Статистическое программное обеспечение помогает это делать, предоставляя *p*-значения для каждого коэффициента (нулевая гипотеза, H_0: коэффициент = 0) и сообщая о наличии статистически значимого

[86] Для простой регрессии с одним входным параметром R^2 представляет собой квадрат коэффициента корреляции, который мы обсуждали в главе 5. Однако значение R^2 может быть и отрицательным. Такое бывает, когда модель линейной регрессии оказывается менее эффективной, чем предсказание среднего значения.

отличия коэффициента от нуля. Например, коэффициент 0,000003 очень близок к нулю и для практических целей может считаться нулевым в вашей модели.

Иными словами, если коэффициент значимо не отличается от нуля, вы можете исключить соответствующий признак из своей модели, поскольку входное значение не влияет на выходное. Разумеется, уроки статистики из главы 6 не теряют при этом своей актуальности. Коэффициент может быть статистически, но не практически значимым. Всегда выясняйте коэффициенты моделей, влияющих на ваш бизнес.

Включение множества признаков

Мы предполагаем, что ваш бизнес не ограничивается простой торговлей лимонадом. Ваши продажи, скорее всего, зависят не только от температуры (если это сезонный бизнес), но и от многих других факторов. К счастью, простую модель линейной регрессии, о которой мы говорили выше, можно расширить, включив в нее множество признаков[87]. Регрессия с одним входным параметром называется простой линейной регрессией, а с несколькими — множественной линейной регрессией.

Рассмотрим пример множественной линейной регрессии на основе данных о жилье, которые мы анализировали в главе 5. Этот набор данных содержит 1234 дома и 81 входной параметр, из которых для упрощения примера мы рассмотрим только 6. (Мы также могли бы использовать АГК для снижения размерности, но не стали этого делать, чтобы не усложнять пример.)

Давайте построим модель для прогнозирования цены продажи дома (выходной параметр) на основе площади участка, года постройки, площади 1-го, 2-го этажа и подвала в квадратных футах и количества полноценных ванных комнат. На основе данных алгоритм линейной регрессии вычисляет наилучшие значения точки пересечения и коэффициентов, перечисленные в табл. 9.2.

---

[87] Верхний предел количества признаков/входных параметров в модели линейной регрессии составляет N — 1, где N — количество строк в наборе данных. Таким образом, для прогнозирования ежемесячных объемов продаж на 12-месячный период вы можете использовать до 11 входных параметров.

Табл. 9.2. Модель множественной линейной регрессии для описания данных о недвижимости. Все соответствующие *p*-значения статистически значимы на уровне 0,05

Входной параметр	Коэффициент	*p*-значение
(Intercept) (точка пересечения)	–1614841,60	<0,000
LotArea (площадь участка)	0,54	<0,000
YearBuilt (год постройки)	818,38	<0,000
1stFlrSF (площадь 1 этажа в кв. футах)	87,43	<0,000
2ndFlrSF(площадь 2 этажа в кв. футах)	90,00	<0,000
TotalBsmtSF (площадь подвала в кв. футах)	53,24	<0,000
FullBath (полноценные ванные комнаты)	–7398,13	0,017

Основной принцип модели множественной регрессии состоит в том, чтобы изолировать влияние одной переменной, контролируя при этом остальные. Например, мы можем сказать, что при прочих неизменных значениях входных данных цена продажи дома, построенного годом позднее (в среднем), будет выше на 818,38 доллара. Коэффициенты каждого признака показывают величину и направление его воздействия на цену. Обязательно учитывайте единицы измерения. Добавление 1 единицы площади в квадратных футах отличается от добавления 1 единицы к количеству ванных комнат. Статистик может масштабировать данные при необходимости сравнения сопоставимых коэффициентов.

Каждый коэффициент также подвергается соответствующему статистическому тесту, который сообщает нам о том, имеет ли его значение статистически значимое отличие от нуля. Если нет, мы можем без опасений исключить его из модели, поскольку он не добавляет никакую информацию и не влияет на результат.

ЛИНЕЙНАЯ РЕГРЕССИЯ: КАКУЮ ПУТАНИЦУ ОНА ВЫЗЫВАЕТ

Если бы мы были какими-нибудь аферистами, мы бы закончили главу предыдущим разделом, предложив вам приобрести программу для расчета линейной регрессии в качестве панацеи, позволяющей решить все проблемы вашего бизнеса. Наш рекламный слоган был бы таким: «Введите данные, получите модель и начните делать прогнозы относительно своего бизнеса уже сегодня!» Звучит фантастически просто — однако к этому моменту вы уже наверняка понимаете, что при работе с данными ничто не так просто, как

кажется (или рекламируется). Как говорилось в эпиграфе к этой главе, при неправильном применении линейная регрессия может оказаться потенциально опасной. Поэтому при создании или использовании регрессионных моделей всегда сохраняйте здоровый скептицизм. Уравнения, терминология и вычисления создают впечатление, будто модель линейной регрессии способна автоматически исправить любую проблему в вашем наборе данных. Но это не так.

Давайте рассмотрим некоторые подводные камни использования линейной регрессии.

Пропущенные переменные

Модели контролируемого обучения не могут выявить взаимосвязь между входной и выходной переменной в случае исключения входной переменной из модели. Рассмотрим нашу простую модель, которая предсказывала уровень продаж лимонада на основе средних значений прошлых продаж без учета температуры.

Главные по данным, будучи осведомленными об этой проблеме, могут предложить для включения в модели информативные, релевантные признаки. Однако не стоит отдавать выбор признаков на откуп аналитикам. Ключ к созданию успешной модели контролируемого обучения — включение в нее правильных данных и наличие опыта в интересующей предметной области.

Например, модель с ценами на жилье, описанная в предыдущем разделе, имеет значение R^2, равное 0,75. Это означает, что с помощью нашей модели мы объяснили 75% вариаций цены продажи. Теперь подумайте о не включенных в эту модель признаках, которые помогли бы предсказать цену дома, — например о таких вещах, как экономические условия, процентные ставки, рейтинги начальных школ и так далее. Эти пропущенные переменные не только влияют на прогнозы модели, но и могут привести к сомнительным толкованиям. Вы заметили, что указанный в табл. 9.2 коэффициент, связанный с количеством ванных комнат, является отрицательным? Это не имеет никакого смысла.

Вот еще один пример. Рассмотрим модель линейной регрессии, которая выявила положительную корреляцию между размером обуви и количеством слов, которые человек может прочитать за минуту. Очевидно, что в данной модели отсутствует такая входная переменная, как возраст, включение которой позволило бы обойтись без входной переменной «размер обуви».

Разумеется, в своей работе вы редко будете сталкиваться со столь очевидными примерами, однако поверьте нам на слово: пропущенные переменные могут и будут порождать проблемы и неверные интерпретации. Кроме того, многие вещи коррелируют с такой часто опускаемой переменной, как время.

Мы надеемся, что при чтении этого раздела вы вспомните о том, что «корреляция не говорит о наличии причинно-следственной связи». Если одна переменная помечена в модели как входная, а другая — как выходная, это не означает, что входные данные обусловливают выходные.

Мультиколлинеарность

Если при использовании линейной регрессии вашей целью является интерпретируемость — возможность определить влияние входных переменных на выходные путем изучения коэффициентов — то вам следует знать о так называемой мультиколлинеарности. Мультиколлинеарность означает, что несколько переменных коррелируют друг с другом — и это создает проблему для интерпретируемости вашей модели.

Как вы помните, цель множественной регрессии — изолировать влияние одной из переменных при сохранении постоянных значений остальных входных переменных. Однако это возможно лишь в том случае, если данные являются некоррелированными.

Например, предположим, что данные о продажах лимонада, которые мы анализировали ранее, имеют температуру как в градусах Цельсия, так и в градусах Фаренгейта. Очевидно, что эти два показателя полностью коррелированы, поскольку один является функцией другого. Однако допустим, что показания температуры регистрируются с помощью разных приборов с целью внесения некоторой вариации[88]. В этом случае модель превратится:

1. *Из* Продажи = 1,03 (Температура) — 71,07.
2. *В* Продажи = –0,2(Температура в градусах Фаренгейта) + 2,1(Температура в градусах Цельсия) — 30,8.

Теперь похоже, что температура в градусах Фаренгейта отрицательно коррелирует с выходным параметром! Однако нам известно, что входные

[88] Модели линейной регрессии не вычисляются, если два входных параметра идеально коррелированы, поэтому мы добавили шум в данные в этом примере.

данные смешаны, даже избыточны, и линейная регрессия не в состоянии разрушить имеющуюся взаимосвязь. Мультиколлинеарность имеет место в большинстве наборов данных наблюдений, так что считайте это предупреждением. Что касается экспериментальных данных, то они обычно собираются таким образом, чтобы предотвратить мультиколлинеарность, насколько это возможно[89].

Утечка данных

Вернемся к построению модели для прогнозирования цены продажи дома. Но на этот раз допустим, что набор обучающих данных включает не только характеристики дома (площадь, количество спален и так далее), но и первоначально предложенную цену. Этот набор данных показан в табл. 9.3.

Запустив модель на этих данных, вы можете заметить, что начальное предложение очень хорошо предсказывает цену продажи. «Отлично!» — думаете вы и решаете положиться на это, чтобы спрогнозировать цены на жилье для своей компании.

Затем модель запускается в производство. При попытке использовать модель вы обнаруживаете, что у вас нет доступа к первоначальному предложению по домам, цены продажи которых вы пытаетесь предсказать, поскольку они еще не проданы! Это пример утечки данных[90], которая происходит, когда некая сопутствующая выходная переменная маскируется под входную.

Проблема с использованием первоначального предложения заключается во времени. Представьте, что вы можете узнать значение начального предложения только после фактической продажи дома.

Поскольку мы погружены в данные, нам очень трудно заметить подобные утечки. К сожалению, во многих учебниках этой проблеме не уделяется внимание, потому что там используются идеализированные наборы данных, тогда как в реальных наборах вероятность утечки есть всегда. Как главный по данным, вы должны следить за тем, чтобы ваши входные и выходные параметры не содержали перекрывающуюся информацию.

Мы вернемся к обсуждению проблемы утечки данных в следующих главах.

---

[89] Этой идее посвящена целая область статистики под названием «Планирование экспериментов».

[90] **https://en.wikipedia.org/wiki/Leakage_(machine_learning)**

Табл. 9.3. Выборка данных о домах

Площадь в квадратных футах	Количество спален	Количество ванных комнат	Первоначальное предложение, $	Цена продажи, $
1500	2	1	190 000	200 000
2000	3	2	240 000	250 000
2500	4	3	300 000	300 000

Ошибки экстраполяции

Экстраполяция — это прогнозирование значения за пределами диапазона входных данных, использованных для построения модели. В случае торговли лимонадом при температуре 0 °F модель предсказала бы объем продаж на уровне –71,07 доллара. Если бы в доме не было ни одного квадратного фута площади и ни одной ванной комнаты (то есть если бы дома фактически не существовало), модель предсказала бы цену продажи в –1 614 841,60 доллара. Оба значения не имеют смысла.

Модели делают прогнозы за пределами диапазона данных, на которых они «учились». В отличие от людей, уравнения не имеют здравого смысла, позволяющего понять, что их результаты неверны. Математические уравнения не способны думать. Если вы подставите в них числа в качестве входных данных, они выдадут вам некий численный результат. Именно вы как главный по данным должны понимать, что имеет место экстраполяция.

Следует подчеркнуть, что результаты модели всегда рассчитываются на основе конкретных данных. То есть вы не должны делать прогнозы на основе данных, которые «соответствуют» диапазону обучающих данных, но не соответствуют контексту, в котором эти данные были собраны, поскольку модель ничего не знает о происходящих в мире изменениях.

Если бы вы построили модель, предсказывающую цены на дома в 2007 году, она показала бы себя ужасно в 2008 году после обвала рынка жилья. Использование такой модели в 2008 году означало бы экстраполяцию данных о рыночных условиях 2007 года, которые весьма сильно отличались от условий 2008 года. В 2021 году, пока мы писали эту книгу, многие отрасли сталкивались с этой проблемой из-за пандемии COVID-19. Модели, обученные на данных, собранных до ее начала, больше не отражают многие из вновь возникших взаимосвязей, а значит, больше не актуальны.

Многие взаимосвязи не являются линейными

Линейная регрессия не годится для моделирования поведения фондового рынка, который на протяжении всей своей истории рос экспоненциально, а не линейно. Статистический отдел компании Procter & Gamble посоветовал бы «не подгонять прямую линию к кривой в форме банана».

В арсенале статистиков есть инструменты для преобразования некоторых нелинейных данных в линейные. Тем не менее иногда стоит просто признать тот факт, что линейная регрессия не подходит для решения стоящей перед вами задачи.

Вы объясняете или предсказываете?

В этой главе мы обсуждали две цели применения регрессионных моделей — объяснение взаимосвязей и прогнозирование. Судя по всему, модели линейной регрессии могут делать и то и другое. Коэффициенты модели линейной регрессии (при правильных условиях) обеспечивают интерпретируемость, которой уделяется большое внимание во многих отраслях — например, в клинических испытаниях, когда исследователи пытаются понять точную величину и направление влияния входного параметра (дозировка лекарства) на выходной (кровяное давление). В данном случае необходимо проявлять большую осторожность, чтобы избежать негативного влияния мультиколлинеарности и пропущенных переменных на объяснительную способность модели.

В других областях, таких как машинное обучение, целью является точное предсказание[91]. Например, наличие мультиколлинеарности может не представлять проблемы, если модель способна хорошо предсказывать будущие результаты. Когда цель модели — предсказать новые выходные данные, вы должны быть очень осторожны, чтобы избежать так называемого переобучения.

Как вы помните, модели — это упрощенные версии реальности. Хорошая модель хорошо аппроксимирует взаимосвязи между входными и выходными данными. По сути, она регистрирует некое скрытое явление, выражением которого и являются данные.

---

[91] Разница между объяснением и предсказанием с помощью моделей подробно описана в статье: Shmueli, G. (2010). To explain or to predict? *Statistical science, 25*(3), 289–310.

Однако переобученная модель фиксирует не взаимосвязь, существование которой мы предполагаем, а взаимодействие обучающих данных со всем присутствующим в них шумом и вариациями. Поэтому ее прогнозы — не результат моделирования, а просто некий набор точек данных, которые у нас уже есть.

По сути, переобученные модели запоминают выборочные обучающие данные и плохо обобщают новые наблюдения. Посмотрите на рис. 9.4. Слева вы видите данные о продажах лимонада с моделью линейной регрессии. Справа представлена сложная регрессионная модель, которая прекрасно предсказывает некоторые точки. Какую из них вы хотели бы использовать для прогнозирования?

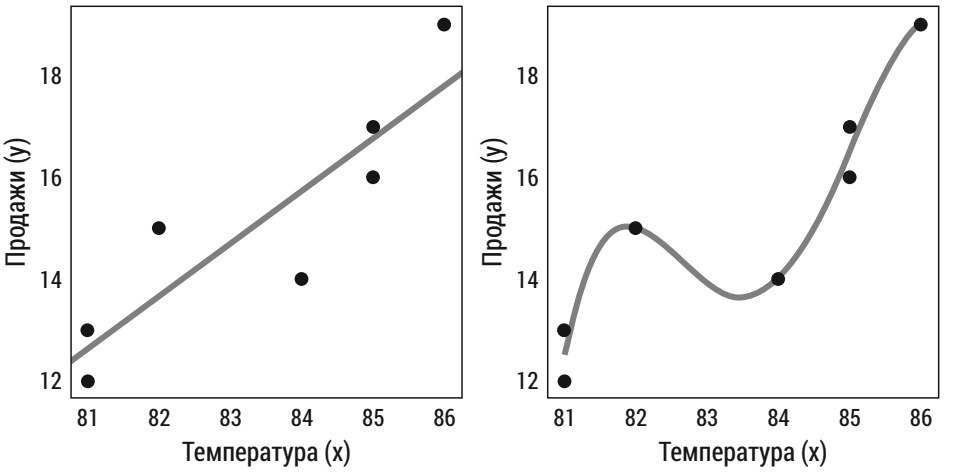


Рис. 9.4. Две конкурирующие модели. Модель слева хорошо обобщает, а переобученная модель справа, по сути, просто запоминает данные. Из-за вариаций модель справа не сможет хорошо предсказывать новые точки

Для предотвращения переобучения можно разделить набор данных на две части: обучающий набор, используемый для построения модели, и тестовый набор, позволяющий оценить ее эффективность. Производительность модели на тестовом наборе данных, на которых она не училась, позволяет оценить ее предсказательную способность.

Производительность регрессионной модели

Сталкиваясь с регрессионной моделью на работе, будь то модель множественной линейной регрессии или что-то более изощренное, вы можете

оценить ее соответствие вашим данным с помощью графика сравнения фактического и прогнозируемого значений. Некоторые полагают, что мы не можем визуализировать производительность модели регрессии при наличии слишком большого количества входных данных — однако помните о том, что сделала модель. Она преобразовала входные параметры (один или несколько) в выходные.

Итак, для каждой строки в наборе данных у вас есть фактическое значение и связанное с ним прогнозируемое значение. Создайте на их основе диаграмму рассеяния. Они должны быть хорошо коррелированы. Такой визуальный тест позволит вам быстро оценить эффективность вашей модели. Ваши специалисты по работе с данными могут предоставить несколько связанных показателей (одним из которых является R-квадрат), однако не стоит смотреть только на эти цифры. Всегда, всегда требуйте предоставить вам график сравнения фактических и прогнозируемых значений. Пример с использованием модели, построенной на основе данных о жилье, представлен на рис. 9.5.

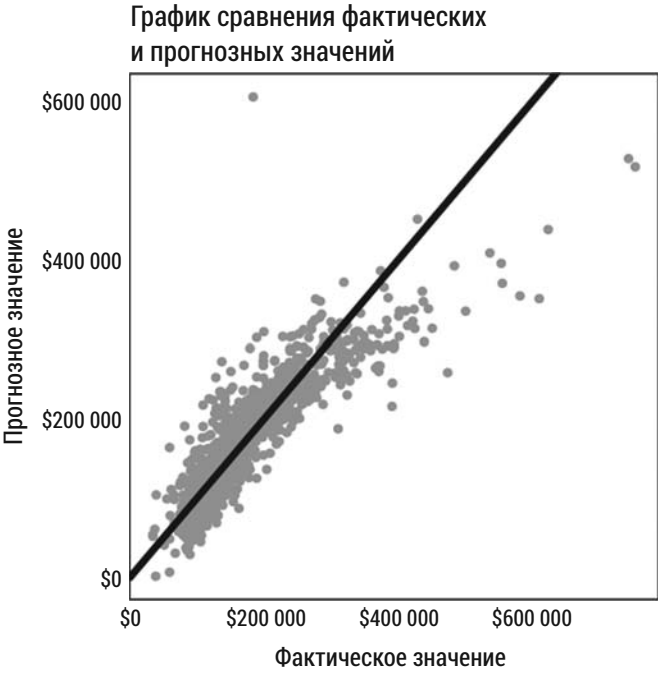


Рис. 9.5. На этом графике вы видите, что модель плохо прогнозирует цены на самые дорогие дома. Можете ли вы определить другие недостатки этой модели, используя этот график?

ПРОЧИЕ МОДЕЛИ РЕГРЕССИИ

Вы также можете столкнуться с такими вариантами линейной регрессии, как LASSO и Ridge (ридж-регрессия, или гребневая регрессия), которые могут помочь при наличии большого количества коррелированных входных данных (мультиколлинеарность) или в тех случаях, когда количество входных переменных превышает количество строк в вашем наборе данных. В результате получается модель, похожая на модели множественной регрессии.

Другие модели регрессии выглядят совершенно иначе. Метод *k*-ближайших соседей, обсуждавшийся во введении, применялся к задаче классификации, но его также можно применить и к задаче регрессии. Например, для предсказания цены продажи любого дома мы могли бы взять среднюю цену продажи трех ближайших к нему недавно проданных домов. В этом случае для решения задачи регрессии использовался бы метод *k*-ближайших соседей.

В следующей главе мы рассмотрим некоторые из этих моделей, поскольку их можно использовать для решения задач как классификации, так и регрессии.

ПОДВЕДЕНИЕ ИТОГОВ

Наша цель в этой главе состояла в том, чтобы помочь вам развить интуитивное понимание принципа контролируемого обучения и его фундаментального алгоритма — линейной регрессии. Мы также рассмотрели множество причин, по которым регрессионные модели могут ошибаться. Помните об этих проблемах и подводных камнях, потому что вопрос не в том, какие из них повлияют на ваши регрессионные модели, а в том, сколько это сделают.

Как вы, вероятно, уже догадались, обучающие данные обусловливают как мощь, так и ограниченность контролируемого обучения. К сожалению, зачастую компании уделяют больше внимания новейшим алгоритмам контролируемого обучения, чем сбору актуальных, точных и достаточных данных для подачи на вход этих алгоритмов. Пожалуйста, помните мантру «мусор на входе, мусор на выходе». Хорошие данные — ключевое условие эффективности моделей контролируемого обучения.

Если вы понимаете принцип обучения без учителя и обучения с учителем, то вы понимаете суть машинного обучения. Поздравляем! И извините за отсутствие больших открытий. Мы стремились преподать вам основы машинного обучения, обойдясь без громких рекламных слоганов и лишней шумихи. Машинное обучение предполагает как обучение без учителя, так и обучение с учителем.

Мы продолжим разговор о машинном обучении в следующей главе в контексте обсуждения моделей классификации.

ГЛАВА 10

Освойте модели классификации

Алгоритм машинного обучения заходит в бар.
Бармен спрашивает: «Что будете?»
Алгоритм отвечает: «А что заказали остальные?»

_ *Чет Хаасе (@chethaase)*

В предыдущей главе мы говорили о контролируемом обучении с помощью моделей регрессии, которые позволяют предсказывать численные значения (вроде объема продаж) путем подгонки модели к набору признаков. Но что, если вам требуется предсказать конкретный результат? Например, захочет ли человек, обладающий определенным набором демографических характеристик, купить книгу о данных? Если вы когда-нибудь задавались вопросом о том, как компании оценивают вероятность того, что вы щелкнете по тому или иному рекламному объявлению, купите продукт (и какой именно), не сможете выплатить кредит, взятый на покупку автомобиля, пройдете собеседование или чем-нибудь заболеете, то эта глава для вас.

В таких задачах, где нужно предсказать категориальную переменную (то есть метку), необходимо использовать модели классификации.

ВВЕДЕНИЕ В КЛАССИФИКАЦИЮ

Модели, предсказывающие два варианта, называются моделями бинарной классификации. Модели, используемые для предсказания множества

классов, называются моделями многоклассовой классификации[92]. Оценка вероятности того, что человек не сможет погасить автокредит, — пример задачи бинарной классификации (да/нет), а предсказание того, какую машину купит человек («Honda», «Toyota», «Ford» и так далее), — пример задачи многоклассовой классификации. Чтобы не усложнять, мы сосредоточимся на задачах бинарной классификации. Просто имейте в виду, что дополнительные классы логично продолжают те темы, которые мы обсудим в этой главе.

Результаты применения некоторых моделей классификации часто делятся на «положительные» или «отрицательные». Как вы помните, научная проверка осуществляется в форме утверждения, а значит, вы должны истолковывать положительное и отрицательное наблюдение как означающее «да» и «нет» соответственно. Это позволяет отделить наблюдения, демонстрирующие действие (щелчок по кнопке, покупка товара, дефолт по кредиту, наличие заболевания), от тех, которые этого не делают. В других случаях — например, при предсказании принадлежности избирателя к политической партии — вам следует четко определить, какой класс является «положительным», а какой «отрицательным», чтобы избежать путаницы. Например, то, что вы определяете принадлежность избирателя к демократам или республиканцам как положительную или отрицательную в своей модели, — не оценка этой принадлежности, а ее произвольное обозначение. Как главный по данным вы должны убедиться в том, что все члены команды одинаково понимают используемые в модели обозначения.

Чему вы научитесь

В этой главе мы будем использовать набор данных о человеческих ресурсах для описания следующих моделей классификации:

— логистическая регрессия;
— деревья решений;
— ансамблевые методы.

[92] Не путайте кластеризацию с классификацией. Помните о том, что кластеризация не предполагает использование меток. При кластеризации если метки и присваиваются, то самим аналитиком и только впоследствии. При решении задач классификации метки изначально присутствуют в наборе данных.

Логистическая регрессия[93] и деревья решений чаще всего изучаются в рамках курсов по науке о данных и широко используются в программном обеспечении. Простота применения и интерпретируемость делают их идеальным выбором для решения некоторых задач. Однако, как и все прочие алгоритмы, описанные в этой книге, они не лишены недостатков.

Мы также познакомим вас с ансамблевыми методами, которые постепенно становятся новым стандартом для специалистов по работе данных, особенно для участников соответствующих соревнований[94].

Во второй половине этой главы мы более подробно рассмотрим проблему утечки данных и переобучения. Обсуждению точности мы посвятим целый раздел в конце главы, поскольку понимание этого термина (в контексте данных) требует рассмотрения ряда нюансов. А мы хотим уберечь вас от самых распространенных ошибок.

Постановка задачи классификации

Представьте, что каждое лето сотни студентов, изучающих науку о данных, подают заявки на стажировку в вашей компании. Просматривать все эти заявки вручную крайне утомительно. Нельзя ли как-то автоматизировать этот процесс?

К счастью, у вашей компании есть набор исторических данных, которые можно использовать для обучения модели, — информация о каждом соискателе и метка «да/нет», говорящая о том, был ли он приглашен на собеседование. Используя исторические данные и такой инструмент, как логистическая регрессия, вы могли бы разработать прогностическую модель, которая использует содержащуюся в заявке информацию в качестве входных данных, например, средний балл, год обучения, специализация, количество внеклассных занятий, и сообщает о том, стоит ли предлагать соискателю пройти собеседование. Если она окажется эффективной, это избавит вас от необходимости просматривать резюме вручную.

Как можно решить эту задачу? Для начала познакомимся с логистической регрессией.

---

[93] Логистическая регрессия, как вы узнаете далее, предсказывает вероятности. При добавлении решающего правила она превращается в алгоритм классификации.

[94] Описанные в этой главе деревья решений и ансамблевые методы можно использовать для решения задач регрессии. Так что, если выходной параметр вашего набора данных является числом, попробуйте их применить.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Начнем с простого и рассмотрим данные из десяти прошлых заявок, используя в качестве входных параметров только средний балл. Поскольку компьютеры понимают лишь числа, вы можете преобразовать ответы «да» и «нет» в значения 1 и 0 соответственно, то есть 1 обозначает положительный класс. Данные представлены в табл. 10.1. Общая тенденция неудивительна: студенты с более высоким средним баллом имеют больше шансов получить приглашение на собеседование.

Табл. 10.1. Простой набор данных для логистической регрессии: использование среднего балла для прогнозирования вероятности приглашения на собеседование

Идентификатор заявки	Средний балл	Приглашение (да/нет)	Приглашение (1/0)
1	2,00	нет	0
2	2,20	нет	0
3	2,50	нет	0
4	2,80	да	1
5	2,85	нет	0
6	3,50	да	1
7	3,60	нет	0
8	3,70	да	1
9	3,80	да	1
10	4,00	да	1

Если бы вы попытались применить к этим данным метод линейной регрессии, описанный в предыдущей главе, вы получили бы весьма странные результаты. Например, если мы введем данные из табл. 10.1 в статистическую программу и сгенерируем регрессионную модель, то получим уравнение следующего вида:

$$\text{Приглашение} = (0{,}5) \times \text{Средний балл} - 1{,}1$$

Однако давайте задумаемся об этой модели. Предположим, что средний балл нового соискателя составляет 2,0. В этом случае регрессионная модель выдала бы результат: $\text{Приглашение} = (0{,}5) \times (2{,}0) - 1{,}1 = -0{,}1$. А если бы

средний балл кандидата составлял 4,0, то результат был бы равен 0,9. Но что означают числа –0,1 и 0,9 в контексте предсказания того, получит ли кандидат приглашение на собеседование? (Мы тоже точно не знаем.)

Что могло бы оказаться полезным, так это прогноз вероятности получения такого приглашения на основе среднего балла соискателя. Например, вы знаете, что для соискателей со средним баллом 2,0 вероятность получения приглашения на собеседование составляет 4%, а для соискателей со средним баллом 4,0 — 92%. Эта информация имеет отношение к поставленной задаче, поскольку позволяет вам ввести правила классификации будущих кандидатов. Однако помните о том, что значения вероятности должны находиться в пределах от 0 до 1 включительно, а модели регрессии не работают в рамках этих ограничений и могут выдавать абсолютно любое значение. Поэтому линейная регрессия не является оптимальным методом для решения данной задачи.

Таким образом, вам нужно как-то ограничить результат решения уравнения вида $y = mx + b$, чтобы гарантировать его нахождение в подходящем диапазоне вероятностей. Именно это и делает логистическая регрессия: она «втискивает» выходные данные в диапазон от 0 до 1, предоставляя пользователю вероятность принадлежности результата к положительному классу (в данном случае: приглашение = «да»).

Рассмотрим уравнение логистической регрессии:

Вероятность принадлежности к положительному классу при условии

$$x = \frac{1}{1 + e^{-(mx+b)}} \tag{1}$$

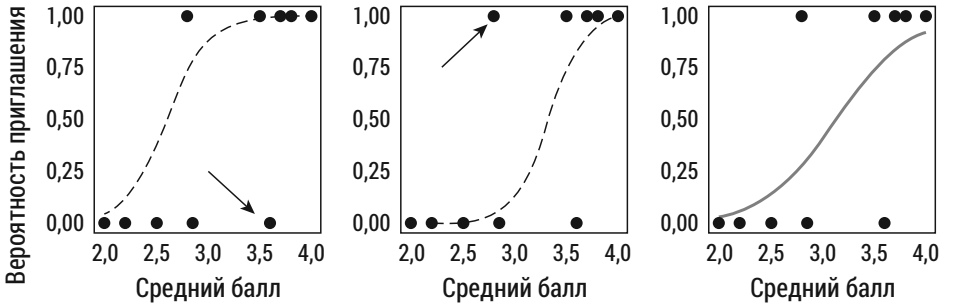


Рис. 10.1. Подгонка различных моделей логистической регрессии к данным. Модель справа соответствует им лучше всего

Вам наверняка уже знаком фрагмент $mx+b$, поскольку это формула линейной регрессии. Только теперь она является частью уравнения, называемого логистической функцией (отсюда и название логистической регрессии)[95], которая гарантирует то, что полученное число является значением вероятности.

Для большей ясности рассмотрим несколько графиков. На рис. 10.1 представлены три диаграммы рассеяния, построенные на основе данных из табл. 10.1. (В предыдущей главе при построении «линии наилучшего соответствия» мы получили похожий набор из трех графиков.) Каждый из этих графиков отражает разный набор входных значений для m и b в уравнении (1). Напомним, что в случае линейной регрессии значения m и b модулировали оптимальное положение линии, минимизирующее величину ошибки, выражаемую суммой квадратов. Но мы установили, что прямая линия линейной регрессии не может хорошо соответствовать этим данным, поскольку выходит за пределы 0 слева и за пределы 1 справа. Однако уравнение (1) вне зависимости от значений m и b всегда будет давать S-образную кривую, лежащую в диапазоне от 0 до 1.

Проанализируйте левый и средний графики на рис. 10.1 и определите их слабые места. На левом графике пунктирной линией показана модель, которая слишком уверенно предсказывает то, что высокий средний балл приведет к приглашению на собеседование, упуская при этом кандидата со средним баллом 3,5, который это приглашение не получил. Модель, показанная на среднем графике, выдает неоправданно низкую вероятность для студентов с низким средним баллом. Согласно ей, студент со средним баллом в 2,8, которого пригласили на собеседование, имел на это почти нулевой шанс. Крайний правый график на рис. 10.1 может похвастаться оптимальным балансом. Этот результат применения алгоритма логистической регрессии наилучшим образом уравновешивает левую и среднюю диаграммы и с математической точки зрения является оптимизированным решением для имеющихся точек данных. Полученная в результате модель логистической регрессии имеет следующее уравнение:

$$\text{Вероятность получения приглашения при данном среднем балле} = \frac{1}{1+e^{-(2{,}9*\text{Ср.балл}-9{,}0)}}. \quad (2)$$

[95] Число e в уравнении — математическая константа вроде π, которая применяется далеко не только в логистической регрессии. Это так называемая постоянная Эйлера, приблизительно равная 2,71828.

Логистическая регрессия уменьшает так называемую логистическую функцию потерь, которая представляет собой способ измерения степени близости предсказанных вероятностей к фактическим меткам. Хотя линейная и логистическая регрессии используют разные методы, их цель одна и та же — максимально приблизить совокупность предсказанных моделью значений к фактическим.

Логистическая регрессия: что дальше?

Логистическая регрессия дает два преимущества: мы получаем формулу, которая помогает делать прогнозы на основе данных, а коэффициенты этой формулы объясняют взаимосвязи между входными и выходными параметрами.

Применить ее можно следующим образом. На рис. 10.2 показана вероятность приглашения на собеседование для студента со средним баллом 2,0, согласно нашей модели логистической регрессии. Шанс получить такое приглашение для этого человека составляет около 4%. Кандидат, повышающий свой средний балл с 2,0 до 3,0, повышает вероятность получения приглашения на собеседование с 4 до 41%, то есть разница составляет 37%. Однако увеличение среднего балла еще на одну единицу, с 3,0 до 4,0, повышает вероятность с 41 до 92%; здесь разница составляет целых 51%! Обратите внимание на то, что при использовании моделей логистической регрессии влияние дополнительного балла на вероятность приглашения не является постоянным. В этом заключается еще одно отличие логистической регрессии от линейной: в случае линейной регрессии увеличение входной переменной на одну единицу всегда одинаково влияет на результат, каким бы ни было начальное значение.

Сама по себе логистическая регрессия не скажет вам, следует ли пригласить на собеседование того или иного человека или нет. Скорее она сообщает вам вероятность такого приглашения. Если вы хотите автоматизировать процесс принятия решений с помощью логистической регрессии, вам необходимо задать точку отсечения (пороговое значение), также известное как решающее правило; оно определяет реализацию того, чему научилась ваша модель. Если вы зададите точку отсечения на отметке 90%, то есть будете рассматривать только те заявки, средний балл в которых предполагает 90%-ную вероятность приглашения на собеседование, то, скорее всего, сделаете меньше предложений. С другой стороны, если вы готовы рассматривать заявки соискателей, шанс на приглашение которых, исходя из прошлых

данных, составляет 60%, то увидите гораздо больше кандидатов. Задание точек отсечения требует участия экспертов в предметной области.

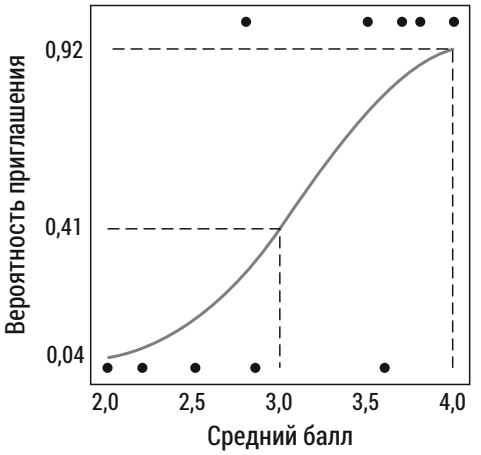


Рис. 10.2. Применение модели логистической регрессии для прогнозирования вероятности приглашения при среднем балле равном 2, 3 и 4

Как говорилось ранее, коэффициент любой регрессионной функции говорит о взаимосвязях между входными и выходными данными. С первого взгляда понятно, что значение коэффициента для среднего балла в уравнении (2) является положительным и составляет 2,9. Это говорит о том, что более высокий средний балл повышает шансы человека на получение приглашения. В данном случае это не столь уж сногсшибательная новость, однако для исследователей, предсказывающих вероятность развития рака на основе определенных биомаркеров, это может иметь большое значение[96].

На что следует обратить внимание при работе с логистической регрессией

Моделям логистической регрессии свойственны те же проблемы, что и моделям линейной регрессии, которые мы подробно рассмотрели в предыдущей главе, а именно:

[96] Чтобы по-настоящему понять эту формулу, необходимо познакомиться с концепцией логарифма отношения шансов, рассмотрение которой выходит за рамки данной книги.

- Пропущенные переменные. Алгоритм не может учиться на данных, которых нет.
- Мультиколлинеарность. Коррелированные входные признаки могут сильно исказить вашу интерпретацию коэффициентов модели, а иногда даже сделать положительный коэффициент отрицательным (или наоборот).
- Экстраполяция. В случае с логистической регрессией проблема с экстраполяцией стоит не столь остро, как в случае с линейной, потому что ее выходные данные всегда находятся в пределах диапазона от 0 до 1. Однако расслабляться все-таки не следует. Предсказание значений за пределами диапазона обучающих данных может привести к чрезмерно уверенным оценкам вероятностей, поскольку эти прогнозные значения асимптотически приближаются к единице.

Разумеется, при использовании логистической регрессии следует избегать и других ошибок, которые мы обсудим в конце главы.

ДЕРЕВЬЯ РЕШЕНИЙ

Некоторых людей отталкивает (и, возможно, пугает) математика, связанная с использованием логистической регрессии. Кроме того, далеко не каждую взаимосвязь между входными и выходными данными можно описать с помощью линейной модели $y = mx + b$. Альтернативный, более понятный и простой для визуализации подход — дерево решений. Деревья решений разбивают набор данных на несколько частей, предоставляя список правил наподобие блок-схемы, которыми можно руководствоваться при прогнозировании.

Возьмем, к примеру, набор данных, приведенных в табл. 10.2. Здесь вы видите выборку данных о десяти студентах (из 300), которые подали заявку и были приглашены на собеседование в вашу компанию. Вместо того чтобы использовать средний балл в качестве единственного входного параметра для своей модели, вы решаете проанализировать все признаки, чтобы выяснить, как приглашения на интервью делались в прошлом. Обратите внимание на то, что в этом наборе данных на собеседование были приглашены 120 студентов (то есть 40%).

Табл. 10.2. Фрагмент набора данных о стажерах. Специализации студентов таковы: Инф. = Информатика, Экон. = Экономика, Стат. = Статистика и Биз. = Бизнес.

ID студента	Средний балл	Курс	Специализация	Количество внеклассных занятий	Приглашение?
1	3,41	1	Инф.	1	Нет
2	3,33	3	Экон.	2	Нет
3	2,96	3	Инф.	5	Да
4	3,28	2	Стат.	4	Да
5	2,78	2	Инф.	3	Нет
6	3,01	4	Экон.	0	Нет
7	2,56	3	Стат.	2	Нет
8	2,72	3	Инф.	4	Да
9	2,00	3	Стат.	2	Нет
10	2,42	1	Биз.	3	Нет
⋮	⋮	⋮	⋮	⋮	⋮

Если вы хотите использовать эти признаки, чтобы понять, кто получил приглашение, а кто нет, вы можете самостоятельно вывести несколько правил. Например, студенты с высоким средним баллом, участвующие во внеклассных занятиях, вероятно, имеют больше шансов получить приглашение. Но какой средний балл вы использовали бы для «разделения» совокупности студентов? 3,0? 3,5? И с помощью какой информации вы бы обосновали свое решение? Как вы уже, вероятно, поняли, самостоятельное выведение правил — чрезвычайно сложная задача. К счастью, алгоритм для создания дерева решений может позаботиться об этом за вас. Он ищет входной признак и его значение, которое наилучшим образом отличает студентов, получивших приглашение на интервью, от тех, кто его не получил. Затем он находит следующий признак, позволяющий разделить уже эти две группы и так далее.

Мы прогнали наш набор данных через алгоритм под названием CART[97] и сгенерировали дерево решений, изображенное на рис. 10.3. Оно больше похоже на перевернутое дерево, состоящее из «узлов», «ветвей» и «листьев»

[97] Существует несколько алгоритмов для создания деревьев решений, но наиболее популярный из них — CART (Classification and Regression Trees, деревья классификации и регрессии). Подробную информацию о нем можно найти в работе Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

решений, в котором окончательный прогноз определяется листом. Давайте обойдем это дерево, чтобы разобраться в том, как оно работает.

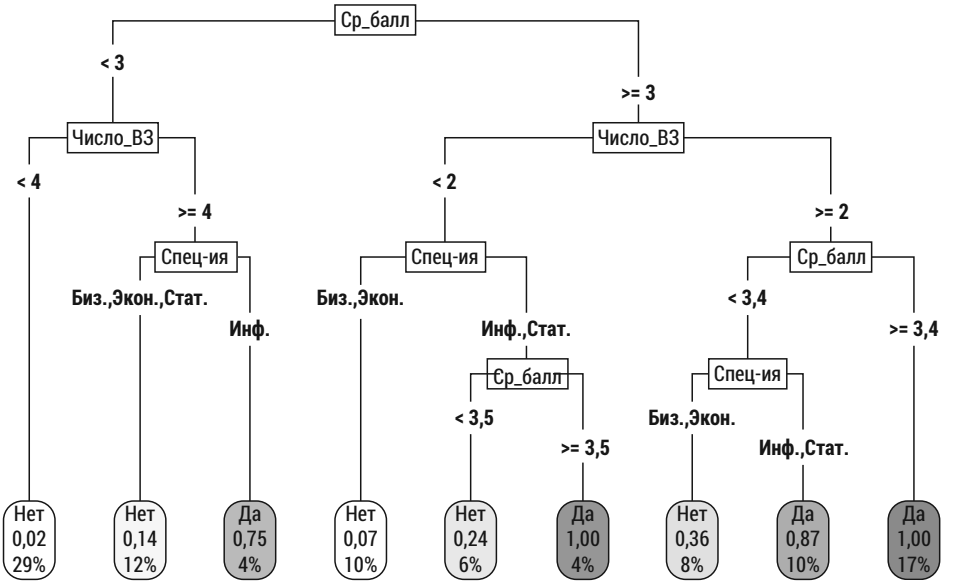


Рис. 10.3. Простой алгоритм дерева решений, примененный к набору данных о стажерах

Предположим, что соискатель по имени Эллен учится на втором курсе, имеет средний балл 3,6, специализируется на изучении информатики и занимается спортом. Эти данные можно закодировать следующим образом: {Ср_балл = 3,6, Курс = 2, Специальность = Инф., Число_ВЗ = 1}, где Число_ВЗ означает «число внеклассных занятий».

В самом верху рис. 10.3 находится корневой узел, указывающий на признак, позволяющий лучше всего разделить данные; здесь это средний балл. Средний балл Эллен составляет 3,6, поэтому она переходит в правую ветвь к следующему узлу принятия решений: Число_ВЗ. Ее Число_ВЗ равно 1, поэтому она переходит в левую ветвь к следующему узлу принятия решений: Специальность. Студенты, специализирующиеся на изучении информатики, переходят направо, и тут в дело снова вступает такой признак, как средний балл. Средний балл Эллен составляет не менее 3,5, поэтому вы можете предсказать «Да», она будет приглашена на собеседование.

Обратите внимание на то, как это дерево раскрывает взаимодействия входных признаков. В данном случае небольшое количество внеклассных

занятий компенсируется наличием высокого среднего балла по информатике или статистике.

Числа на листьях в нижней части рис. 10.3 подытоживают разделение обучающих данных, произведенное деревом решений. Крайний справа лист имеет три точки данных: {Да, 1,00 и 17%}. Это говорит о том, что в прошлом 100% соискателей со средним баллом ≥ 3,4 и как минимум тремя внеклассными занятиями были приглашены на собеседование вне зависимости от своей специализации. (Вы можете это увидеть, если проследите путь от этого листа до корневого узла.) Для любого нового кандидата, соответствующего этим критериям, прогнозным значением будет «Да», потому что в прошлом процент соискателей, попадающих в этот лист, превышал 50%. В данном случае речь идет о 51 соискателе, на которых приходилось 17% обучающих данных.

Крайний левый лист показывает, что 29% бывших претендентов имели средний балл ниже 3,0 и менее 4 внеклассных занятий, и из них только 2% были приглашены на собеседование. Поэтому данный узел содержит прогноз «Нет»[98].

Деревья решений отлично подходят для отображения разведочных данных и позволяют легко и быстро убедиться в том, что входные данные в вашем наборе связаны с выходными.

Однако одного дерева редко бывает достаточно для прогнозирования. Давайте посмотрим, как одно дерево решений (вроде изображенного на рис. 10.3) может ввести в заблуждение. С одной стороны, дерево может продолжать расти вглубь до тех пор, пока каждый кандидат не окажется в своем отдельном листе, что будет представлять идеальные правила принятия решений для всех 300 претендентов из набора обучающих данных. И если следующая группа стажеров будет иметь точно такие же характеристики, то ваше дерево будет идеальным. Однако, учитывая то, что в данных всегда присутствуют вариации, здравый смысл подсказывает нам, что это невозможно. Новые кандидаты будут отличаться от тех, на которых мы учились, а переобученное дерево будет очень уверенно предлагать вам потенциально неверные решения.

Действительно, одиночные деревья решений склонны к переобучению, при котором модель описывает набор обучающих данных гораздо лучше,

[98] Мы создали это дерево и его визуализацию с помощью (бесплатной) статистической программы R с открытым исходным кодом и пакетов «rpart» и «rpart.plot». Не все деревья решений, с которыми вы столкнетесь, будут иметь подобный уровень детализации.

чем ту реальность, для предсказания которой она была создана. Один из способов устранения этой проблемы — так называемая обрезка, однако одиночные деревья остаются весьма чувствительными к своим обучающим данным. Если бы вы отобрали 100 кандидатов из своего набора данных и построили новое дерево решений, то, вероятно, обнаружили бы другие узлы решений и разделительные значения внутри дерева. Например, для разделения корневого узла может использоваться значение среднего балла 3,2 вместо 3,0.

Для исправления проблем, свойственных деревьям решений, можно использовать ансамблевые методы.

АНСАМБЛЕВЫЕ МЕТОДЫ

Ансамблевые методы, предполагающие агрегирование результатов десятков, а то и тысяч запусков алгоритма, пользуются популярностью среди специалистов по работе с данными благодаря своей способности делать значимые прогнозы на детальном уровне.

Абсолютных фаворитов у дата-сайентистов два — случайные леса и деревья решений с градиентным усилением. Они часто используются командами, побеждающими в соревнованиях по науке о данных, проводимых на веб-сайте **Kaggle.com**; компании размещают наборы данных и вручают дата-сайентистам, создавшим на их основе максимально точные модели, солидные денежные призы. В этом разделе мы предоставим вам краткое интуитивно понятное объяснение этих методов.

Случайные леса

Если вы понаблюдаете за любыми двумя опытными интервьюерами, то заметите, что каждый из них использует собственные правила принятия решений, основанные на их личном опыте и типах кандидатов, с которыми они взаимодействовали. Проще говоря, они оценивают кандидатов по-разному. Вот почему во многих компаниях за отбор новых сотрудников отвечают целые команды, а решение принимается на основе консенсуса, позволяющего сбалансировать различия в оценках нескольких человек.

Случайный лес[99] — это эквивалент данной идеи в виде дерева решений. Этот алгоритм берет случайную выборку данных и строит дерево решений,

[99] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

а затем повторяет этот процесс еще несколько сотен раз[100]. В результате получается «лес», состоящий из деревьев, которые имитируют множество независимых оценщиков вашего набора данных, а окончательный прогноз — это консенсус, то есть решение, принятое большинством голосов. (Случайные леса также могут использоваться для расчета среднего значения предсказанных вероятностей при решении задач классификации или среднее значение непрерывных числовых данных при решении задач регрессии.)

На рис. 10.4 показаны четыре дерева в нашем лесу. Присмотритесь, и вы заметите еще одну особенность случайных лесов. В двух деревьях первым разделителем является средний балл, в одном — специализация, еще в одном — количество внеклассных занятий. Так и должно быть. Случайные леса случайным образом выбирают не только наблюдения (строки) для построения дерева, но и признаки (столбцы). Это устраняет корреляцию составляющих лес деревьев, позволяя каждому из них находить новые взаимосвязи в данных. В противном случае найденная деревьями информация оказалась бы избыточной.

Деревья решений с градиентным усилением

Деревья решений с градиентным усилением[101] используют другой подход. В то время как случайный лес создает сотни отдельных деревьев и в конце усредняет их результаты, деревья с градиентным усилением строятся последовательно.

В ситуации приема на работу это означает, что несколько интервьюеров выстраиваются в очередь за дверью, чтобы последовательно побеседовать с кандидатом. Каждый интервьюер входит в комнату, задает кандидату один-два вопроса, выходит и говорит следующему интервьюеру что-то вроде: «На данный момент я склоняюсь к приему этого человека на работу, но нам нужно задать больше наводящих вопросов, касающихся таких-то областей», и так далее. Результат — единая рекомендация, основанная на совокупности рекомендаций всей группы, а не множество отдельных рекомендаций, объединенных в одну.

---

[100] Построение моделей на основе случайных выборок данных называется «бэггингом». Случайные леса — один из вариантов применения данного метода.

[101] Дополнительную информацию о градиентном усилении (бустинге) можно найти в главе 10 книги Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics, и в указанных там источниках. Однако имейте в виду, что это довольно сложный текст.

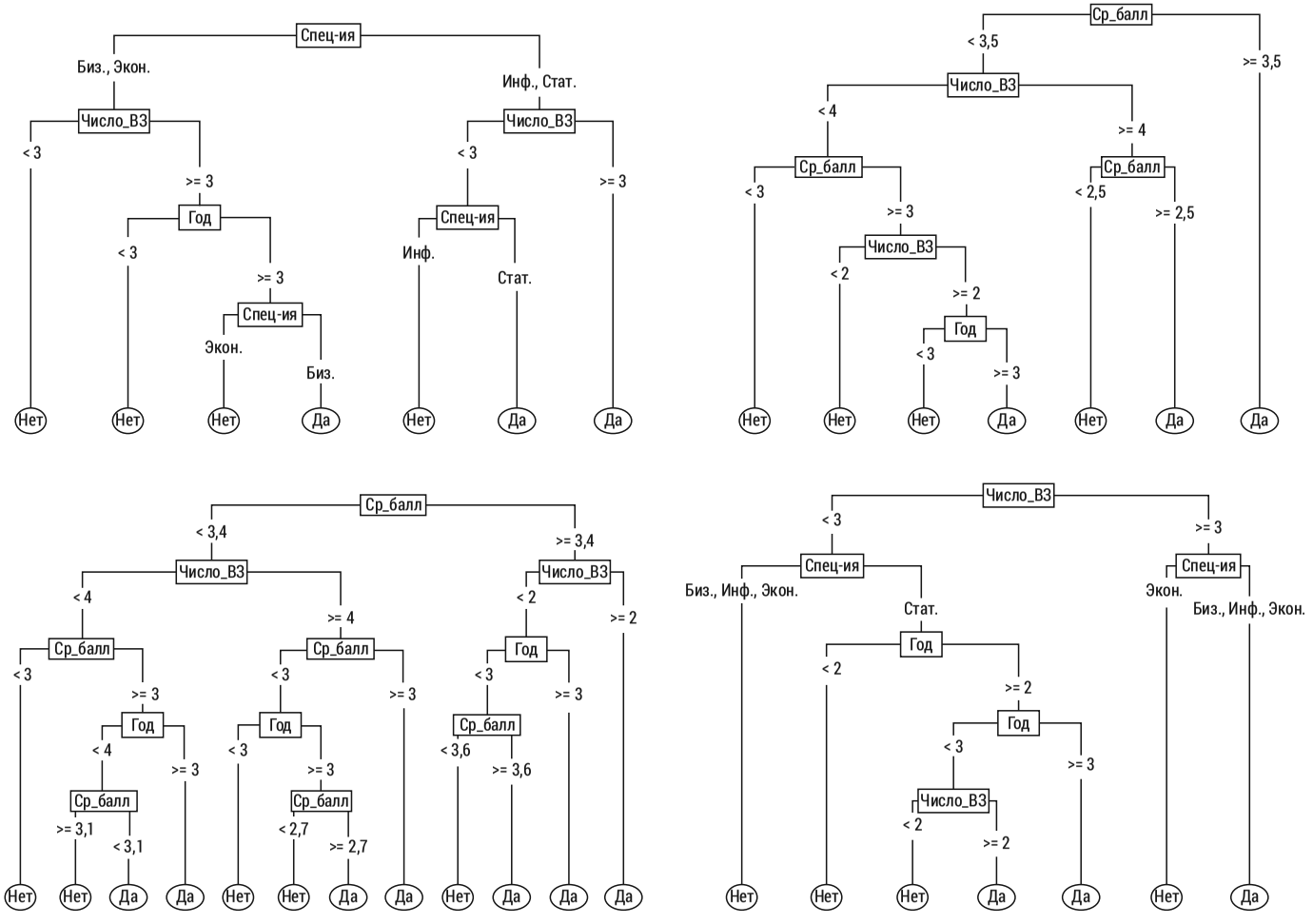


Рис. 10.4. Случайный лес — это «лес», состоящий из нескольких деревьев решений (как правило, сотен), в котором каждое дерево построено на основе случайного подмножества данных. Итоговый прогноз представляет собой консенсус всех составляющих лес деревьев

Как правило, деревья решений с градиентным усилением начинаются с построения так называемого неглубокого дерева с несколькими ветвями и узлами. Эта первая итерация по своей сути весьма наивна и не позволяет правильно разделить набор данных. На следующем этапе с учетом ошибок, допущенных первым деревом, строится новое дерево с усиленными наблюдениями, для которых были характерны особенно большие ошибки (в этом проявляется действие градиента). Для создания усиленной модели этот процесс выполняется тысячи раз с использованием больших наборов данных.

Как правило, эти ансамблевые методы не предназначены для анализа «небольших данных», поэтому специалистам следует применять их при работе с сотнями, а не десятками наблюдений.

Интерпретируемость ансамблевых моделей

Только представьте, как сложно было бы разобраться в тысячах листьев и узлов деревьев, правила которых чувствительны к мельчайшим изменениям в данных. Эти модели часто называют «черными ящиками», поскольку их внутреннее устройство очень трудно понять. При использовании случайных лесов и деревьев решений с градиентным усилением вместо логистической регрессии, выигрывая в точности, вы теряете в интерпретируемости. Это компромисс[102].

Другие модели типа «черный ящик» мы обсудим в главе 12.

ОСТЕРЕГАЙТЕСЬ ЛОВУШЕК

Какой бы мощью ни обладали модели классификации, при их неправильном применении можно угодить в несколько ловушек. Не заблуждайтесь: модели, которые имеют недостатки, описанные в следующих разделах, не могут быть «достаточно хорошими». Как главный по данным, вы должны хорошо понимать следующие потенциальные ловушки:

— неправильное определение типа задачи;
— утечка данных;
— отсутствие разделения данных;

[102] Хороший обзор можно найти в статье “Ideas on interpreting machine learning” на сайте **www.oreilly.com/radar/ideas-on-interpreting-machine-learning**. В настоящее время проводятся исследования, направленные на улучшение понимания работы этих методов.

— выбор неправильного порогового значения для принятия решения;
— неправильное понимание точности.

Последнее мы обсудим в следующем разделе.

Неправильное определение типа задачи

Совершенно очевидно, что если вы хотите предсказать категориальную переменную, вам не следует использовать линейную регрессию. Например, вспомните, что в табл. 10.1 мы заменили «да» и «нет» значениями 1 и 0 при постановке задачи, решаемой с помощью логистической регрессии.

Ваша статистическая программа не поправит вас, если вы неправильно примените линейную регрессию к этим данным. Она не знает, что ваши 1 и 0 означают «Да» и «Нет». Мы видели, как подобное происходит множество раз. Главным по данным следует иметь это в виду и незамедлительно исправлять ошибку.

Утечка данных

Что, если для построения классификационной модели вы собрали все возможные исторические данные о заявках соискателей на прохождение стажировки, в том числе о том, были ли они в конечном итоге наняты на работу или нет (0 означает «Нет», а 1 — «Да»). А затем вы применили логистическую регрессию, чтобы предсказать, получит ли предложение тот или иной кандидат.

Как вы думаете, что не так с использованием атрибута «нанят»?

Слово «нанят» означает, что соискатель согласился устроиться на постоянную работу после стажировки (вот и утечка). Только соискатели, получившие предложение пройти стажировку (что вы и пытаетесь предсказать), будут иметь входной параметр «Нанят?» = 1. Если значение атрибута «Нанят?» = 1, то значение целевого параметра «Пригласить?» также всегда должно быть равно 1. Данная модель абсолютно бесполезна, потому что обучена на данных, которые не могут быть доступны во время прогнозирования.

Критическое осмысление имеющихся данных и входных признаков в алгоритме контролируемого обучения — это то, что программа не может сделать за специалистов по работе с данными.

Отсутствие разделения данных

Если вы не разделите свои данные на набор для обучения и набор для тестирования, вы рискуете переобучить свою модель и получить ужасную производительность при анализе новых данных. Как правило, рекомендуется использовать 80% наблюдений в наборе данных для обучения модели, а остальные 20% — для тестирования ее производительности.

Янн ЛеКун, главный научный сотрудник Meta* по вопросам искусственного интеллекта, сформулировал это так: «Тестирование модели на обучающем наборе предается анафеме в мире машинного обучения, поскольку это величайший грех, который только можно совершить»[103]. Поэтому убедитесь в том, что вы тестируете свои модели на данных, с которыми они раньше не сталкивались. Если ваш алгоритм машинного обучения демонстрирует практически идеальные прогнозы, что возможно в случае деревьев решений с градиентным усилением — значит, ваша модель, скорее всего, переобучена.

Выбор неправильного порогового значения для принятия решения

Большинство моделей классификации выдают не метку, а вероятность принадлежности к положительному классу. Как вы помните, для студента со средним баллом 2,0 шанс получить приглашение на стажировку составлял 4%, а для студента со средним баллом 3,0 — 41%. Однако на основе этой информации нельзя ничего сделать до тех пор, пока не будет сформулировано решающее правило.

Именно здесь в игру вступаете вы. Выбор порогового значения вероятности для выполнения окончательной классификации — это решение, которое должен принять человек, а не машина. Многие программные пакеты в качестве такого значения по умолчанию выбирают 0,5 или 50%. Однако это значение не обязательно соответствует особенностям стоящей перед вами проблемы.

Не стоит относиться к его выбору легкомысленно. Для модели, которая определяет, кому стоит отправить по почте предложение кредитной карты, можно задать низкое пороговое значение (судя по нашим почтовым ящикам, так и происходит), тогда как модель, оценивающая претендентов на прохождение дорогостоящего лечения, может иметь высокое значение. Вам

* Признана экстремисткой на территории РФ.

[103] Цитата из поста.

необходимо учитывать эти компромиссы, которые сильно зависят от специфики вашей бизнес-задачи.

Теперь давайте поговорим о точности в контексте классификации и о том, что мы вообще подразумеваем под этим словом.

Неправильное понимание точности

Поскольку вы и остальные сотрудники вашей компании занимаетесь построением, развертыванием и поддержанием работы моделей классификации, предназначенных для автоматизации процесса принятия решений, вы должны уметь оценивать эти модели.

Ваша первая задача — сделать паузу и провести инвентаризацию исторических данных. Когда вы приступите к развертыванию своей модели, вам понадобится задать критерий для ее оценки, то есть создать «средство контроля». И это необходимо сделать для любой модели классификации, которую создаете вы или ваши специалисты по работе с данными. В случае бинарной классификации для этого достаточно определить долю класса большинства в наборе данных. В наборе данных о стажерах этим классом был «Нет», так как 60% кандидатов не получили предложения пройти стажировку (а 40% получили).

Теперь предположим, что кто-то из вашей команды применяет XGBoost (алгоритм градиентного усиления деревьев решений) к 80% данных (обучающий набор), и модель классификации предсказывает верные результаты в 60% случаев на оставшихся 20% данных (тестовый набор). Поскольку это больше, чем 50/50, такой результат может показаться вам вполне хорошим, так как в долгосрочной перспективе эта модель обещает работать лучше, чем подбрасывание монеты.

Однако на самом деле это указывает на то, что признаки в вашем наборе данных никак не связаны с выходными параметрами. Как в этом можно убедиться? Ну, если бы вы обратились к своему исходному набору данных, полностью проигнорировали входные параметры и попытались просто угадать класс большинства для каждого прогноза («Нет»), то вы оказались бы правы в 60% случаев! Так что алгоритм XGBoost ничем вам не помог. Метрика точности 60% в каком-то неточна, поскольку не превышает контрольный показатель.

Подумайте о событиях, которые случаются нечасто. Например, рекламное объявление в Интернете может быть показано тысячам пользователей,

но лишь несколько человек кликнут по нему. Мы бы назвали эти данные несбалансированными, поскольку слишком большую долю обучающего набора составляют объекты одного класса (большинство пользователей «не щелкнули» по объявлению). Если, например, 99,5% людей не щелкают по объявлению, то прогноз по умолчанию, говорящий о том, что никто никогда по нему не щелкнет, окажется верным в 99,5% случаев.

По этой причине вам не следует оценивать производительность алгоритма машинного обучения исключительно по критерию точности. Гораздо более эффективный способ оценки модели классификации — использование матрицы ошибок.

Матрицы ошибок

Матрица ошибок — это способ визуализации результатов модели классификации и определенного порога принятия решений. Представьте, что модель, построенная на основе алгоритма случайного леса, была обучена на 80% данных о стажерах (240 кандидатов) и протестирована на оставшихся 20% данных (60 кандидатов) с целью имитации процесса ее использования в реальном мире. Матрица ошибок, приведенная в табл. 10.3, демонстрирует результаты, полученные при использовании порога отсечения по умолчанию, равного 0,5. Обратите внимание на то, что сумма всех значений составляет 60, что соответствует количеству наблюдений в тестовом наборе. В этой выборке 23 кандидата получили приглашение на стажировку, а 37 — нет. Насколько хорошо алгоритм справился с классификацией этих данных?

Матрица ошибок предоставляет несколько критериев для оценки производительности модели. Обычная точность — это всего лишь один из них.

Точность = Процент верных прогнозов = (36 + 19)/60 = 91,6%

Однако точность — это не то, на чем вам стоит сосредоточивать внимание, особенно учитывая ее уязвимость к проблеме несбалансированных данных. В большинстве случаев вас, скорее всего, будет волновать то, насколько хорошо ваш алгоритм предсказывает истинно положительные и истинно отрицательные значения. Другими словами, находит ли классификатор те случаи, которые должен находить (истинные положительные), и игнорирует ли те наблюдения, которые должен игнорировать (истинно отрицательные)?

Табл. 10.3. Матрица ошибок для прогнозов модели классификации с порогом отсечения 0,5

		Фактические значения	
		Да	Нет
Предсказанные значения	Да	19	1
	Нет	4	36

Доля истинно положительных результатов (она же «Чувствительность» или «Отзывчивость») = Количество соискателей, приглашенных на стажировку, деленное на количество соискателей, которые должны были получить такое приглашение = 19/(19 + 4) = 83%. Вам нужно, чтобы это значение было максимально близко к 100%.

Доля истинно отрицательных результатов («Специфичность») = Количество соискателей, которым было отказано в приглашении на собеседование, деленное на количество соискателей, которым должно было быть в нем отказано = 36/(36 + 1) = 97%. Это значение также должно быть максимально близко к 100%.

Напомним, что для создания матрицы ошибок по умолчанию, как правило, используется порог отсечения 0,5. Если бы мы увеличили это значение до 0,75, то для получения приглашения соискатель должен был бы соответствовать более строгим критериям. Новая матрица показана в табл. 10.4.

Обратите внимание, как изменились показатели.

Доля истинно положительных результатов = Количество соискателей, приглашенных на стажировку, деленное на количество соискателей, которые должны были получить такое приглашение = 12/(12 + 11) = 52%.

Доля истинно отрицательных результатов = Количество соискателей, которым было отказано в приглашении на собеседование, деленное на количество соискателей, которым должно было быть в нем отказано = 37/37 = 100%.

Увеличение порогового значения привело к уменьшению доли истинно положительных результатов, что, в свою очередь, увеличило долю истинно отрицательных результатов. Более высокий порог позволяет отсеять

неподходящих кандидатов, но за это приходится заплатить отсевом нескольких подходящих кандидатов.

Мы хотели продемонстрировать компромисс, на который приходится идти при определении порога отсечения. В конечном счете выбор подходящего порогового значения требует экспертных знаний в предметной области. Как главный по данным вы должны потратить время на обдумывание порога отсечения, лучше всего подходящего для решения стоящей перед вами задачи.

Табл. 10.4. Матрица ошибок для прогнозов модели классификации с порогом отсечения 0,75

		Фактические значения	
		Да	Нет
Предсказанные значения	Да	12	0
	Нет	11	37

Путаница в терминах, связанных с матрицей ошибок

Доля истинно положительных и истинно отрицательных результатов — это далеко не все показатели, которые можно получить на основе матрицы ошибок.

Статистики и врачи называют долю истинно положительных результатов «чувствительностью», а специалисты по работе с данными и машинному обучению — «отзывчивостью». В разных областях для одних и тех же показателей используются разные термины.

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы обсудили логистическую регрессию, деревья решений и ансамблевые методы. Кроме того, мы поговорили о множестве подводных камней, с которыми вы можете столкнуться при работе с моделями классификации. В частности, мы обсудили такие распространенные ловушки классификации, как:

— неправильное определение типа задачи;
— утечка данных;
— отсутствие разделения данных;
— выбор неправильного порогового значения для принятия решения;
— неправильное понимание точности.

Для лучшего понимания точности мы описали матрицу ошибок и то, как ее можно использовать для оценки производительности модели. В следующей главе мы поговорим о неструктурированных данных и текстовой аналитике.

ГЛАВА 11

Освойте текстовую аналитику

«Стремитесь к успеху, но готовьтесь к овощам»

_ *InspireBot, бот на основе искусственного интеллекта, «предназначенный для создания неограниченного количества уникальных вдохновляющих цитат»*[104]

В нескольких предыдущих главах мы говорили о данных в их традиционном понимании. Для большинства людей наборы данных представляют собой таблицы, состоящие из строк и столбцов. Это структурированные данные. Однако в реальном мире большая часть данных, с которыми вы взаимодействуете каждый день, является неструктурированной. Эти данные содержатся в текстах, которые вы читаете, в словах и предложениях электронных писем, новостных статей, сообщений в социальных сетях, обзоров продуктов на Amazon, статей в «Википедии» и книги, которую вы держите в руках.

Эти неструктурированные текстовые данные также можно проанализировать, но с ними нужно обращаться несколько иначе — о чем мы и поговорим в этой главе.

ОЖИДАНИЯ ОТ ТЕКСТОВОЙ АНАЛИТИКИ

Прежде чем углубиться в тему, мы хотели бы поговорить об ожиданиях от текстовой аналитики. На протяжении многих лет этому виду аналитики уделялось достаточно большое внимание. Одним из способов ее применения является анализ настроений, позволяющий определять эмоции автора

[104] Сгенерируйте собственные вдохновляющие цитаты на сайте **inspirobot.me**.

публикации в социальных сетях, комментария или жалобы. Однако, как вы увидите далее, проанализировать текст не так-то просто. К концу этой главы вы поймете, почему некоторые компании преуспевают в использовании текстовой аналитики, а другие — нет.

Многие люди уже представляют, на что способны компьютеры, анализирующие человеческий язык, благодаря огромному успеху компьютера IBM Watson в викторине *Jeopardy!* в 2011 году[105] и более поздним достижениям в области разработки систем распознавания речи (например, Alexa от Amazon, Siri от Apple и Assistant от Google). Такие системы перевода, как Google Translate, достигли уровня производительности, близкого к человеческому, за счет использования машинного обучения (в частности, контролируемого). Эти приложения по праву считаются одними из самых выдающихся достижений в области компьютерных наук, лингвистики и машинного обучения.

Именно поэтому предприятия имеют чрезвычайно большие ожидания, когда начинают анализировать имеющиеся у них текстовые данные: комментарии клиентов, результаты опросов, медицинские записи — любой текст, хранящийся в базах данных. Если уж путешественники могут перевести свою речь на один из сотни языков за долю секунды, то и компания, безусловно, сможет проанализировать тысячи комментариев клиентов, чтобы выявить самые насущные проблемы. Верно?

Ну, может, и так.

Технологии анализа текста, хоть и позволяют решать масштабные и сложные задачи, вроде преобразования голоса в текст и речевого перевода, но часто не справляются с задачами, которые кажутся гораздо более простыми. И мы по опыту знаем, что, когда компании приступают к анализу собственных текстовых данных, их часто постигает разочарование. Короче говоря, анализировать текст сложнее, чем может показаться. И как главный по данным, вы должны учитывать это при формулировании своих ожиданий.

Цель этой главы — преподать вам основы текстовой аналитики[106], которая позволяет извлекать полезную информацию из необработанного текста. Имейте в виду, что мы коснемся этой развивающейся области лишь вскользь. Однако мы надеемся, что это позволит вам получить некоторое

---

[105] Отличное описание системы вопросов-ответов, используемой компьютером Watson, можно найти в книге: Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons.

[106] Текстовая аналитика также иногда называется текст-майнингом.

представление о ее возможностях и проблемах. Благодаря этому по мере появления новых разработок в этой области вы сумеете понять, что из них может оказаться полезным, а что — нет. Как и в случае с любым другим направлением, чем больше вы его изучаете, тем лучше представляете его возможности, а также вырабатываете некоторый скептицизм, вполне приличествующий главному по данным.

В следующих разделах мы поговорим о том, как обнаружить структуру в неструктурированных текстовых данных, какому анализу вы можете их подвергнуть, а затем вернемся к вопросу о том, почему крупнейшие технологические компании могут добиться научно-фантастического прогресса в анализе своих текстовых данных, в то время как остальные могут испытывать с этим трудности.

КАК ТЕКСТ ПРЕВРАЩАЕТСЯ В ЧИСЛА

Читая текст, люди понимают настроение, сарказм, намеки, нюансы и смысл. Иногда это даже невозможно объяснить: стихотворение вызывает в памяти воспоминание, шутка заставляет смеяться.

Так что совсем не удивительно, что компьютер не понимает смысла так же, как это делает человек. Компьютеры могут лишь «видеть» и «считывать» числа. Чтобы проанализировать массу неструктурированных текстовых данных, их необходимо сначала преобразовать в числа и уже знакомые вам структурированные наборы данных. Это преобразование неструктурированного и запутанного текста, содержащего орфографические ошибки, сленг, смайлики или аббревиатуры, в аккуратный структурированный набор данных из строк и столбцов может быть весьма субъективным и трудоемким процессом. Сделать это можно несколькими способами; три из них мы рассмотрим далее.

Большой мешок слов

Самый простой способ преобразования текста в числа предполагает создание модели «мешка слов», которая игнорирует порядок слов и грамматику. В результате фраза «Это предложение является очень большим мешком слов» преобразуется в набор, называемый документом, в котором каждое слово является идентификатором, а количество слов — признаком. Порядок слов не имеет значения, поэтому мы сортируем содержимое мешка по алфавиту: {большим: 1, мешком: 1, очень: 1, предложение: 1, слов: 1, это: 1, является: 1}.

Табл. 11.1. Преобразование текста в числа методом «мешка слов». Числа обозначают количество того или иного слова (токена) в соответствующем предложении (документе)

Исходный текст	большим	двух	из	мешком	очень	предложение	продуктами	с	слов	состоит	это	является
Это предложение является очень большим мешком слов.	1	0	0	1	1	1	0	0	1	0	1	1
Это является большим мешком с продуктами.	1	0	0	1	0	0	1	1	0	0	1	1
Это предложение состоит из двух слов.	0	1	1	0	0	1	0	0	1	1	1	0

Кратко об облаках слов

Прежде чем двигаться дальше, давайте поговорим об облаках слов; это первое, с чем сталкиваются люди, интересующиеся текстовой аналитикой. Облако слов — это простое визуальное представление, в котором размер слова зависит от частоты, с которой оно встречается в словаре. Облако слов из текста данной главы показано на рис. 11.1[107].

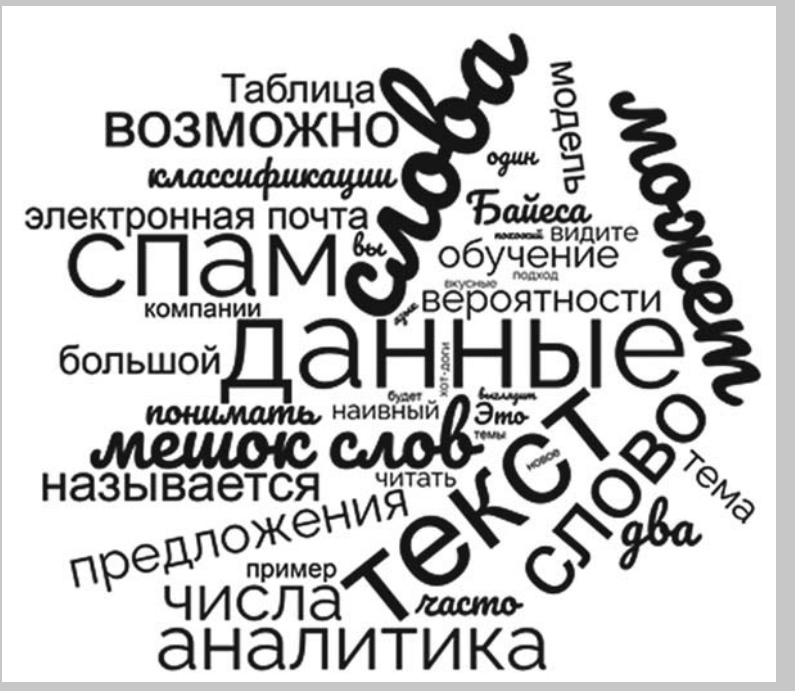


Рис. 11.1. Облако слов из текста данной главы

Вам удалось извлечь какую-нибудь полезную информацию из рис. 11.1? Вероятно, нет. Мы понимаем, что облака слов — отличный маркетинговый материал, но мы не их фанаты и не рекомендуем их использовать даже в качестве инструмента визуализации, поскольку анализировать частоту встречаемости слова в тексте по размеру шрифта гораздо сложнее, чем по длине столбика гистограммы.

[107] Облако слов создано с помощью сайта **wordclouds.com**.

Каждый идентификатор называется токеном. Набор токенов из всех документов — словарем.

Разумеется, ваши текстовые данные будут содержать не один документ, поэтому мешок слов может стать очень большим. Каждое уникальное слово и вариант написания станет новым токеном. Вот как будет выглядеть таблица, в каждой строке которой содержится предложение (комментарий, отзыв о продукте и так далее).

Для необработанного текста:

— Это предложение является очень большим мешком слов.
— Это является большим мешком с продуктами.
— Это предложение состоит из двух слов.

Мешок слов будет выглядеть так, как показано в табл. 11.1, где точки данных — количество того или иного слова в предложении.

Глядя на табл. 11.1, называемую матрицей «документ — термин» (один документ в строке, один термин в столбце), становится понятно, что базовая текстовая аналитика может сводиться к подсчету количества повторений каждого из слов (самое популярное слово — «это») и определению предложения, содержащего максимальное количество токенов (первое предложение). Хотя приведенный пример не особенно интересен, именно так рассчитывается базовая сводная статистика для документов.

Скорее всего, вы также заметили некоторые недостатки в табл. 11.1 (и в облаке слов!). По мере добавления новых документов количество столбцов в таблице будет увеличиваться, поскольку вам придется добавлять новый столбец для каждого нового токена. Кроме того, в результате этого таблица станет разреженной, то есть заполненной нулями, потому что каждое отдельное предложение будет содержать лишь несколько слов из словаря.

Стандартный способ решения этой проблемы — удалить такие слова-заполнители, как «и», «но», «что», «или», «это» и так далее, которые сами по себе не добавляют смысла. Это так называемые стоп-слова. Также принято удалять знаки препинания и цифры, преобразовывать все символы в нижний регистр и применять стемминг, то есть отбрасывать суффиксы и окончания слов, что позволяет сопоставлять такие слова, как продукты и продуктов, с одной и той же основой «продукт-», а слова читать, читаю, читает — с основой «чита-». Более продвинутый аналог стемминга — так называемая лемматизация, которая позволяет сопоставлять слова «хороший», «лучше»,

«лучший» со словом «хороший». В этом смысле лемматизация «умнее» стемминга, но занимает гораздо больше времени.

Подобные корректировки позволяют значительно уменьшить размер словаря и упростить процесс анализа. На рис. 11.2 показано, как этот процесс выглядит для одного предложения.

Глядя на рис. 11.2, становится понятно, в чем сложность анализа текста. Процесс преобразования текста в числа отфильтровал эмоциональную составляющую, контекст и порядок слов. Если вам кажется, что это повлияет на результаты любого последующего анализа, то вы правы. И в данном случае нам еще повезло, что в тексте отсутствуют орфографические ошибки, представляющие дополнительную проблему для специалистов по работе с данными.

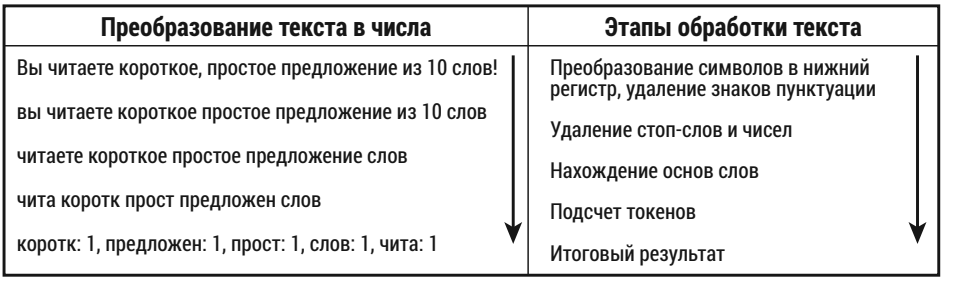


Рис. 11.2. Преобразование текста в мешок слов

Использование метода «мешка слов», доступного в свободном программном обеспечении и изучаемого на курсах по текстовой аналитике, привело бы к получению одинакового результата при числовом кодировании следующих двух предложений, несмотря на очевидные различия в их смысле:

1. Джордан любит хот-доги, но ненавидит гамбургеры[108].
2. Джордан ненавидит хот-доги, но любит гамбургеры.

Люди понимают разницу между этими двумя предложениями, а модель «мешка слов» — нет. Однако не спешите списывать ее со счетов. Несмотря на свой примитивный принцип, мешок слов может оказаться весьма полезным при обобщении разрозненных тем, которые мы рассмотрим в следующих разделах.

[108] Любимое блюдо Джордана — это хот-дог.

N-граммы

На примере с Джорданом и его любимыми хот-догами легко увидеть недостаток метода мешка слов. Фраза из двух слов «любит хот-доги» по смыслу противоположна фразе «ненавидит хот-доги», однако при использовании метода мешка слов контекст и порядок слов игнорируются. В данном случае могут помочь N-граммы. *N-грамма* — это последовательность из N-слов, поэтому 2-граммы (формально называемые биграммами) для фразы «Джордан любит хот-доги, но ненавидит гамбургеры» будут такими: {Джордан любит: 1, любит хот-доги: 1, ненавидит гамбургеры: 1, но ненавидит: 1, хот-доги, но: 1}.

Данное расширение метода мешка слов добавляет контекст, необходимый для различения фраз из одинаковых слов, но расположенных в разном порядке. Токены биграмм часто добавляются в мешок слов, что еще сильнее увеличивает размер и разреженность матрицы «документ-термин». С практической точки зрения это означает необходимость хранить большую (широкую) таблицу, содержащую относительно небольшой объем информации. Мы добавили несколько биграмм в табл. 11.1 и получили табл. 11.2.

О том, следует ли отфильтровывать биграммы со стоп-словами, ведутся споры, поскольку такой подход может привести к потере контекста. В некоторых программах слова «мой» и «ваш» считаются стоп-словами, однако смысл биграммных фраз «мое предпочтение» и «ваше предпочтение» при отбрасывании стоп-слов полностью исчезает. Это еще одно решение, которое должны принять ваши специалисты при анализе текстовых данных. (Вероятно, вы уже начинаете понимать, насколько сложна текстовая аналитика.)

Табл. 11.2. Пополнение «мешка слов» биграммами. Получившаяся в результате матрица «документ-термин» является очень большой

большим	двух	из	мешком	очень	предложение	продуктами	с	слов	состоит	это	является	большим мешком	мешком с	с продуктами	это предложение	…
1	0	0	1	1	1	0	0	1	0	1	1	1	0	0	1	…
1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	0	…
0	1	1	0	0	1	0	0	1	1	1	0	0	0	0	1	…

Однако после такой подготовки обобщение текста можно произвести путем простого подсчета. Такие веб-сайты, как **Tripadvisor.com**, применяют эти подходы и предоставляют пользователям возможность быстро находить отзывы по часто упоминаемым словам или фразам. Например, на сайте вашего местного стейк-хауса среди предлагаемых поисковых запросов вы можете встретить такие биграммы, как «запеченный картофель» или «идеально приготовленный».

Векторное представление слов

С помощью метода мешка слов и N-грамм можно обнаружить сходства между документами. Если несколько документов содержат похожие наборы слов или N-грамм, вы можете предположить, что предложения связаны между собой (в разумных пределах, конечно: помните о любви/ненависти Джордана к хот-догам). В этом случае строки в матрице «документ-термин» будут численно похожи.

Но как можно численно определить то, какие слова в словаре — не документы, а именно слова — связаны между собой?

В 2013 году компания Google проанализировала миллиарды пар слов (два слова, находящихся в непосредственной близости друг от друга в предложении) в своей огромной базе данных статей Google News[109]. Проанализировав частоту встречаемости пар слов, — например, (вкусная, говядина) и (вкусная, свинина) встречались чаще, чем (вкусная, корова) и (вкусная, свинья), — специалисты компании смогли сгенерировать так называемое векторное представление слов, то есть представление слов в виде списка чисел (или векторов). Если слова «говядина» и «свинина» часто сопровождаются словом «вкусная», то математически они будут представлены как похожие в том элементе вектора, который связан с вещами, обычно описываемыми как «вкусные», то есть с тем, что мы, люди, называем едой.

Чтобы объяснить, как это работает, мы будем использовать небольшое количество пар слов (компания Google использовала миллиарды). Представьте, что при просмотре статьи в местной газете мы обнаруживаем следующие пары слов: (вкусная, говядина), (вкусный, салат), (корм, корова), (говядина, корова), (свинья, свинина), (свинина, салат), (салат, говядина), (есть,

[109] Более подробное описание модели Word2vec можно найти в главе 11 замечательной книги: Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin UK.

свинина), (корова, ферма) и так далее. (Придумайте еще несколько пар слов по аналогии.) Каждое слово попадает в словарь: {говядина, корова, вкусный, ферма, корм, свинья, свинина, салат}[110].

Слово «корова», например, может быть представлено в виде вектора, длина которого соответствует длине приведенного выше словаря, со значением 1 на месте слова «корова» и значением 0 на месте всех остальных слов: (0, 1, 0, 0, 0, 0, 0, 0). Этот вектор представляет собой входные данные для алгоритма контролируемого обучения и сопоставляется с соответствующим выходным вектором (длина которого также соответствует длине словаря). В нем содержатся вероятности появления других слов из словаря рядом с входным словом. Таким образом, входному слову «корова» может соответствовать выходной вектор (0,3, 0, 0, 0,5, 0,1, 0,1, 0, 0), говорящий о том, что слово «корова» сопровождалось словом «говядина» в 30% случаев, словом «ферма» — в 50% случаев, а словами «корм» и «свинья» — в 10% случаев.

Как и при решении любой другой задачи контролируемого обучения, модель пытается как можно лучше сопоставить входные данные (слова-векторы) с выходными (векторами вероятностей). Только в данном случае нас интересует не сама модель, а сгенерированная ею таблица чисел, демонстрирующая то, как каждое из слов в словаре связано со всеми остальными словами. Это и есть векторное представление слов, которое можно рассматривать как числовое представление слова, кодирующее его «смысл». В табл. 11.3 показаны некоторые слова из нашего словаря наряду с их векторным представлением. Слово «корова», например, в данном случае записано в виде трехмерного вектора (1,0, 0,1, 1,0). Раньше оно было записано в виде более длинного и разреженного вектора (0, 1, 0, 0, 0, 0, 0, 0).

Табл. 11.3. Представление слов в виде векторов

Слово	Измерение 1	Измерение 2	Измерение 3
Корова	**1,0**	0,1	**1,0**
Говядина	0,1	**1,0**	**0,9**
Свинья	**1,0**	0,1	0,0
Свинина	0,1	**1,0**	0,0
Салат	0,0	**1,0**	0,0

[110] Да, здесь мы игнорируем множество пар слов, которые могут присутствовать даже в самых коротких статьях. Уже одно это должно дать вам представление о той вычислительной сложности, с которой пришлось столкнуться компании Google.

Интересная особенность векторного представления слов — измерения содержат их смысл подобно тому, как снижение размерности при анализе главных компонент позволяет отразить темы признаков.

Посмотрите на содержимое столбца «Измерение 1» в табл. 11.3. Видите закономерность? Что бы ни означало данное измерение, Корова и Свинья обладают соответствующим свойством в полной мере, а Салат вообще не имеет к нему отношения. Мы могли бы назвать это измерение Животные. Измерение 2 можно было бы назвать Еда, учитывая высокие значения для слов Говядина, Свинина и Салат, а Измерение 3 мы могли бы назвать Коровьим, потому что в нем выделяются слова, имеющие отношение к коровам. Благодаря этой структуре можно увидеть сходства в том, как используются слова, и даже составить из них простые уравнения (каким бы странным это ни казалось).

В качестве упражнения, используя табл. 11.3, попробуйте убедиться в том, что Говядина — Корова + Свинья ≈ Свинина[111].

Этот способ называется Word2vec[112] (преобразование слов в векторы), а векторы, сгенерированные Google, доступны для бесплатной загрузки[113]. Разумеется, не стоит ждать, что все соотношения будут идеальными. Как вы уже убедились, вариации есть во всем, и в тексте тоже. Вкусными могут называться не только продукты питания, но и совершенно несъедобные вещи. Кроме того, ситуацию усугубляет множество существующих в языке омонимов.

Векторное представление слов с его числовой структурой, позволяющей выполнять вычисления, применяются в поисковых и рекомендательных системах. Однако векторы, созданные на основе текста новостей сервиса Google News, могут не иметь отношения к стоящим перед вами проблемам. Например, торговые названия стирального порошка Tide® (в переводе с англ. «прилив») и крекеров Goldfish® («золотая рыбка») в системе типа Word2vec могут быть семантически близки словам «океан» и «домашнее животное» соответственно. Однако для продуктового магазина эти товары могут оказаться семантически более близкими таким конкурирующим брендам, как стиральный порошок Gain® и крекеры Barnum's Animals®.

---

[111] Говядина = (0,1, 1,0, 0,9), Корова = (1,0, 0,1, 1,0), Свинья = (1,0, 0,1, 0,0). Если произвести сложение и вычитание соответствующих элементов, то получится Говядина — Корова + Свинья = (0,1, 1,0–0,1), что довольно близко к значению слова Свинина = (0,1, 1,0, 0).

[112] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[113] **code.google.com/archive/p/word2vec**.

Вы можете применить модель Word2vec к своим текстовым данным и сгенерировать собственные векторы слов. При этом вы можете обнаружить темы и концепции, которые в противном случае даже не пришли бы вам в голову. Однако выявление достаточного количества смысловых взаимосвязей может стать проблемой. Не каждая компания имеет доступ к такому же количеству данных, как Google. Вам может просто не хватить текстовых данных для получения осмысленных векторов.

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

После превращения текста в осмысленный набор данных можно приступать к его анализу. Инвестиция в преобразование неструктурированных текстовых данных в структурированный набор из строк и столбцов окупается возможностью применения методов анализа, описанных в предыдущих главах (с некоторыми изменениями). Этим методам будут посвящены следующие несколько разделов.

В главе 8 мы говорили о неконтролируемом обучении, суть которого заключается в поиске естественных закономерностей в строках и столбцах набора данных. Применив алгоритм кластеризации вроде метода *k*-средних к матрице «документ-термин» наподобие табл. 11.1 и 11.2, мы получим набор из *k* различных, но в чем-то похожих групп текстовых данных. В некоторых случаях это может оказаться полезным. Однако выполнять кластеризацию текстовых данных методом *k*-средних довольно сложно. Рассмотрим, к примеру, следующие три предложения:

1. Министерству обороны следует определиться с официальной политикой в отношении космоса.
2. Договор о нераспространении ядерного оружия важен для национальной обороны.
3. Соединенные Штаты недавно отправили в космос двух астронавтов в рамках своей космической программы.

На наш взгляд, здесь обсуждаются две общие темы — национальная оборона и космос. В первом предложении затрагиваются обе темы, а во втором и третьем — только одна. (Если вы с этим не согласны — а на это у вас, безусловно, есть право, — то вы уже представляете главную проблему кластеризации текстовых данных: четкое разделение тем возможно далеко не всегда.

Кроме того, относительно текста люди могут составить разные мнения, чего не скажешь о числах.)

Тематическое моделирование[114] похоже на метод *k*-средних в том смысле, что это алгоритм неконтролируемого обучения, который пытается сгруппировать похожие наблюдения, но при этом не требует, чтобы каждый документ был явно отнесен к одному кластеру. Вместо этого он выдает значения вероятности, говорящие о том, как один документ связан с несколькими темами. Предложение 1, например, может получить оценку 60% по теме национальной обороны и 40% по теме космоса.

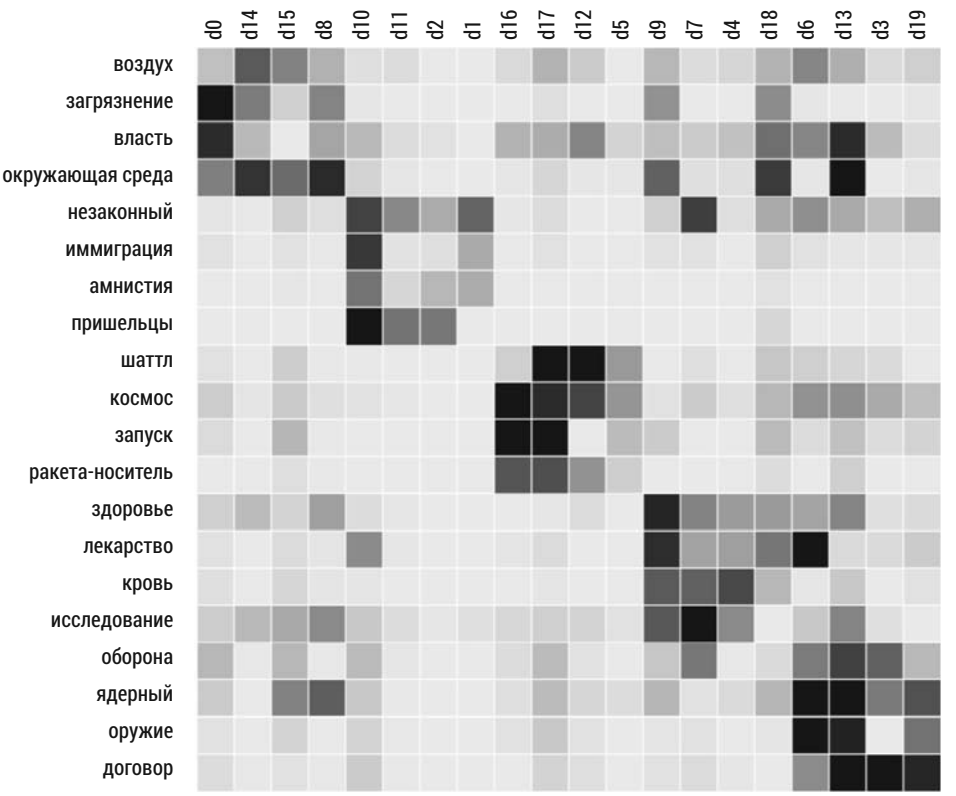


Рис. 11.3. Кластеризация документов и терминов путем тематического моделирования. Можете ли вы выделить пять основных тем на этом изображении? Как бы вы их назвали?

[114] Два популярных метода тематического моделирования — латентно-семантический анализ (ЛСА) и латентное размещение Дирихле (ЛРД).

Чтобы лучше в этом разобраться, рассмотрим еще один пример. На рис. 11.3 представлена перевернутая набок матрица «документ-термин»[115]. Слева вы видите термины, которые встречаются в 20 документах d0–d19. Каждая ячейка отражает частоту встречаемости слова в документе: более темные ячейки соответствуют более высокой частоте. Причем термины и документы были упорядочены с помощью метода тематического моделирования.

Глядя на это изображение, вы можете заметить слова, часто встречающиеся в документах и вместе образующих возможные темы, а также документы, содержащие термины, связанные с несколькими темами (в частности, обратите внимание на d13, третий столбец справа). Однако имейте в виду: данный метод не гарантирует получения точных результатов, как и другие методы неконтролируемого обучения.

С практической точки зрения тематическое моделирование работает лучше всего, когда в наборе документов представлены различные темы. Это может показаться очевидным, однако нам известны случаи, когда тематическое моделирование применялось к подмножествам текстов, которые перед анализом были отфильтрованы по конкретной интересующей аналитиков теме. Это все равно что взять группу новостных статей, отобрать только те, которые содержат слова «баскетбол» и «Леброн Джеймс», а затем ожидать значимых результатов от применения тематического моделирования к отобранным статьям. Результаты вас разочаруют. Отфильтровывая тексты, вы, по сути, задаете одну тему для оставшихся статей. Помните об этом нюансе, продолжайте спорить со своими данными и корректируйте ожидания по мере необходимости.

КЛАССИФИКАЦИЯ ТЕКСТОВ

В этом разделе мы поговорим о контролируемом обучении на матрице «документ-термин» (при условии наличия известных целевых атрибутов). В случае с текстом мы, как правило, пытаемся предсказать категориальную переменную, поэтому данная задача решается с помощью моделей классификации, о которых мы говорили в предыдущей главе, а не моделей

[115] Это изображение взято с сайта **en.wikipedia.org/wiki/File:Topic_model_scheme.webm**, создано Кристофом Карлом Кингом и распространяется по лицензии Creative Commons Attribution-Share Alike 4.0 International.

регрессии, которые предсказывают числа. Один из самых известных примеров успешного применения метода классификации текстов — спам-фильтр, используемый в сервисах электронной почты, входными данными для которого является текст сообщения, а выходными — бинарный флаг «спам» или «не спам»[116]. Пример применения многоклассовой классификации текстов — автоматическое распределение новостных онлайн-статей по категориям: местные новости, политика, мир, спорт, развлечения и так далее.

Чтобы получить представление о том, как происходит классификация текстов с помощью метода мешка слов, давайте рассмотрим один (упрощенный) случай. В табл. 11.4 показаны пять различных тем электронных писем, разбитых на токены, и метки, указывающие, является ли письмо спамом или нет. (Следует помнить о том, сколько усилий компании тратят на сбор подобных данных. Всякий раз, когда провайдер электронной почты спрашивает вас, являются ли те или иные письма спамом, вы предоставляете данные для алгоритмов машинного обучения!)

Табл. 11.4. Простейший пример классификации спама

Тема письма	совет	лысый	день рождения	долги	бесплатно	помощь	мама	вечеринка	избавление	акции	виагра	спам?
Совет для вечеринки по случаю дня рождения мамы	1	0	1	0	0	0	1	1	0	0	0	0
Бесплатная «Виагра!»	0	0	0	0	1	0	0	0	0	0	1	1
Бесплатный совет по поводу акций	1	0	0	0	1	0	0	0	0	1	0	1
Бесплатный совет по избавлению от долгов	1	0	0	1	1	0	0	0	1	0	0	1
Лысеете? Мы можем помочь!	0	1	0	0	0	1	0	0	0	0	0	1

[116] Одна из самых значимых статей в данной области — Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural networks, 10*(5), 1048–1054.

Как можно использовать алгоритм для извлечения уроков из данных в табл. 11.4 и формирования прогнозов относительно новых, невиданных ранее тем электронных писем?

Возможно, вы подумали о логистической регрессии, которая бывает полезна для предсказания бинарных результатов. Но, к сожалению, здесь она не сработает из-за слишком большого количества слов и недостаточного количества примеров для обучения. В табл. 11.4 столбцов больше, чем строк, а логистическая регрессия этого не любит[117].

Наивный байесовский алгоритм

В подобной ситуации обычно применяется так называемый наивный байесовский алгоритм классификации (названный так в честь того самого Байеса, которого мы упоминали в главе 6). Данный алгоритм пытается ответить на простой вопрос: где с большей вероятностью встречаются слова, используемые в теме электронного письма, — в спаме или в обычном письме? Вы, скорее всего, пытаетесь сделать то же самое, когда просматриваете содержимое своего почтового ящика: вы по собственному опыту знаете, что слово «бесплатно», как правило, встречается в спаме. То же касается слов «деньги», «Виагра» или «богатство». Если большинство слов ассоциируется со спамом, то электронное письмо, скорее всего, им и является. Все просто.

Другими словами, вы пытаетесь рассчитать вероятность того, что электронное письмо — спам, опираясь на слова, содержащиеся в строке темы, (w_1, w_2, w_3,…). Если эта вероятность превышает вероятность того, что письмо спамом не является, мы отмечаем его как спам. Эти конкурирующие вероятности записываются с помощью следующей нотации:

— Вероятность того, что электронное письмо является спамом = P(спам | w_1, w_2, w_3,…).
— Вероятность того, что электронное письмо не является спамом = P(не спам | w_1, w_2, w_3,…).

Прежде чем двигаться дальше, давайте изучим данные в табл. 11.4. Нам известна вероятность, с которой каждое из слов встречается в спам

[117] Линейная регрессия не работает и в том случае, если в наборе данных содержится больше признаков, чем наблюдений. Тем не менее существуют разновидности линейной и логистической регрессии, позволяющие справиться с такой ситуацией.

(и не спам) письмах. Слово «бесплатно» присутствовало в трех из четырех спам-сообщений, поэтому вероятность встретить данное слово при условии, что письмо является спамом, составляет P(бесплатно | спам) = 0,75. Выполнив аналогичные расчеты для слов «долг» и «мама», мы получим: P(долг | спам) = 0,25, P(мама | не спам) = 1 и так далее.

Что нам это дает? Мы хотим знать вероятность того, что то или иное электронное письмо — спам при условии наличия в нем определенных слов. При этом нам известна вероятность встретить то или иное слово в письме при условии того, что оно является спамом. Эти две вероятности не одинаковы, но они связаны теоремой Байеса (см. главу 6). Как вы помните, основная идея данной теоремы — поменять условные вероятности местами. Таким образом, вместо P(спам | $w_1, w_2, w_3,\ldots$) мы можем использовать $P(w_1, w_2, w_3,\ldots$ | спам). Благодаря дополнительным расчетам (которые мы опускаем для краткости[118]) принятие решения относительно классификации нового электронного письма как спам-сообщения сводится к выяснению того, какое из двух значений выше:

1. Оценка «спам» = P(спам) × P(w_1 | спам) × P(w_2 | спам) × P(w_3 | спам).
2. Оценка «не спам» = P(не спам) × P(w_1 | не спам) × P(w_2 | не спам) × P(w_3 | не спам).

Вся эта информация содержится в табл. 11.4. Вероятности P(спам) и P(не спам) отражают долю спама и не спама в обучающих данных — 80% и 20% соответственно. Другими словами, если бы вы хотели делать предположения, не глядя на строку темы, вы бы предполагали, что письмо — «спам», потому что такие письма составляют класс большинства в обучающих данных.

Чтобы прийти к приведенным выше формулам, наивный байесовский алгоритм совершил то, что обычно считается вопиющей ошибкой при работе с вероятностями, а именно — допустил отсутствие зависимости между событиями. Вероятность встретить в спам-сообщении оба слова «бесплатно» и «Виагра», обозначаемая как P(бесплатно, виагра | спам), зависит от того, насколько часто эти слова встречаются в одном и том же письме, однако это значительно усложняет вычисления. «Наивность» наивного байесовского алгоритма выражается в предположении независимости всех вероятностей: P(бесплатно, виагра | спам) = P(бесплатно | спам) × P(виагра | спам).

---

[118] Дополнительную информацию вы можете найти в статье **https://ru.wikipedia.org/wiki/Байесовская_фильтрация_спама**

Более глубокий взгляд

Увидев электронное письмо с темой «Избавьтесь от долгов с помощью наших советов по торговле акциями!», вы бы сосредоточились на словах «избавьтесь», «долги», «акции» и «совет» и вычислили бы следующие конкурирующие значения:

1. Оценка «спам» = P(спам) × P(избавьтесь | спам) × P(долги | спам) × P(акции | спам) × P(совет | спам).
2. Оценка «не спам» = P(не спам) × P(избавьтесь | не спам) × P(долги | не спам) × P(акции | не спам) × P(совет | не спам).

Однако есть небольшая проблема. Новые и редкие слова требуют некоторой корректировки расчетов, чтобы вероятности не умножились на ноль. В крошечном наборе данных, приведенных в табл. 11.4, слово «избавьтесь» вообще не встречается, тогда как слова «долги», «акции» и «совет» встречаются только в спам-сообщениях. Из-за подобных нюансов оценки «спам» и «не спам» окажутся равными нулю. Чтобы это исправить, давайте представим, что мы встречали каждое слово хотя бы один раз, прибавив 1 к частоте встречаемости. Кроме того, мы прибавим 2 к частоте встречаемости спама (и не спама), чтобы значения не были равны 1[119].

Теперь мы можем произвести вычисления:

1. Оценка «спам»: $(0{,}8)\times\frac{0+1}{4+2}\times\frac{1+1}{4+2}\times\frac{1+1}{4+2}\times\frac{2+1}{4+2}=0{,}0074$.
2. Оценка «не спам»: $(0{,}2)\times\frac{0+1}{1+2}\times\frac{0+1}{1+2}\times\frac{0+1}{1+2}\times\frac{1+1}{1+2}=0{,}0049$.

Первое значение больше второго, поэтому электронное письмо с темой: «Избавьтесь от долгов с помощью наших советов по торговле акциями!» мы классифицируем как спам.

[119] Это называется поправкой Лапласа, которая помогает предотвратить высокую вариацию в небольших количествах значений, о которой мы говорили в главе 3.

Анализ настроений

Анализ настроений — это популярный способ применения алгоритмов классификации текстов к данным социальных сетей. Если вы введете в поисковую строку Google запрос «анализ настроений по сообщениям в Twitter», то количество результатов вас наверняка удивит; складывается впечатление, что этим заняты все. Суть идеи в данном случае та же, что и в рассмотренном выше примере со спамовыми/не спамовыми письмами и сводится к ответу на вопрос о том, являются ли слова в сообщении в социальной сети (обзоре продукта или опросе) скорее «положительными» или скорее «отрицательными». То, что вы будете делать с полученной информацией, зависит от конкретного бизнес-кейса. Однако следует отметить, что при анализе настроений не стоит выполнять экстраполяцию за пределы контекста обучающих данных, рассчитывая на получение осмысленных результатов.

Что мы имеем в виду? Дело в том, что многие классификаторы для «анализа настроений» обучаются на данных, находящихся в свободном доступе в Интернете. Популярный набор данных для студентов — большая коллекция рецензий на фильмы из базы данных IMDb.com. Этот набор данных и любая модель, созданная на его основе, будут иметь отношение исключительно к обзорам фильмов. Разумеется, она будет ассоциировать такие слова, как «великолепный» и «замечательный», с положительными эмоциями, однако не стоит ожидать, что эта модель будет хорошо работать при ее применении к уникальному бизнес-кейсу, которому присуща особая терминология.

А как насчет методов работы с текстом на основе деревьев?

Методы на основе деревьев, такие как случайный лес и бустинг (усиление), могут применяться для решения задач классификации текстов и, как правило, работают лучше, чем наивный байесовский алгоритм с некоторыми наборами данных. Однако наивный байесовский алгоритм обычно становится хорошей отправной точкой и отличается прозрачной интерпретацией.

ПРАКТИЧЕСКИЕ СООБРАЖЕНИЯ ПРИ РАБОТЕ С ТЕКСТОМ

Теперь, когда вы познакомились с несколькими инструментами текстовой аналитики, давайте сделаем шаг назад и поговорим об анализе текста на более высоком уровне.

При работе с текстом вам доступна роскошь чтения данных. Если тематическое моделирование намекает на то, что те или иные предложения относятся к определенным темам, вы можете оценить эти результаты. Если кто-то строит модель классификации текста, попросите представить как хорошие, так и плохие результаты.

По опыту нам известно, что презентовать успешный проект текстовой аналитики заинтересованным сторонам довольно весело, поскольку в данном случае результаты представляют не ряды чисел, а то, что аудитория может прочитать, понять и обсудить. Однако докладчики склонны акцентировать внимание на захватывающих и легких победах, а не на явных промахах. При представлении результатов анализа текста главный по данным должен стремиться к максимальной прозрачности. Также при обработке результатов запросите примеры, когда алгоритмы не сработали. Поверьте, так бывает.

Это возвращает нас к замечанию, которое мы сделали в начале главы: когда компании приступают к анализу собственных текстовых данных, их часто постигает разочарование. Оно было сделано вовсе не для того, чтобы отвратить вас от текстовой аналитики. Открыто говоря о недостатках, мы надеемся предотвратить возможную негативную реакцию со стороны вас или вашей компании, которая может возникнуть, когда вы начнете анализировать текст, поймете, что это сложнее, чем вы думали, и откажетесь от этой идеи или удовлетворитесь слабой аналитикой.

К этому моменту вы уже должны были выработать достаточно скепсиса, чтобы понимать, где именно могут возникнуть проблемы. Однако некоторые крупные технологические компании, по-видимому, преодолели эти трудности и добились лидерства в области текстовой аналитики и обработки естественного языка (NLP, Natural Language Processing), которая имеет дело со всеми аспектами языка, включая звук (в отличие от просто письменного текста).

Преимущества технологических гигантов

В отличие от многих других компаний, такие технологические гиганты, как Apple, Amazon, Google и Microsoft, обладают обилием текстовых и голосовых данных (данных, снабженных метками, которые можно использовать для контролируемого обучения моделей), мощными компьютерами, группами преданных делу исследователей мирового уровня и деньгами.

Благодаря таким ресурсам они добились значительного прогресса в области анализа не только текста, но и звука. В последние годы произошли заметные улучшения в следующих сферах:

— Преобразование речи в текст. Голосовые помощники и функции преобразования голоса в текст на смартфонах стали работать более точно.
— Преобразование текста в речь. Голоса в программах для чтения с экрана компьютера теперь больше напоминают человеческие.
— Преобразование текста в текст. Перевод с одного языка на другой выполняется мгновенно и с достаточно высокой точностью.
— Чат-боты. Окна чата, которые теперь автоматически открываются на каждом веб-сайте с вопросом: «Чем я могу вам помочь?», стали (чуть) более полезными.
— Генерация понятного человеку текста. Языковая модель GPT-3[120] от компании OpenAI способна генерировать текст, напоминающий человеческий, отвечать на вопросы, а также генерировать компьютерный код по запросу. На момент написания этой книги данная модель самая продвинутая в своем роде. Согласно оценкам, стоимость ее обучения (здесь имеется в виду только использование компьютеров без учета оплаты труда исследователей) составила 4,6 миллиона долларов США[121].

Добавьте к этому наличие доступа к данным и группы экспертов-исследователей, и вы поймете, почему обработка естественного языка (пока) остается недоступной большинству компаний. Хотя алгоритмы имеют открытый исходный код, массовый сбор данных и доступ к суперкомпьютерам остается прерогативой технологических гигантов.

---

[120] Generative Pre-trained Transformer 3

[121] **https://www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/?sh=2f45a93b481a**

Кроме того, при формулировании своих ожиданий следует учитывать то, что приложения, создаваемые технологическими гигантами, универсальны для миллионов людей, то есть предназначены для решения задач, общих для представителей всех слоев общества. Например, голосовой помощник Alexa от компании Amazon предназначен для всех, включая детей. А текстовый перевод осуществляется с учетом жестких правил, встроенных в наборы обучающих данных. Слову «вечеринка» в английском языке соответствует слово «фиеста» в испанском. Суть в том, что все пользователи этих систем ожидают того, что они будут работать одинаково.

Сравните это с задачей классификации текста, специфической для того или иного бизнеса. Например, тональность фразы «телефон Samsung лучше, чем iPhone» зависит от того, в какой компании вы работаете, — Apple или Samsung. Данные, к которым у вас есть доступ, могут отличаться особенным, уникальным только для вашей компании языком. Кроме того, размер данных будет значительно меньше, чем у технологических компаний. Соответственно, результаты могут оказаться не такими четкими, как вы ожидаете.

Тем не менее мы настоятельно рекомендуем вам воспользоваться всеми доступными алгоритмами, включая методы текстовой аналитики. Понимание зависит не столько от мощности компьютеров, сколько от контекста и ожиданий. Если вы осознаете ограничения текстовой аналитики до того, как приступите к ней, вы будете готовы правильно ее использовать.

ПОДВЕДЕНИЕ ИТОГОВ

Мы надеемся, что с помощью данной главы нам удалось убедить вас в том, что компьютеры понимают язык не так, как люди, — для компьютера это просто цифры. На наш взгляд, понимание одного этого факта уже очень важно. В следующий раз вы с меньшей вероятностью клюнете на маркетинговую удочку, когда услышите о том, что искусственный интеллект способен решить любую бизнес-задачу, связанную с текстом: в процессе преобразования текста в числа теряется часть смысла, который мы вкладываем в слова и предложения. В этой главе мы обсудили три метода:

— мешок слов;
— N-граммы;
— векторное представление слов.

После преобразования текста в числа вы можете применять методы обучения без учителя, например, тематическое моделирование, или методы обучения с учителем, такие как классификация текстов. Наконец, мы поговорили о том, почему технологические гиганты одерживают верх, и порекомендовали вам формулировать свои ожидания исходя из имеющихся данных и ресурсов.

В следующей главе мы продолжим анализировать неструктурированные данные и поговорим о нейронных сетях и глубоком обучении.

ГЛАВА 12

Концептуализируйте глубокое обучение

«Появление искусственного интеллекта иногда называют новой промышленной революцией. И если глубокое обучение — это паровой двигатель этой революции, то данные — это уголь: топливо, питающее наши интеллектуальные машины, без которого ничего не было бы возможно»

_ *Франсуа Шолле, исследователь ИИ и автор книг*[122]

Поздравляем: вам удалось добраться до главы, которая во многих отношениях является кульминацией вашего пути становления главным по данным. В ней мы соберем вместе различные фрагменты мозаики и погрузимся в развивающуюся область машинного обучения, называемую глубоким обучением.

Сегодня использование глубокого обучения стимулирует развитие передовых технологий, а его человекоподобные проявления периодически вызывают восхищение общественности. Сфера глубокого обучения охватывает технологии, лежащие в основе работы систем распознавания лиц, автономного вождения, обнаружения рака и перевода речи. То есть они помогают принимать решения, которые некогда считались прерогативой человека. Однако, как будет показано далее, глубокое обучение не является чем-то новым и не настолько похоже на работу человеческого разума, как может показаться на первый взгляд.

[122] Шолле Франсуа, «Глубокое обучение на Python» (Издательство: Питер, 2018).

Большая часть ожиданий и ажиотажа в сфере работы с данными обусловлена потенциалом глубокого обучения. Неудивительно, что представители делового мира тратят много денег на внедрение этой технологии, что в ближайшие годы может повлиять на многие отрасли. Однако по мере развития сферы глубокого обучения нарастает и шумиха вокруг нее. При этом из виду часто упускаются порождаемые ею этические проблемы.

В этой главе мы рассмотрим компоненты глубокого обучения, начав с его структуры. В основе глубокого обучения лежит семейство моделей, называемых искусственными нейронными сетями. Считается, что эти алгоритмы имитируют то, как мозг обдумывает идеи, — однако, как мы увидим далее, это верно лишь отчасти. Затем мы поговорим о том, как нейронные сети можно модифицировать для решения более сложных задач (вроде распознавания образов). В конце главы мы коснемся практических проблем, связанных с применением технологии глубокого обучения, поговорим о ее неправильном использовании и более широких последствиях применения моделей типа «черный ящик».

НЕЙРОННЫЕ СЕТИ

Прежде чем концептуализировать глубокое обучение, сначала необходимо познакомиться с его строительными блоками — искусственными нейронными сетями.

Чем нейронные сети похожи на мозг?

Человеческий мозг — это сеть, состоящая из биологических нейронов. Считается, что эти нейроны «поглощают информацию» в виде химических сигналов и электрических импульсов. В определенный момент — мы не до конца понимаем, в какой именно — эта информация «активирует» нейрон, то есть заставляет его среагировать. Если вы ведете машину, и на дорогу внезапно выбегает олень, ваш мозг быстро обрабатывает входные данные (вашу скорость, расстояние до оленя, присутствие машин поблизости), активируя миллионы нейронов, которые, в свою очередь, принимают решение (нажать на тормоз или свернуть с дороги)[123].

[123] Разумеется, продемонстрировать резкие и ожидаемые изменения в химии мозга можно не только с помощью такого экстремального примера, как выбегающий на дорогу олень. Дело в том, что ваш мозг обрабатывает входные и выходные данные прямо сейчас. Миллионы нейронов активируются в процессе чтения этих строк.

В связи с этим возникает вопрос: можно ли создать серию моделей и алгоритмов, способных учиться так же, как это делает наш мозг? Можно ли быстро преобразовать входные параметры в виде данных, изображений или звука в осмысленные выходные данные? Только подумайте, какие возможности предоставил бы нам алгоритм, имитирующий работу нашего мозга. Сколько оценок, делаемых нами каждую секунду, мы могли бы оптимизировать, поручив компьютеру решение соответствующих задач?

Для ответа на этот вопрос и были созданы искусственные нейронные сети — вычислительный аналог биологических нейронных сетей.

Все это звучит невероятно, и, разумеется, авторы этой книги считают все связанное с нейронными сетями чрезвычайно захватывающим. Первые нейронные сети были созданы в 1940-х годах для имитации биологии человека в ее тогдашнем понимании. Большая часть ажиотажа вокруг нейронных сетей — а значит, и глубокого обучения — обусловлена тем, что они вдохновлены работой нашего мозга. Однако уподоблять нейронную сеть мозгу весьма рискованно, поскольку такая аналогия приписывает уровень абстракции и общих знаний моделям нейронных сетей, которые, по сути, являются всего лишь гигантскими математическими уравнениями.

Таким образом, несмотря на заявления СМИ и маркетологов, мы не должны обманываться, думая, будто последние достижения в области нейронных сетей и глубокого обучения отражают их более тесную связь с человеческим мозгом. Успех этих алгоритмов скорее обусловлен более быстрыми компьютерами, огромными объемами данных и множеством исследований, проводимых в области машинного обучения, статистики и математики.

Давайте рассмотрим принцип работы нейронных сетей на двух примерах.

Простая нейронная сеть

Как вы помните, в главе 10 мы создали модель, предсказывающую, получит ли кандидат приглашение на собеседование, исходя из его среднего балла, курса обучения, специализации и количества внеклассных занятий. На рис. 12.1 показана ее визуализация в виде простейшей нейронной сети.

На рис. 12.1 представлены четыре входных параметра:

— Средний балл = 3,90
— Курс = 4
— Специализация = «Статистика» (закодирована с помощью цифры 2)
— Внеклассные занятия = 5 (общее количество внеклассных занятий)

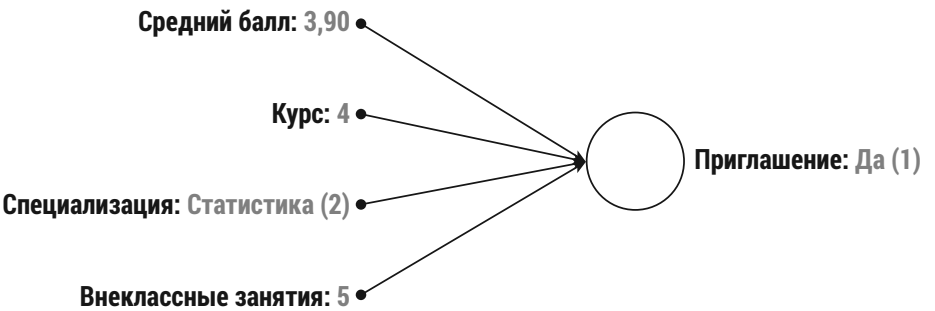


Рис. 12.1. Простейшая нейронная сеть из всех возможных. Четыре входных параметра обрабатываются функцией активации одного нейрона, определяющей выходной сигнал

Эти значения передаются в вычислительную единицу — нейрон, который представлен на рисунке в виде кружка. Внутри этого нейрона находится функция активации, которая преобразует четыре входных параметра в единое числовое значение. Если комбинация входных параметров превышает пороговое значение, нейрон «активируется» и предсказывает, что заявитель получит приглашение.

В качестве функции активации можно использовать разные функции в зависимости от решаемой задачи и имеющихся данных. Поскольку в этом примере мы решаем задачу классификации, то есть предсказываем, получит ли данный стажер приглашение, — наша функция активации должна предоставить нам значение вероятности получения приглашения, аналогично тому, как это делалось в главе 10 с помощью логистической регрессии[124].

В уравнениях 12.1 и 12.2 показана наиболее распространенная функция активации. Мы разбили ее на две части, чтобы ее было легче использовать (и набирать):

$$\text{Вероятность приглашения} = \frac{1}{1+e^{-(X)}}, \tag{12.1}$$

где $X = w_1 \times \text{Средний балл} + w_2 \times \text{Курс} + w_3 \times \text{Специализация} + w_4 \times \text{Внекл. занятия} + b.$ (12.2)

[124] Нейронные сети можно использовать и для решения задач регрессии. Только при этом будет применяться другая функция активации, поскольку итоговое вычисление, по сути, будет сводиться к модели линейной регрессии.

Мы надеемся, что эти уравнения уже кажутся вам знакомыми. Уравнение 12.1 — это логистическая функция из главы 10, а уравнение 12.2 — функция линейной регрессии, представленная в главе 9. Таким образом, с математической точки зрения нейронная сеть просто содержит компоненты машинного обучения и статистических алгоритмов. Уравнение линии линейной регрессии 12.2 позволяет объединить четыре входных параметра в один, а логистическая функция 12.1 «втискивает» результат в диапазон от 0 до 1, в котором должны находиться значения вероятности.

Цель сети, как и логистической регрессии, — найти наилучшие значения весов и постоянного члена (представленных в уравнении 12.2 буквами *w* и *b* соответственно и вместе называемых параметрами), которые делают выходные данные, прогнозируемые сетью, максимально близкими к фактическим выходным данным в совокупности[125]. Под «обучением» нейронной сети (как и в случае машинного обучения в целом) понимается процесс оптимизации параметров уравнений, подобных 12.2, для получения прогноза.

Как учится нейронная сеть

Главный вопрос — какими должны быть эти параметры для достижения оптимума? Ответ на него превращает нейронную сеть в полезную машину для прогнозирования. Однако в самом начале процесса обучения параметры могут быть любыми. Поэтому наш алгоритм присваивает им случайные значения в качестве отправной точки. Если бы вам нужно было вымыть руки под краном, который вы видите впервые и у которого отсутствуют метки горячей и холодной воды, то вы просто включили бы его, оценили температуру и отрегулировали ее. То же самое и здесь.

Эти начальные случайные веса неверны в том смысле, что они были установлены произвольно, а не определены в процессе обучения. Однако они позволяют с чего-то начать, а самое главное — получить числовой результат. Например, предположим, что вы берете данные двух выдающихся бывших стажеров, Уилла и Элли, и подаете их на вход нейронной сети (то есть подставляете в приведенные выше уравнения). Используя случайные параметры, мы получили результат 0,2 для Уилла и 0,3 для Элли. Другими словами, при использовании случайных значений весов и постоянных членов

[125] Веса также называются коэффициентами. Для одних и тех же понятий существует несколько названий.

вероятность получения приглашения обоими претендентами оказалась низкой. Однако, поскольку это исторические обучающие данные, мы знаем, какими должны были быть выходные значения. И Уилл, и Элли получили приглашения. Истинные выходные значения в обоих случаях были равны 1, но модель предсказала низкие значения для каждого из стажеров. Итак, исходное качество предсказаний нейронной сети ужасно.

На этом этапе модель анализирует истинные значения выходных параметров (1 и 1) и отправляет сообщение о том, что текущие параметры неверны, и их необходимо скорректировать. Но как именно мы должны изменить каждый вес? Алгоритм, называемый обратным распространением ошибки[126], корректирует эти веса и решает, на сколько их необходимо увеличить или уменьшить. Может быть, придать большее значение среднему баллу? А может быть, стоит уменьшить важность курса? Затем процесс повторяется: обновленные веса снова применяются для оценки данных Уилла и Элли, и на этот раз выдаются результаты 0,4 и 0,6. Уже лучше, но еще не идеально. Алгоритм обратного распространения ошибки отправляет сигнал обратно по сети и снова корректирует веса, после чего процесс повторяется. Со временем параметры приближаются к своему гипотетическому оптимуму, при котором предсказанные значения в среднем наиболее близки к фактическим меткам[127].

Чуть более сложная нейронная сеть

В предыдущем примере мы просто взяли логистическую регрессию и превратили ее в визуализацию нейронной сети. Математика была идентичной. В связи с этим возникает вопрос: зачем вообще это делать? Зачем представлять логистическую регрессию в виде чего-то нового, называемого нейронной сетью?

Ответ на этот вопрос и реальная польза нейронных сетей заключается в том, что происходит, когда мы добавляем в сеть «скрытые» слои. Итак,

---

[126] Для поклонников исчисления сообщаем, что обратное распространение ошибки, по сути, представляет собой цепное правило, предоставляющее инструменты для оптимизации вложенных уравнений, подобных тем, которые используются в нейронных сетях.

[127] В случае линейной регрессии для параметров существует настоящий математический оптимум (то есть точка, в которой сумма квадратов является минимальной). К сожалению, при работе с нейронными сетями у нас часто нет никакого способа узнать, достигла ли наша нейронная сеть математического оптимума или просто «достаточно хорошего» результата.

давайте добавим в предыдущую сеть скрытый слой с тремя нейронами, каждый из которых содержит логистическую функцию активации. В результате мы получим сетевую структуру, показанную на рис. 12.2.

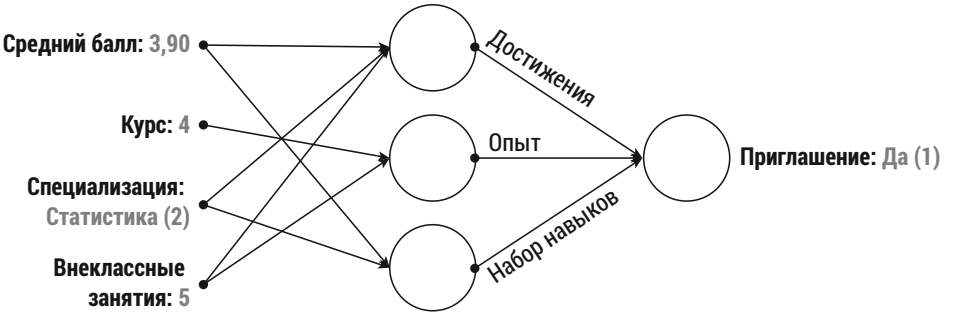


Рис. 12.2. Нейронная сеть со скрытым слоем. Средний слой «спрятан» между основным входным слоем слева и выходным слоем справа

Идея состоит в том, что нейроны в скрытом слое будут «изучать» новые и совершенно иные представления входных данных, что упростит задачу прогнозирования. Давайте посмотрим на верхний нейрон скрытого слоя: находится в середине и состоит из трех нейронов. Предположим, что в результате обучения на исторических данных верхний нейрон «научился» понимать, что средний балл кандидата, его специализация и количество внеклассных занятий — важные факторы, влияющие на вероятность получения приглашения. Это значит, что данные кандидата с высоким средним баллом, большим количеством внеклассных занятий и сложной специализацией заставят этот нейрон «активироваться» и послать сигнал, обозначающий новый признак в данных. Этот признак можно было бы назвать достижением. С математической точки зрения это означает, что в функции активации этого нейрона веса таких параметров, как средний балл, специализация и внеклассные занятия, являются «большими».

Аналогичным образом комбинация количества лет обучения и количества внеклассных занятий может заставить средний нейрон подать сигнал, соответствующий такому признаку, как опыт, а нижний нейрон может активироваться при наличии у студента подходящего набора навыков. Разумеется, все это чисто гипотетически: как и в случае анализа главных компонент, мы категоризируем признаки на основе комбинации полей, которые влияют на них сильнее всего.

Напомним, что четыре исходных входных параметра подаются на скрытый слой и выводятся в виде трех новых признаков. Затем признаки «достижения», «опыт» и «набор навыков» становятся входными данными для последнего нейрона, который берет взвешенную комбинацию этих входных данных, пропускает ее через еще одну функцию активации и выдает прогноз.

С точки зрения вычислений сеть можно рассматривать как серию моделей логистической регрессии, заключенных в каждом из нейронов[128]. В скрытом слое присутствуют три модели логистической регрессии, каждая из которых придает разный вес таким параметрам, как средний балл, курс, специализация и количество внеклассных занятий. (Чтобы не усложнять визуализацию, мы указали не все связи с нейронами скрытого слоя, проигнорировав те, которые имели бы несущественные веса.) Выходные сигналы этих трех моделей становятся входными для последнего нейрона, который взвешивает комбинацию входных данных и выдает окончательный результат.

При этом получаются уравнения в уравнениях — математическое подобие матрешки. Вот как это может выглядеть.

«Внешняя» функция представляет собой функцию активации в последнем слое сети. Для сети, изображенной на рис. 12.2, это будет:

$$\text{Вероятность приглашения} = \frac{1}{e^{-(w_1 \times \textit{Достижения} + w_2 \times \textit{Опыт} + w_3 \times \textit{Набор навыков} + b)}}$$

Однако каждый из признаков в этом уравнении — достижения, опыт и набор навыков — представляют собой отдельные уравнения. Если мы заменим в приведенном выше уравнении только «Достижения», то получим (не пугайтесь):

$$\text{Вероятность приглашения} = \frac{1}{e^{-\left(w_1 \times \left(\frac{1}{1+e^{-(w_{11} \times \textit{Средний балл} + w_{21} \times \textit{Курс} + w_{31} \times \textit{Специализация} + w_{41} \times \textit{ВЗ} + b1)}}\right) + w_2 \times \textit{Опыт} + w_3 \times \textit{Набор навыков} + b\right)}}$$

[128] Здесь мы должны сделать оговорку. Если функция активации не логистическая, то это утверждение неверно.

И это если мы ограничимся только признаком «Достижения»! Мы не стали заменять остальные, но надеемся, что нам удалось проиллюстрировать высказанную ранее мысль о том, что нейронные сети представляют собой гигантские математические уравнения.

Результат этой вложенной структуры — огромное уравнение с множеством параметров, которое принимает входной набор данных и различными способами их комбинирует. Именно наслоение этих функций и позволяет сети выявлять более сложные представления в данных, что делает возможным более детальные предсказания.

Описать процесс работы нейронной сети так же сложно, как и процесс мышления. На практике скрытый слой, скорее всего, не будет выдавать понятные для людей представления, показанные здесь (достижения, опыт и набор навыков). Хуже того, сложность будет возрастать по мере добавления дополнительных слоев и нейронов. Иногда эти модели называют черными ящиками из-за сложности понимания процесса их работы на уровне слоев и нейронов.

Поэтому при объяснении принципа работы нейронных сетей не стоит прибегать к драматическим сравнениям с человеческим мозгом. Более реалистично представление о нейронных сетях как о больших математических уравнениях, обычно используемых при решении задач контролируемого обучения (классификация или регрессия) и способных находить новые представления входных данных, упрощая тем самым процесс прогнозирования.

Что же такое глубокое обучение?

ПРИМЕНЕНИЕ ГЛУБОКОГО ОБУЧЕНИЯ

Глубокое обучение — это семейство алгоритмов, использующих структуру искусственной нейронной сети с двумя или более скрытыми слоями. (Другими словами, это искусственная нейронная сеть с улучшенным брендингом.) Идея углубления (или расширения, как показано на рис. 12.3) нейронной сети состоит в последовательном наложении скрытых слоев, при котором выходные сигналы одного слоя становятся входными сигналами для следующего. В каждом слое реализуются новые абстракции и представления данных, в результате чего из набора входных данных создаются все более нюансированные признаки.

Это сложный процесс, который не всегда было легко осуществить. В 1989 году исследователи под руководством Янна ЛеКуна[129] создали модель глубокого обучения, которая принимала в качестве входных данных рукописные цифры и автоматически присваивала им соответствующую выходную числовую метку. Это было сделано для автоматического распознавания индексов на почтовых отправлениях.

Эта сеть содержала более 1200 нейронов и почти 10 000 параметров. (Только задумайтесь об этом. Модель в уравнении 12.2 содержит всего пять параметров.) Команде исследователей требовался обучающий набор данных, содержащий тысячи рукописных цифр с метками. Все это нужно было осуществить, используя технологии 1980-х годов.

Казалось, что новейшие компьютеры, большой набор размеченных данных и терпение должны были гарантировать успех в сфере глубокого обучения. Но несмотря на некоторый прогресс в этой области, настоящих прорывов пришлось ждать еще несколько лет, потому что (1) обучение глубокой нейронной сети происходило мучительно медленно даже на самых мощных и дорогих компьютерах того времени, и (2) доступ к наборам размеченных входных-выходных данных был ограничен. А на одном терпении далеко не уедешь.

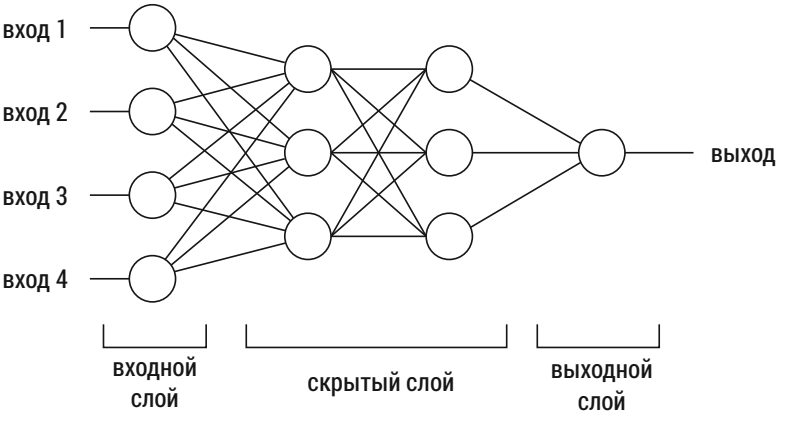


Рис. 12.3. Глубокая нейронная сеть с двумя скрытыми слоями

[129] LeCun, Y., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.

Однако в 2010-х годах сочетание больших наборов данных (благодаря Интернету), усовершенствования алгоритмов (в частности, применения более эффективных по сравнению с логистическими функций активации) и компьютерного оборудования, известного как графические процессоры (GPU), произвело настоящую революцию в сфере глубокого обучения. Графические процессоры позволили сократить время обучения в сто раз[130]. Благодаря этому для изучения тысяч параметров глубокой нейронной сети требовались уже не недели или месяцы, а всего несколько часов или дней. С тех пор успехи в области глубокого обучения росли как снежный ком, особенно в том, что касалось таких неструктурированных данных, как текст, изображения и звук. Это проявилось в создании систем, позволяющих решать всевозможные задачи — от идентификации и маркировки лиц до преобразования аудио в текст.

Преимущества глубокого обучения

Прежде чем приступать к обсуждению того, как методы глубокого обучения позволяют обрабатывать неструктурированные данные, давайте поговорим о том, чем глубокое обучение отличается от алгоритмов, с которыми вы познакомились ранее. Мы уже сказали, что скрытые нейроны способны генерировать новые и более нюансированные представления набора данных, взаимодействия моделей и нелинейные взаимосвязи, тем самым позволяя обнаруживать тонкости, упускаемые другими методами.

С практической точки зрения это может быть невероятно полезно для специалистов по работе с данными, поскольку сокращает время на ручное конструирование признаков.

Конструирование признаков — это процесс объединения или преобразования необработанных данных в новые признаки (новые столбцы) в наборе данных с использованием экспертных знаний. Например, в случае с набором данных, предсказывающим вероятность дефолта по кредиту, создание показателя доступности путем деления стоимости жилья на доход домохозяйства может повысить эффективность модели. Однако этот процесс может оказаться очень трудоемким и запутанным. Благодаря использованию скрытых слоев методы глубокого обучения часто позволяют автоматизировать процесс

[130] См. статью “From not working to neural networking” на странице: **https://www.economist.com/news/special-report/21700756-artificial-intelligence-boom-based-old-idea-modern-twist-not**

конструирования признаков, создавая представления данных, более подходящие для решения задачи прогнозирования.

При использовании большого количества данных, все более глубоких сетей, автоматического конструирования признаков и последовательного наслоения нейронов в данных могут обнаруживаться все более сложные и богатые представления, которые улучшают производительность модели по мере того, как она обучается на все более объемных наборах данных. Это показано на рис. 12.4.

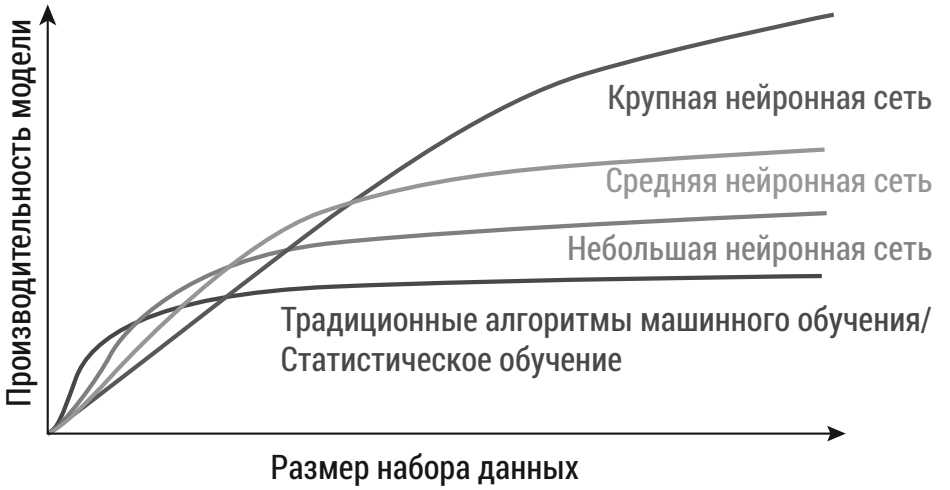


Рис. 12.4. Теоретическое сравнение динамики роста производительности традиционных алгоритмов регрессии и классификации и нейронных сетей разного масштаба по мере увеличения размера набора размеченных данных[131]

На данном изображении показаны теоретические кривые производительности различных алгоритмов и то, что производительность традиционных методов (например, логистической и линейной регрессии) может стагнировать даже при увеличении размера набора размеченных обучающих данных. Линейные методы способны захватывать очень ограниченное количество сигналов. В то же время углубление архитектуры нейронной сети

[131] Изображение взято из статьи **lilianweng.github.io/lil-log/2017/06/21/an-overview-of-deep-learning.html** и вдохновлено изображением из книги Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. Доступ получен через сайт **mlyearning.org**.

позволяет «выжимать» из данных все больше информации и повышать прогностическую эффективность. И по мере увеличения размера набора данных производительность крупных глубоких нейронных сетей может продолжать расти. На практике, разумеется, есть предел, поскольку каждый набор данных ограничен. Из любого лимона в конечном итоге будет выжат весь сок.

Однако по поводу рис. 12.4 следует сделать важную оговорку. Производительность модели будет расти только в том случае, если в данных присутствует значимый сигнал или информация. А гарантировать этого нельзя.

Глубокое обучение с его автоматизированным конструированием признаков и способностью улавливать нюансированные закономерности в данных хорошо справляется с решением задач восприятия. В следующих разделах мы поговорим о том, как это работает.

Как компьютеры «видят» изображения

В предыдущей главе вы узнали о том, как компьютер «читает» текст. В этом разделе вы узнаете, как компьютеры «видят» изображения, а также получите представление о том, как методы глубокого обучения применяются в области компьютерного зрения.

На рис. 12.5 показано, как простое изображение в градациях серого — написанная от руки цифра — воспринимается компьютером[132]. Каждый пиксел изображения был преобразован в значение в диапазоне от 0 (белый цвет) до 255 (черный цвет), который включает все оттенки серого. На рис. 12.5 показано изображение размером 8 на 8 пикселов с низким разрешением, представленное в виде матрицы с 64 значениями в диапазоне от 0 до 255. Люди видят написанную от руки цифру слева, а компьютер — электронную таблицу с числами, показанную в середине.

Теперь представьте себе базу данных, включающую несколько тысяч примеров — рукописных цифр от 0 до 9, отличающихся стилем написания. Люди, включая детей, способны прочитать и распознать эти цифры без особых усилий. Но как компьютер может осуществить классификацию изображений?

[132] Автоматическое распознавание рукописных цифр — это настоящий обряд посвящения для тех, кто стремится освоить методы глубокого обучения. Янн ЛеКун решил эту задачу в 1989 году. Сегодня этот процесс можно реализовать на ноутбуке. База данных рукописных цифр доступна по адресу: **yann.lecun.com/exdb/mnist**.

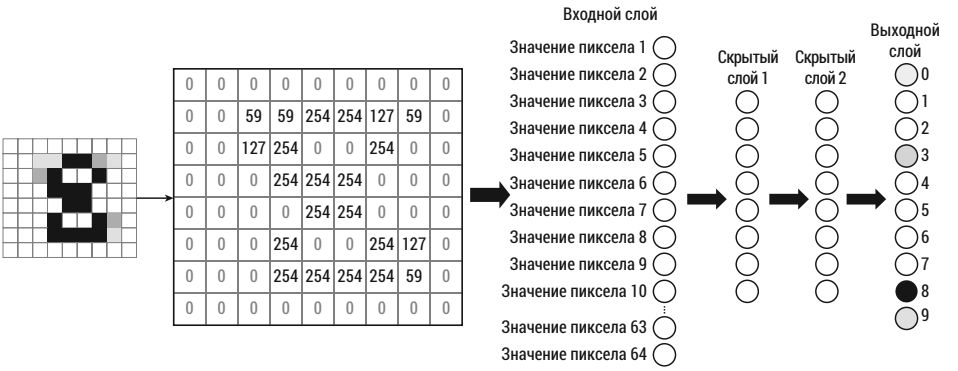


Рис. 12.5. Здесь показано, как изображение в градациях серого «воспринимается» компьютером и как эти данные подаются на вход глубокой нейронной сети. Более темные оттенки в выходном слое указывают на наиболее вероятное предположение

Мы могли бы применить к этому набору данных алгоритм глубокого обучения, способный обучаться на тысячах рукописных цифр. Нейроны скрытого слоя могли бы «активироваться» в том случае, если, скажем, в цифре присутствовала петля (0, 6, 8 или 9), вертикальная (1, 4) или горизонтальная линия (2, 4, 7) или их комбинация.

Здесь мы снова схитрили, чтобы дать вам представление о том, что могут представлять нейроны. Однако, как уже говорилось, скрытые слои зачастую очень трудно интерпретировать, и они могут создавать представления, не имеющие непосредственно воспринимаемого смысла. Но концептуальная идея остается прежней. Внутренние нейроны действительно способны выявлять закономерности в написании цифр, но они имеют только математический, а не визуальный смысл.

Сверточные нейронные сети

Теперь давайте рассмотрим более продвинутый способ анализа изображений с помощью сверточных нейронных сетей, которые используются исследователями для построения систем классификации цветных и крупных изображений, состоящих из большого количества пикселов.

Мы начнем с объяснения того, как компьютер «видит» цветное изображение. Каждый пиксел цветного цифрового изображения состоит из трех цветов — красного, зеленого и синего. Мы называем их цветовыми каналами. Красный канал содержит матрицу значений от 0 (красный отсутствует)

до 255 (красный); то же самое касается зеленого и синего каналов. Таким образом, вместо одной матрицы чисел мы имеем три, как показано на рис. 12.6.

Рис. 12.6. Цветные изображения представлены в виде трехмерных матриц, содержащих значения пикселов красного, зеленого и синего цветов

Соответственно, 10-мегапиксельное изображение будет содержать 30 миллионов значений (значение красного, синего и зеленого цветов для каждого из 10 миллионов пикселов). И если эти 30 миллионов входных данных будут поданы на вход нейронной сети со скрытым слоем, состоящим из 1000 нейронов, то вашему компьютеру потребуется изучить целых 30 миллиардов весовых параметров[133]. Если у вас нет доступа к самому мощному в мире суперкомпьютеру (и даже если бы он у вас был), вам лучше придумать более эффективный способ решения этой задачи.

Исследователи и специалисты по глубокому обучению делают это с помощью процесса, называемого сверткой. Свертка — это математический аналог анализа изображения с помощью ряда увеличительных стекол, каждое из которых предназначено для разных целей. Перемещая увеличительное стекло по изображению слева направо и сверху вниз, вы заметите множество локальных паттернов: линий, углов, закругленных краев, текстур и так далее (рис. 12.7). Свертка осуществляет это математически, то есть выполняет вычисления с использованием локализованного набора значений пикселов, находя края (например, значения 0 рядом с большими значениями) и другие паттерны. Затем она «объединяет» их для нахождения наиболее

[133] Каждый из 1000 нейронов в скрытом слое представлял бы собой взвешенную сумму 30 миллионов входных значений.

выразительных отличительных черт, чтобы уменьшить количество участвующих в процессе чисел.

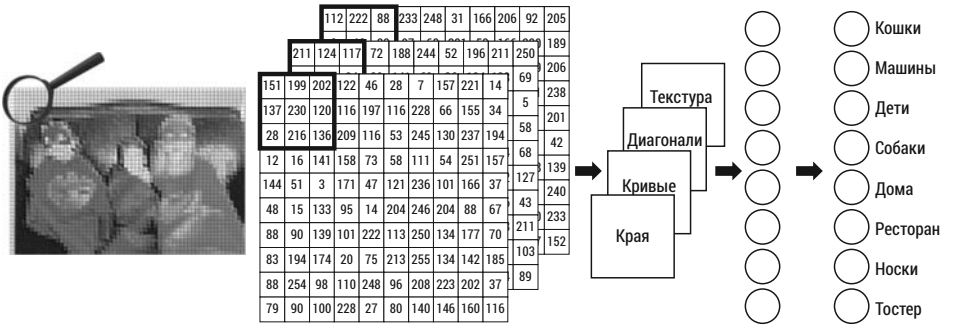


Рис. 12.7. Процесс свертки подобен использованию серии увеличительных стекол для выявления различных форм на изображении, которые подаются в скрытые слои нейронной сети для классификации

После обнаружения локальных паттернов (вроде горизонтальных или диагональных краев) с помощью свертки нейроны скрытого слоя начинают соединять важные края (в математическом смысле) и отфильтровывать информацию, не имеющую отношения к целевому выходу. При этом данные обрабатываются так, что сеть научается определять, есть ли на фотографии дети, а также находить различия в лицах. А в случае с беспилотными автомобилями — отличать припаркованные автомобили от движущихся, пешеходов от строителей, а знак «стоп» от знака «уступи дорогу».

Процесс свертки не только уменьшает количество значений, поступающих в уже знакомую вам структуру нейронной сети (никто не захочет вычислять миллиарды параметров, если этого можно избежать), но и «ищет» похожие признаки на изображениях. В отличие от структурированных наборов данных, признаки в которых имеют фиксированное местоположение в столбцах, признаки на изображениях необходимо не только проанализировать, но и обнаружить. Именно поэтому алгоритмы социальных сетей способны находить ваше лицо на изображении независимо от того, где вы прячетесь.

Глубокое обучение для обработки языка и последовательностей

Глубокое обучение также способствовало прорывам в сфере обработки языка и последовательностей благодаря использованию структуры, называемой рекуррентной нейронной сетью. Как вы помните из предыдущей

главы, традиционные методы анализа текста оказываются неэффективными по причине игнорирования порядка слов, которые просто помещаются в «мешок слов».

Но порядок слов имеет большое значение. Рассмотрим следующие два предложения со словом «апельсиновый». Можете ли вы предсказать последнее слово в каждом из них?

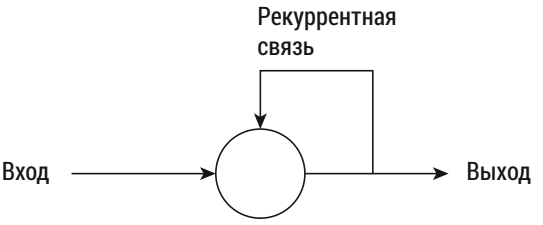


Рис. 12.8. Простое представление рекуррентной нейронной сети

1. За завтраком я люблю пить апельсиновый _____.
2. Мой двоюродный брат живет в калифорнийском «апельсиновом _____».

Вероятно, ваш мозг без промедления вставил пропущенные слова: «сок» и «округ». К тому моменту, когда вы достигли пропущенного слова в первом предложении, слова «завтрак» и «пить» уже находились в вашей кратковременной памяти. Очевидным ответом был «сок».

Во втором примере ваша память о «Калифорнии» и слова «живет в», скорее всего, натолкнули вас на мысль о калифорнийском «апельсиновом округе» (Ориндж Каунти). В обоих случаях ваш мозг удерживал уже известную ему информацию во время обработки новой. Рекуррентная нейронная сеть[134] — вычислительный эквивалент этого процесса.

На рис. 12.8 показана простая рекуррентная нейронная сеть, в которой выходные данные подаются обратно в сеть, в результате чего создается «память».

В случае с вышеописанной задачей классификации текста обучающую выборку образуют тысячи или миллионы входных-выходных пар, пересекающихся последовательностей слов. Например, вход «За завтраком я» будет сопоставлен с выходом «завтраком я люблю». Когда система просматривает

[134] Существует несколько типов рекуррентных нейронных сетей. Самая популярная из них называется сетью с долгой кратковременной памятью (LSTM, Long short-term memory).

последовательности слов в предложении, она также «запоминает» те слова, которые встретились ей ранее. Таким образом, когда сеть видит вход «пить апельсиновый», то выходом, скорее всего, будет «пить апельсиновый сок», если исторические данные содержат предложения с фразой «апельсиновый сок» или «пить сок».

Вы уже наверняка понимаете, как подобные алгоритмы глубокого обучения можно использовать для более быстрого сочинения ответов на письма и исправления грамматических ошибок прямо в процессе набора текста. Эта технология была реализована в функции «Smart Compose» сервиса Google Gmail в 2018 году; она предлагает текст, помогая вам заканчивать свои _____, и работает на основе рекуррентных нейронных сетей[135].

Теперь, когда мы обсудили технические подробности процесса глубокого обучения, давайте рассмотрим его практические аспекты.

ГЛУБОКОЕ ОБУЧЕНИЕ НА ПРАКТИКЕ

По поводу глубокого обучения очень сложно не проникнуться энтузиазмом. Мы коснулись лишь поверхности его потенциала. А крупнейшие мыслители современности убедительно доказывают, что эта технология во многом будет определять наше будущее. Однако этот энтузиазм может отвлечь нас от проблем, свойственных работе с данными.

Есть ли у вас данные?

Какой бы захватывающей ни казалась технология глубокого обучения, вероятно, самую большую трудность для компаний представляет недостаток размеченных обучающих данных. Как говорилось в эпиграфе к этой главе, данные — это «топливо, питающее наши интеллектуальные машины, без которого ничего не было бы возможно». Тем не менее, как уже не раз было сказано, мы снова и снова видим, как многие компании торопятся приступить к глубокому обучению, не имея достаточного объема размеченных данных для решения стоящей перед ними специфической задачи.

Эксперт по глубокому обучению и ИИ Эндрю Ын сформулировал эту проблему так:[136]

---

[135] **www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail**

[136] **deeplearning.ai/the-batch/issue-62**

Главное препятствие для использования преимуществ ИИ в экономике — огромный объем необходимой кастомизации. Чтобы использовать компьютерное зрение для проверки промышленных товаров, нам нужно обучать разные модели для каждого продукта, который мы хотим проверить: для каждой модели смартфона, для каждого полупроводникового чипа, для каждого бытового прибора и так далее.

И для каждой из этих моделей потребуется собственный — и, скорее всего, очень большой набор размеченных изображений.

Трансферное обучение, или как работать с небольшими наборами данных

При наличии некоторого количества размеченных данных, например, не тысяч, а сотен изображений, вашу команду может выручить так называемое трансферное обучение.

Идея трансферного обучения заключается в загрузке модели, обученной распознавать такие повседневные объекты, как воздушные шары, кошки, собаки и так далее[137]. Это означает, что тысячи значений параметров в сети были оптимизированы для работы с группой изображений. Как вы помните, первые слои нейронных сетей, обученных на изображениях, изучают такие общие представления, как формы и линии. А последующие, более глубокие слои соединяют эти края и линии, формируя ожидаемое выходное изображение.

Суть трансферного обучения — выделить несколько последних слоев, которые изучают то, как линии и края образуют, например, изображения кошек и собак, и заменить их новыми слоями, которые в результате очередного раунда обучения становятся способны объединять эти формы в очертания опухолей на медицинских изображениях. Имейте в виду то, что трансферное обучение может уменьшить количество размеченных изображений в 10 раз, но оно не позволяет обойтись несколькими десятками.

[137] Многие практики используют для трансферного обучения модели, обученные на базе данных ImageNet (**https://ru.wikipedia.org/wiki/ImageNet**).

Являются ли ваши данные структурированными?

Ореол таинственности вокруг глубокого обучения в значительной степени обусловлен его прогностической эффективностью в отношении перцептивных данных: изображений, видео, текста и аудио, то есть тех данных, которые мы можем понять, оценить и использовать результаты, не заглядывая в электронную таблицу. В случае со структурированными данными, то есть типичными строками и столбцами, глубокое обучение далеко не всегда может повысить производительность модели.

Если при попытке построения модели контролируемого обучения на структурированном наборе данных ваши специалисты обращаются к глубокому обучению как к методу последней надежды, потому что «все остальное не сработало» — вас, скорее всего, постигнет разочарование.

При работе со структурированными данными глубокие нейронные сети часто проигрывают методам, основанным на использовании деревьев решений (см. главу 10). Без сомнения, существуют исключения, но если точность вашей модели на основе деревьев решений удручающе низкая, то вам лучше потратить время на устранение проблем в ваших данных и оценку принципиальной возможности решения поставленной задачи. (Помните, что наличие размеченных данных еще не гарантирует обнаружения связей между входными и выходными параметрами.) Залог эффективности глубокого обучения — существование отношений между входными и выходными данными. Однако данный метод не может сгенерировать нечто из ничего.

Качество и полнота ваших данных по-прежнему имеют значение.

Как будет выглядеть сеть?

Несмотря на кажущуюся простоту настройки глубоких нейронных сетей, в ходе этого процесса необходимо принять множество решений. Например:

— Сколько слоев должно быть в сети?
— Сколько нейронов должно быть в каждом слое?
— Какие функции активации следует использовать?

Мы не будем останавливаться на этих вопросах (им посвящено множество отличных книг — см. врезку). Просто отметим, что специалисты по работе с данными могут потратить несколько недель на эксперименты с этими

параметрами и общей архитектурой сети. При построении большой сети старайтесь не допускать ее переобучения; в этом вам помогут уроки из глав 9 и 10.

Глубокое обучение для практиков

Если вы хотите научиться самостоятельно создавать модели глубокого обучения, мы настоятельно рекомендуем серию книг Франсуа Шолле, посвященных использованию библиотеки глубокого обучения Keras для языков R и Python.

- Шолле Ф. «Глубокое обучение на Python» (Издательство: Питер, 2018).
- Шолле Ф. и Аллер Дж. Дж. «Глубокое обучение на R». (Издательство: Питер, 2018).

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ВЫ

В завершение этой главы мы хотим немного поговорить об искусственном интеллекте (ИИ) и его приложениях. Как главный по данным, вы должны знать о существовании двух типов искусственного интеллекта. Первый из них — общий искусственный интеллект (ОИИ), призванный воспроизвести процесс человеческого познания. Здесь вы можете вспомнить свой любимый научно-фантастический фильм. Однако прогресс в области ОИИ столь незначительный, что поводов для беспокойства пока нет.

Тем не менее значительный прогресс был достигнут в области искусственного интеллекта узкого назначения (или слабого ИИ). Она охватывает компьютерные системы, которые хорошо справляются с какой-то одной задачей, например, с распознаванием лиц, переводом речи или обнаружением признаков мошенничества. Эффективность слабого ИИ обусловлена эффективностью машинного обучения. Можно сказать, что ИИ — это и есть машинное обучение. Говоря об ИИ, мы на самом деле говорим о машинном обучении. А если задача связана с обработкой перцептивных, неструктурированных данных, то речь идет о глубоком обучении. Машинное обучение — это подраздел ИИ, а глубокое обучение — подраздел машинного обучения (рис. 12.9).

Некоторые люди используют термин ИИ более свободно, чем другие. Например, в обществе принято называть систему рекомендаций фильмов искусственным интеллектом, тогда как в ее основе лежит скорее машинное или статистическое обучение. Почему это важно? Дело в том, что понимание того, что создание «ИИ», о котором говорится в новостях, требует больших наборов данных, собранных у таких людей, как вы и я, ставит вопрос о качестве этих данных, изменчивости, возможной утечке, переобучении и множестве других практических проблем. ИИ усиливает закономерности, содержащиеся в данных, собранных в прошлом; речь не идет о создании чего-то напоминающего человеческое сознание.

Рис. 12.9. Глубокое обучение — это подраздел машинного обучения, которое является подразделом искусственного интеллекта

Преимущества технологических гигантов

Однако существование этой дихотомии объясняется преимуществами технологических гигантов, которые на протяжении многих лет незаметно собирали размеченные данные для своих моделей машинного и глубокого обучения.

Помните, как много лет назад вы щелкали по своим фотографиям в социальных сетях? То же самое делали миллионы других людей, предоставляя этим платформам множество изображений (входных данных) с расположением лиц (выходных данных). Теперь благодаря глубокому обучению система способна нарисовать рамку вокруг вашего лица и отличить вас от вашего

друга. А надоедливые капчи, предлагающие вам доказать, что вы человек, при посещении определенных сайтов («Выберите все изображения с пересечением улиц»), используются для глубокого обучения сетей, лежащих в основе работы систем беспилотных транспортных средств[138]. Возможно, вы решите воздержаться от поездок на беспилотном автомобиле до тех пор, пока веб-сайты не перестанут просить вас идентифицировать знаки «стоп» на изображениях.

При обсуждении глубокого обучения сбору данных уделяется наименьшее внимание, поскольку эта тема гораздо менее захватывающая по сравнению с разговорами о человеческом мозге и автоматической классификации изображений. Но если вас интересует то, как ваша компания может извлечь выгоду из глубокого обучения или машинного обучения вообще, то вашим первым шагом будет сбор размеченных данных. Если у вас есть данные (например, изображения, которые нужно разметить), но вы не хотите тратить на это время — не проблема. Для решения этой задачи создана целая индустрия, и вы можете заплатить сущие копейки за то, чтобы другие люди разметили ваши данные за вас. Так что будущее, в котором можно легко получить доступ к необходимым наборам данных, может быть гораздо ближе, чем кажется.

Этический аспект глубокого обучения

Авторы данной книги — не специалисты по этике и не те люди, которые вправе вести эту дискуссию. С другой стороны, главный по данным не обязан вести дискуссию для того, чтобы в ней участвовать. Поскольку вы находитесь на переднем крае работы с данными, вы должны обеспечить их добросовестное использование.

Объем данных растет гораздо быстрее, чем наша способность формулировать связанные с этим проблемы. Помимо того, что вызывает критику в случае со всеми новыми технологиями, использование данных порождает дополнительные проблемы. Они связаны с нашей ошибочной верой в то, что они всегда отражают непоколебимую истину, и с тем, что алгоритмы зачаровывают нас кажущейся почти человеческой способностью к принятию решений.

---

[138] **medium.com/hackernoon/you-are-building-a-self-driving-ai-without-even-knowing-about-it-62fadbfa5fdf**

И хотя мы не раз подчеркивали, что алгоритмы не воспроизводят процесс человеческого мышления, результаты их применения могут заставить нас в это поверить. Например, хакеры используют такую разновидность алгоритмов глубокого обучения, как генеративно-состязательные сети (GAN), для создания так называемых дипфейков. Это позволяет им накладывать фальшивое лицо поверх лица реального человека, создавая иллюзию того, что этот человек сделал что-то, чего на самом деле не делал. Фейковые новости можно распространять в Twitter, используя для этого реалистичные заголовки, созданные на основе реальных заголовков. Именно так с помощью данной технологии нас можно обмануть.

На более глубоком уровне нам следует проявлять осторожность в том, какие именно человеческие функции мы пытаемся передать алгоритмам глубокого обучения. Например, насколько полезным для судьи может оказаться инструмент, прогнозирующий вероятность рецидива для правонарушителя с помощью ИИ?

Как мы уже говорили, основная причина критики глубокого обучения — огромная путаница с тем, что происходит за кулисами. Очень трудно объяснить гигантское математическое уравнение с миллионами параметров. Однако эти уравнения могут использоваться при вынесении приговора преступникам, лежать в основе работы функции безопасности на телефоне (вроде iPhone Face Scan компании Apple) или применяться системой вашего автомобиля для экстренного торможения или поворота, если на дороге появится олень.

Более того: зачастую то, что мы моделируем, — не просто точки данных, а конкретные люди. Различные аспекты их идентичности кодируются и снабжаются метками. Когда мы получаем данные, они могут не иметь для нас большого значения. Но если мы признаем тот факт, что объем данных растет быстрее, чем наша способность формулировать связанные с этим проблемы, нам не стоит предполагать, будто общество уже дало нам добро на их использование. То, что мы можем собирать определенные признаки и запускать алгоритмы, не всегда значит, что нам стоит это делать. И хотя мы предоставили вам инструменты для понимания хорошо сконструированных приложений глубокого обучения, вам не следует предполагать, что в каждом приложении это сделано правильно. Даже в своей организации стоит скептически относиться к заявлениям о том, что глубокое обучение решает проблемы. Попросите не просто показать вам данные и алгоритмы, но и спросите, кого именно затрагивает полученный результат, а затем решите, насколько вас это устраивает.

Короче говоря, машины становятся умнее, и вам не следует от них отставать. Не воспринимайте собственную роль в процессе использования данных для улучшения результатов бизнеса и общества в целом как нечто само собой разумеющееся.

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы объединили многое из описанного в предыдущих главах, чтобы объяснить принцип работы алгоритмов глубокого обучения. Помните, что в основе глубокого обучения лежат искусственные нейронные сети, состоящие из нейронов, каждый из которых содержит уравнение, называемое функцией активации. Выходной сигнал каждого слоя поступает на вход одного или нескольких нейронов. Эти слои становятся подфункциями для последнего слоя, который, в свою очередь, превращается в одно большое (и впечатляющее!) математическое уравнение, служащее прогностической моделью.

Глубокое обучение — это захватывающая новая глава в машинном обучении. Запуск все более сложных моделей с каждым днем становится все проще и дешевле. Тем не менее, несмотря на потенциал глубокого обучения, нам не следует возлагать на него слишком большие надежды. Данная технология позволяет эффективно решать такие задачи восприятия, как классификация изображений и текстов, состоящих из высококачественных и правильно размеченных данных, но не всегда оказывается оптимальным вариантом для решения небольших задач, предполагающих работу со структурированными данными.

В конечном итоге модели запускают люди. Не позволяйте ореолу таинственности, окружающему алгоритмы глубокого обучения, заставлять вас думать, будто они умнее вас, и верить, что вы используете их нейтральным способом. В конце концов, это ваша работа, и вы должны чувствовать себя комфортно, выполняя ее.

ЧАСТЬ
IV

Гарантируйте успех

В части IV вы узнаете о том, как извлечь максимальную пользу из своего пути становления главным по данным, учась на чужих ошибках, как технических, так и связанных с человеческим фактором.

Эта часть состоит из следующих глав:

Глава 13. Остерегайтесь ловушек.
Глава 14. Знайте людей и типы личностей.
Глава 15. Что дальше?

ГЛАВА 13

Остерегайтесь ловушек

«Первый принцип — не обманывать себя, а себя обмануть легче всего»
_ *Ричард Ф. Фейнман, лауреат Нобелевской премии по физике*

Чтобы понимать, думать и говорить на языке данных, очень важно знать об ошибках, которые вы можете допустить, если потеряете бдительность в ходе работы с ними и их интерпретации. Некоторые подводные камни довольно легко устранить, но их трудно обнаружить, если вы не знаете, где искать. А если вы не будете соблюдать осторожность, то они могут обернуться крупными катастрофами, подобными тем, которые были описаны ранее в этой книге (вспомните аварию шаттла «Челленджер» и крах рынка жилья в 2008 году).

В этой главе мы напомним вам о тех ловушках, о которых вы уже знаете, и познакомим вас с несколькими другими распространенными подводными камнями, которые могут сорвать всю вашу работу или (что еще хуже) заставить вас прийти к неверным выводам.

Прежде чем мы начнем, стоит отметить, что обсуждать чужие ошибки и просчеты в работе с данными очень легко и весело. Однако, хотя мы призываем вас скептически относиться к работе, проделанной другими специалистами в вашей области, стоит помнить о том, что позитивные изменения невозможны без проявления сочувствия и поддержки. Ошибки случаются — и надо отметить, что авторы этой книги пришли к знаниям, изложенным в этой главе, далеко не легким путем. Поэтому давайте признаем, что в основе большинства ловушек лежат не чьи-то гнусные намерения

и недобросовестность. Чаще люди просто не знают о том, что может пойти не так. Именно об этом мы и поговорим в данной главе.

ПРЕДВЗЯТОСТИ И СТРАННОСТИ В ДАННЫХ

Предвзятость — это сложная тема, затрагивающая различные дисциплины. Под предвзятостью мы понимаем однобокое (а иногда даже непоследовательное) предпочтение, отдаваемое идеям и концепциям отдельными людьми и подкрепляемое их группами. В этом разделе мы обсудим распространенные варианты предвзятости в мире данных, а также такие явления, когда при первом взгляде на данные у вас может сложиться одно впечатление, а при повторном их рассмотрении — другое.

Систематическая ошибка выжившего

Представьте, что инвестиционная компания в одном и том же году запускает десятки взаимных фондов, каждый из которых содержит случайный набор акций. Если фонд не покажет целевую доходность за определенный период времени (например, если доходность индекса S&P 500 составит 10%, а доходность одного из фондов — только 3%), то его деятельность будет прекращена. По прошествии нескольких лет останутся только «выжившие» взаимные фонды, отличающиеся впечатляющей доходностью. И тут появляется потенциальный инвестор в вашем лице. Вам демонстрируют показатели фондов компании, превышающие рыночные на протяжении нескольких лет подряд.

Вы бы инвестировали в них свои средства?

Возможно. Компании отказываются от плохих активов, что по своей сути совсем не плохо. Плохо — делать вид, что плохих активов никогда и не существовало, поскольку это создает предвзятость. В этом примере вам не были представлены данные о низкодоходных фондах, потому что компания от них отказалась. Из-за этого результаты деятельности компании показались вам более впечатляющими и заставили вас поверить в то, что в ней работают опытные финансовые аналитики, тогда как наиболее правдоподобное объяснение — простое везение.

Это пример систематической ошибки выжившего, которая представляет собой «разновидность систематической ошибки отбора, когда по одной

группе объектов (условно называемых «выжившие») данных много, а по другой («погибшие») — практически нет»[139].

Классический пример систематической ошибки выжившего — случай статистика Абрахама Вальда, которому было поручено минимизировать потери флота бомбардировщиков союзников во время Второй мировой войны. Самолеты, пережившие жестокие бои, возвращались на базу с серьезными повреждениями и пулевыми отверстиями в фюзеляже и крыльях. Изначально идея заключалась в том, чтобы укрепить те места самолетов, в которых наблюдалось больше всего повреждений. Однако Вальд посчитал ее проявлением ошибки выжившего. Дело в том, что во внимание принимались только вернувшиеся самолеты. Но как быть с теми, которые не смогли вернуться? Что этот характер повреждений говорит о них?

Рекомендация Вальда казалась парадоксальной: он предложил бронировать те участки, которые имели наименьшие повреждения у вернувшихся самолетов. Почему? Потому что самолеты, получившие повреждения в этих местах, так и не вернулись на базу.

Регрессия к среднему

Регрессия к среднему — это явление, суть которого формулируется достаточно просто: за экстремальными значениями случайной величины часто следуют менее экстремальные. Это наблюдение было впервые сформулировано как «регрессия к посредственности» в 1886 году сэром Фрэнсисом Гальтоном[140], который заметил, что дети высоких родителей оказываются менее высокими, чем они (что говорит о регрессии данного показателя), а дети низкорослых родителей — не такими низкорослыми. По сути, он выявил естественную, глубинную стабильность, существующую в росте людей и их потомков: за экстремальными значениями (низкими и высокими) обычно следуют не столь экстремальные (не такие низкие и не такие высокие) значения.

Хотя этот пример может показаться очевидным, регрессия к среднему имеет более широкие последствия для процесса рассуждения. Если вы не смотрите на все имеющиеся данные с высоты птичьего полета, некоторые наблюдения могут показаться экстремальными. В этом случае предвзятость

[139] **https://ru.wikipedia.org/wiki/Систематическая_ошибка_выжившего**

[140] Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246–263.

может проявиться в том, что вы примете решение, основанное на этих экстремальных событиях, не принимая во внимание то, что независимо от ваших действий на горизонте находится более предсказуемое событие, близкое к истинному среднему значению.

Возьмем, к примеру, игрока Национальной футбольной лиги (NFL), который после выдающегося года оказался на обложке популярной видеоигры *Madden NFL*, но в следующем году показал гораздо менее впечатляющие результаты. Этот феномен получил название «проклятья Madden»[141]. Но мы называем это регрессией к среднему значению. Или представьте себе в целом благонадежного сотрудника, у которого выдался тяжелый год, в результате чего его работа получила плохие отзывы. Для него составляется план исправления, и в следующем году его производительность восстанавливается. Менеджер приписывает это улучшение своему мудрому руководству, однако показатели работника, скорее всего, в любом случае улучшилась бы из-за регрессии к среднему значению.

Регрессия к среднему призывает нас не верить выбросам. Ни удача, ни неудача не продлится вечно.

Парадокс Симпсона

Еще одно явление, на которое следует обратить внимание, — парадокс Симпсона. Это потенциально катастрофическая ловушка при работе с данными наблюдений (с которыми вам предстоит работать чаще всего). Парадокс Симпсона возникает в том случае, когда тенденция или связь между переменными меняется на противоположную после добавления третьей переменной. В связи с парадоксом Симпсона вам следует остерегаться двух ошибок: принятия корреляции за причинно-следственную связь и выявления неправильной корреляции.

Рассмотрим данные в табл. 13.1, взятые из исследования 1986 года, посвященного двум типам хирургических методов удаления камней в почках[142]. Обзор медицинских записей показал, что новая, минимально инвазивная процедура удаления камней в почках является более успешной (83%)

---

[141] **https://ru.wikipedia.org/wiki/**Проклятье**_Madden**

[142] Этот пример был впервые использован в работе: Julious, S. A., & Mullee, M. A. (1994). Confounding and Simpson's paradox. *Bmj, 309*(6967), 1480–1481. Мы обнаружили его в отличной книге: Reinhart, A. (2015). *Statistics done wrong: The woefully complete guide*. No Starch Press.

по сравнению с традиционной (78%). Результаты были статистически значимыми и, по общему мнению, вполне убедительными.

Табл. 13.1. Показатели успеха хирургических процедур удаления камней из почек

Способ лечения	Общий показатель успеха
Традиционная процедура	78%
Новая процедура	83%

К сожалению, в этих данных возник парадокс Симпсона. Дальнейший обзор данных показал, что при разбивке камней в почках по размерам, результат меняется на противоположный. Как оказалось, традиционная процедура отличалась высокими показателями успеха как у пациентов с небольшими камнями в почках (диаметром <2 см), так и у пациентов с большими камнями (диаметром ≥ 2 см). Эта разбивка показана в табл. 13.2.

Табл. 13.2. Парадокс Симпсона на примере показателей успеха хирургических процедур удаления камней из почек

Способ лечения	Небольшие камни в почках	Большие камни в почках	Общий показатель успеха
Традиционная процедура	93%	73%	78%
Новая процедура	87%	69%	83%

Как это возможно? Дело в том, что новая процедура была опробована на множестве пациентов с небольшими камнями в почках (то есть на предположительно более легких случаях), в то время как традиционная процедура в основном использовалась для лечения пациентов с более крупными камнями в почках. Несмотря на то что традиционная процедура показала лучшие результаты при удалении небольших камней (93%), новая процедура была выполнена гораздо большему количеству пациентов, а показатель ее успешности составил 87%. Таким образом, общий показатель успеха новой процедуры тяготеет к 87%. В табл. 13.2 мы видим, что общий показатель успешности традиционной процедуры (78%) больше тяготеет к показателям успеха у пациентов с крупными камнями в почках (73%). Новая процедура сработала хуже на этой группе, но она была выполнена слишком небольшому количеству пациентов, чтобы это повлияло на ее общий показатель

успешности. Запутались? Это нормально. Именно поэтому данный феномен и называется парадоксом.

Чтобы снизить риски, связанные с парадоксом Симпсона, разделите наблюдения по группам случайным образом, чтобы исключить смешивание. Другими словами, соберите экспериментальные данные.

Предвзятость подтверждения

Такая ловушка, как предвзятость подтверждения, представляет потенциальную опасность для любого проекта по работе с данными. Ее суть заключается в такой интерпретации данных и результатов, которая подтверждает уже существующие убеждения; при этом противоречащие этим убеждениям доказательства отбрасываются.

В подобной предвзятости легко обвинить руководителей высшего звена, политиков и лиц, заинтересованных в результате деятельности бизнеса, — но признать это за собой гораздо труднее. Тем не менее для многих команд аналитиков предвзятость подтверждения — практически неотъемлемая часть образа жизни. Дело в том, что порой им приходится искать доказательства правильности шагов руководства, которые могут предприниматься еще до анализа достаточного количества данных. По крайней мере, часть работы этих команд направлена на формирование предвзятости подтверждения. Это не просто, но, как главный по данным, вы должны стремиться преодолеть эту предвзятость и сообщить о результатах максимально правдиво. В противном случае команда может использовать предвзятость подтверждения для обоснования решений руководства вместо того, чтобы анализировать все доступные решения, не подвергаясь давлению с его стороны.

Ловушка невозвратных затрат

Суть ловушки невозвратных затрат — желание продолжать работу над проектом, в который уже было вложено огромное количество времени, денег, ресурсов и усилий. В такой ситуации очень трудно отказаться от результатов, даже если вы понимаете, что:

— У вас нет нужных данных для реализации проекта.
— У вас нет подходящей технологии для реализации проекта.

— Исходное содержание проекта не охватывает его основополагающие достоинства.

Некоторые компании предпочли бы, чтобы вы предоставили хоть какие-то результаты, оправдывающие затраченное время и усилия. Подобное давление создает благодатную почву для формирования многих из перечисленных выше предвзятостей.

Алгоритмическая предвзятость

По мере автоматизации все большего количества решений с помощью машинного обучения мы все чаще сталкиваемся с так называемой алгоритмической предвзятостью[143], буквально встроенной в мир данных и вычислений. Хотя исследователи и организации лишь недавно начали внимательно изучать ее происхождение и последствия, такая предвзятость существовала в данных всегда. Часто она является продуктом статус-кво, и ее может быть трудно обнаружить до тех пор, пока этот статус-кво не будет подвергнут фундаментальному пересмотру. Однако, если вы будете осознанно подходить к своей работе, то сможете обнаружить эту предвзятость гораздо раньше.

Вспомните пример из предыдущих глав, в котором мы рассматривали данные о кандидатах на стажировку и пытались предсказать, получат ли они приглашение на собеседование. Если бы набор данных включал такую категориальную переменную, как пол, и исторически мужчины получали бы приглашение на интервью чаще, чем женщины, то каждый алгоритм выявлял бы эту взаимосвязь и чаще отдавал предпочтение соискателям-мужчинам. Для алгоритма существуют лишь единицы и нули, но главные по данным должны знать, что подобная предвзятость имеет место даже в таких ведущих технологических компаниях, занимающих передовые позиции в области машинного обучения, как Amazon[144].

Имейте в виду: какими бы ни были ваши намерения, алгоритмическая предвзятость встречается повсеместно. Прогнозы моделей не являются истиной

---

[143] Алгоритмическая предвзятость: **https://en.wikipedia.org/wiki/Algorithmic_bias**

[144] В статье Reuters 2018 года “Amazon scraps secret AI recruiting tool that showed bias against women” говорится о том, что алгоритмы компании занижали баллы тем кандидатам, в чьих резюме содержалось слово «женский» и названия женских колледжей. **www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G**

в последней инстанции. Все результаты — это продукты предположений. И вы должны действовать так, будто все данные наблюдений предвзяты, потому что так оно и есть. Делая прогнозы, модели подтверждают и усиливают предвзятость и стереотипы, уже присутствующие в данных. Не стоит дожидаться изменения образа мышления, чтобы начать разбираться в предубеждениях, присущих вашей собственной работе. К этому следует приступить уже сегодня[145].

Прочие предубеждения

В этом разделе содержится далеко не полный список предубеждений, парадоксов и странностей, встречающихся в данных. Мы рекомендуем вам обращать внимание на ловушки, не принадлежащие ни одной из категорий. Если вы будете искать лишь конкретные типы предвзятостей или логических ошибок, то можете упустить другие менее заметные предубеждения, которые общество еще не определило. То, что эти ловушки не определены, не значит, что их нет.

БОЛЬШОЙ СПИСОК ЛОВУШЕК

Теперь, когда вы познакомились с некоторыми распространенными предубеждениями и когнитивными ловушками, присущими работе с данными, давайте поговорим о более конкретных подводных камнях, которых следует избегать при реализации подобных проектов. Мы разделили этот большой список на две категории: ловушки статистики и машинного обучения и ловушки проекта.

Ловушки статистики и машинного обучения

Этот раздел содержит список ловушек статистики и машинного обучения, многие из которых мы обсуждали ранее.

Принятие корреляции за причинно-следственную связь. Не поддавайтесь искушению построить причинно-следственный нарратив вокруг коррелированных переменных. Рост уровня продаж компании

[145] В этом вам может помочь ресурс Брукингского института: **www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms**

может коррелировать с увеличением количества просмотров ее рекламы на YouTube, но это не значит, что одно вызвало другое. Как правило, следует избегать разговоров о причинно-следственных связях, если только вы не разработали процесс сбора и анализа данных специально для их поиска (то есть если вы не используете экспериментальные данные). Эти идеи обсуждались в главах 4 и 5.

***p*-хакинг. Представьте, что в некой статье говорится:** «Люди, которые пьют слишком много кофе, имеют повышенный риск развития рака желудка. Результат статистически значим на уровне значимости 0,05»[146]. Как говорилось в главе 7, при уровне значимости 0,05 результаты бывают ложноположительными в 1 случае из 20. *p*-хакинг — это процесс тестирования нескольких закономерностей в данных вплоть до обнаружения статистически значимого *p*-значения. Связь между употреблением кофе и раком желудка была бы менее пугающей, если бы впоследствии вы узнали, что исследователи также изучили взаимосвязь между употреблением кофе и раком мозга, раком мочевого пузыря, раком молочной железы, раком легких или любым другим из 100 видов рака. По чистому совпадению в пяти случаях было бы обнаружено статистически значимое *p*-значение, даже если бы никакой связи не существовало. Учтите, что *p*-хакинг — это разновидность ошибки выжившего, поскольку при этом сообщаются только значимые *p*-значения.

Нерепрезентативная выборка. Результаты опросов во время выборов, которые не представляют голосующее население, будут неверными. Опрос посетителей страницы вашей компании в социальной сети может не отражать мнение большинства ваших клиентов. Не бойтесь спорить с данными (глава 4), поскольку формирование политики или принятие решений на основе выборки данных, не представляющей совокупность, на которую они будут влиять, может привести к серьезным ошибкам. Хуже того, такие данные могут создавать чувство ложного спокойствия, заставляя вас думать, будто вы принимаете обоснованное решение, тогда как отсутствие данных наверняка было бы предпочтительнее использования тех некачественных данных, с которыми вы работаете.

[146] Это всего лишь пример. Авторы данной книги не занимаются исследованием рака.

Утечка данных. Не обучайте модель на данных, недоступных во время прогнозирования. Вам может показаться, что у вашей команды превосходная модель, но она на самом деле может быть совершенно бесполезной. Предсказать, купит ли посетитель вашего сайта продукт, довольно легко, если вы знаете, что при покупке он применил код купона. Главные по данным должны убедиться в наличии каждого признака в модели в момент принятия решения (см. главы 9 и 10).

Переобучение. Как вы помните, модели — это упрощенные версии реальности. Они используют то, что мы знаем, чтобы помочь нам предсказать то, чего мы не знаем. Когда модель хорошо работает на данных, с которыми она уже сталкивалась, но не способна предсказать новые наблюдения, можно сказать, что модель «переобучена». В некотором смысле модель «запоминает» сценарии, определенные обучающими данными, вместо того, чтобы «учиться» на этих обучающих данных и делать прогнозы относительно того, что ей неизвестно (см. главы 9 и 10). Главные по данным могут предотвратить переобучение, разделив данные на обучающие и тестовые наборы. Обучите модель на обучающем наборе и оцените эффективность ее предсказаний на тестовом наборе.

Нерепрезентативные обучающие данные. Эта ловушка предполагает использование «нерепрезентативной выборки» для создания модели машинного обучения. Модели знают только те данные, на которых они были обучены. Модель, обученная на данных о недвижимости в Огайо для предсказания цены продажи домов в Огайо, не способна предсказать цену аренды квартиры в Нью-Йорке. Точно так же интеллектуальный динамик с голосовым помощником, обученный на образцах звука, записанных в студии звукозаписи, может испытывать сложности с разбором команд в шумном доме. Чтобы избежать этой ловушки, главные по данным должны тщательно продумать обстоятельства использования их модели и собрать отражающие их обучающие данные.

Ловушки проекта

В этом разделе мы поговорим о ловушках, в которые можно угодить при реализации проекта по работе с данными.

Отказ от постановки острых вопросов (или решение неправильной задачи). Даже небольшая двусмысленность может привести к путанице и рассогласованию между группой специалистов по работе с данными, бизнес-группой и заинтересованными сторонами проекта. Убедитесь в том, что все четко понимают решаемую бизнес-задачу (глава 1).

Вопрос не адаптируется после провала. Прояснить бизнес-задачу важно — как и быстро признать то, что ее нельзя решить. Многие команды специалистов по работе с данными быстро обнаруживают недостатки исходного вопроса, но продолжают двигаться вперед, подчиняясь внешнему давлению. Чтобы избежать возникновения несоответствий, вопрос необходимо скорректировать.

Данными владеют, а не управляют (то есть данные трудно получить). В некоторых организациях определенные группы (например, ИТ-отдел, финансовый отдел или бухгалтерия) владеют данными, которые требуются вам для работы. Хотя многие из этих организаций практикуют управление данными на бумаге, получить к ним доступ бывает непросто. Ваша компания должна понимать, что при ограничении доступа к данным вы можете сделать не так много.

Данные не содержат необходимой информации. Данные могут быть легко доступными и «опрятными», но они могут не содержать информацию, необходимую для решения поставленной задачи. Если данные не содержат нужной вам информации, постарайтесь собрать более качественные данные.

Отказ от использования недорогих инструментов и технологий с открытым исходным кодом. Прежде чем взяться за реализацию масштабного проекта, связанного с внедрением какой-либо новой технологии, потратьте время на прототипирование. Вполне возможно, что инвестирование в платформу обработки данных для управления будущими операциями изменит очень многое для вашей команды. Однако прежде чем тратить деньги, попробуйте создать минимально жизнеспособный продукт с помощью Microsoft Excel или таких бесплатных технологий с открытым исходным кодом, как R или Python.

Слишком оптимистичные сроки. Проекты по работе с данными часто проваливаются совершенно неожиданным образом. Описанные выше проблемы обнаруживаются только через несколько недель после запуска проекта, а жесткие сроки приводят к срезанию углов и плохому анализу. Сроки реализации проекта должны учитывать неизбежные задержки при работе с данными.

Завышенные ожидания относительно ценности. Компании привыкли многого ожидать от науки о данных, статистики и машинного обучения. Говорите о ценности, которую может принести ваш проект, открыто, но не преувеличивайте ее, чтобы не вызвать отрицательную реакцию; она может негативно сказаться на текущих и будущих проектах.

Ожидание предсказания непредсказуемого. Некоторые вещи невозможно предсказать вне зависимости от количества собранных исторических данных. Документирование каждого вращения каждого колеса рулетки в Лас-Вегасе не поможет вам предсказать результат следующего вращения.

Выход за рамки разумного. Как и вы, авторы этой книги любят работать с данными. Многие из нас готовы ухватиться за очередную идею. Однако очень часто из виду упускается нечто совершенно очевидное: наука о данных, статистика, машинное обучение и ИИ могут решить многие важные проблемы, но далеко не все. При работе с данными, статистикой и алгоритмами нередко можно выйти за рамки разумного. Вы можете задействовать алгоритм классификации для определения бизнес-правил. Однако иногда у нас уже есть набор правил, в соответствии с которыми мы действуем. В таких случаях будет гораздо проще, если окружающие вас люди просто запишут их. По сути, если ваша команда может написать бизнес-правила для автоматизации процесса, то вашу работу можно считать выполненной. В настоящее время эта идея теряется на фоне шумихи вокруг науки о данных. Машинное обучение кажется привлекательным для руководства, но иногда оно — просто излишество.

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы рассмотрели распространенные заблуждения и ловушки. Как уже было сказано, представленный список не является исчерпывающим. И вы должны исходить из того, что такого списка в принципе не существует. Помните, что объем данных растет быстрее, чем наша способность формулировать порождаемые этим ростом проблемы и возможности. Если вы примете эту идею, то поймете, что ни один список не может включить в себя все те ловушки, в которые люди еще не попадали. Однако в этой главе мы предоставили вам отправную точку.

Проекты часто завершаются неудачно. И, скорее всего, у вас будет по крайней мере один неудачный проект, с которым вы будете ассоциироваться (вероятнее всего, их будет гораздо больше). Будьте открыты и откровенны, когда случаются неудачи, и по возможности переключайтесь на реализацию новых идей. Ваш опыт станет вашим лучшим учителем.

ГЛАВА 14

Знайте людей и типы личностей

«Люди переживают, что компьютеры станут слишком умными и захватят мир, но настоящая проблема в том, что они являются слишком глупыми и уже его захватили»[147]

_ *Педро Домингос, исследователь ИИ*

В предыдущей главе вы узнали о распространенных ловушках, в которые можно угодить при реализации проекта по работе с данными. В этой главе мы поговорим о людях и их ролях, а также о том, сколько проектов терпят неудачу не из-за технологий или данных, а из-за конфликтующих личностей и неэффективного общения.

Именно недостатки коммуникации стали причиной провала многих из описанных в этой книге проектов. Наша цель — научить вас ориентироваться в коммуникативных красных флажках благодаря пониманию особенностей личностей, вовлеченных в проект. В этой главе мы обсудим убеждения ключевых фигур и рассмотрим сценарии того, что происходит при нарушении коммуникации между специалистами по работе с данными и бизнес-профессионалами. Понимание ролей других людей и проявление сочувствия поможет вам, как главному по данным, устранить любые пробелы в общении.

В следующем разделе мы рассмотрим дополнительные наблюдения, касающиеся специалистов по работе с данными и бизнес-профессионалов,

[147] Цитата из интервью: **www.washington.edu/news/2015/09/17/a-q-a-with-pedro-domingos-author-of-the-master-algorithm**

и выделим сценарии, в которых пробелы в общении приводят к провалу проектов. Затем мы поговорим о разном отношении людей к данным — энтузиазме, цинизме и скептицизме.

СЕМЬ СЦЕН КОММУНИКАТИВНОГО СБОЯ

Когда коммуникация нарушается в ходе реализации проекта по работе с данными[148], вы можете стать свидетелями одной из семи сцен, описанных в табл. 14.1. В следующих разделах мы подробно рассмотрим сценарии для каждой из них, которые могут показаться вам весьма знакомыми.

Табл. 14.1. Семь сцен коммуникативного сбоя

Сцена	Краткое описание
Постмортем	Старшему дата-сайентисту поручается вернуть проект в нужное русло спустя долгое время после появления первых признаков опасности. Но уже слишком поздно.
Время историй	Умный аналитик исключает технические нюансы из своей презентации, подчиняясь мифу о том, что руководителям нужно объяснять все так же, как детям. При этом аналитик чувствует, что предает свою роль как критического мыслителя.
Игра «Телефон»	Предварительная статистика вырывается из контекста, а затем распространяется настолько широко, что теряет то небольшое значение, которое она изначально имела.
В дебри	Результаты анализа настолько технические, что они утрачивают всякий смысл. Их представление больше похоже на самолюбование, чем истинное описание ситуации.
Проверка реальности	Специалист по работе с данными стремится усовершенствовать непрактичное решение и не рассматривает альтернативы до тех пор, пока этого не потребует руководитель.
Захват власти	Дата-сайентист пытается решить основополагающие бизнес-проблемы, не заботясь о доверии команды или сосредотачиваясь на быстрых победах.
Хвастун	Дата-сайентист придирается к результатам чужой работы, поэтому его больше не привлекают к участию в проектах.

[148] На создание этого раздела нас вдохновила статья Скотта Беринато «Data Science and the Art of Persuasion» (**hbr.org/2019/01/data-science-and-the-art-of-persuasion**), основанная на нашем опыте и опыте наших коллег по бизнесу, которые любезно поделились своими историями.

Постмортем

Важный проект, реализуемый телекоммуникационной компанией, застопорился после шести месяцев работы.

Перед командой проекта, состоящей из одного дата-сайентиста, была поставлена задача прогнозирования оттока клиентов. Ему нужно было предсказать, переключится ли клиент на нового оператора сотовой связи в следующем году. Для этого была разработана модель, которая оценивает всех текущих клиентов компании, основываясь на исторических данных: для клиента_1 вероятность сменить оператора связи составляет 85%, для клиента_2 — 10% и так далее.

На бумаге работа завершена. Модели можно использовать. Код запущен в производство. Но есть маленькая (ну, может, не такая уж маленькая) проблема: модель далеко не так точна, как команда обещала заинтересованным сторонам.

Руководитель проекта на протяжении последних нескольких недель избегал обсуждения текущих проблем с дата-сайентистом, полагая, что они незначительны и легко решаемы. (Компьютеры могут все, верно?) Но проблемы оказались гораздо серьезнее, чем предполагалось, и руководитель начинает нервничать. Возглавить проект предлагают еще более высокопоставленному специалисту по работе с данными.

Но уже слишком поздно.

К этому моменту приняты уже сотни решений, и эксперт не может даже начать распутывать клубок проблем, учитывая, что до представления результатов высшему руководству остается всего неделя. Эксперт не только повторяет опасения дата-сайентиста, но и пополняет список проблем.

Потратив еще один 12-часовой рабочий день на спасение того, что осталось от проекта, старший специалист по работе с данными вспоминает цитату выдающегося статистика Р. Э. Фишера: «Вызвать статистика после завершения эксперимента — это то же самое, что попросить его провести вскрытие и сказать, по какой причине эксперимент закончился неудачей». На руководителя проекта возложена неприятная обязанность сообщить о проблеме высшему руководству.

Время историй

Талантливая студентка, изучавшая сложные технические концепции в университете на протяжении последних пяти лет, работает над своим первым

крупным проектом в маркетинговой фирме. За день до представления результатов ее менеджер предлагает ей представить результаты проведенного анализа в виде «истории» и сократить ее содержание до одного слайда PowerPoint.

«С ними надо говорить, как с пятиклассниками», — заявляет менеджер.

Она неохотно соглашается, хотя знает, что в аудитории будут ученые. Она считает, что презентация уже была достаточно сокращена. Кроме того, она проверила понятность своего выступления на технически не подкованных коллегах.

«Поверь мне, — говорит менеджер, — тебе не нужно, чтобы эта группа задавала какие-либо вопросы». Из-за отбрасывания большинства технических деталей и критических рассуждений результаты работы сводятся к простому заголовку.

Во время представления результатов в ходе презентации большинство зрителей кивает. Некоторые задаются вопросом: «Нам действительно нужны дата-сайентисты, находящие простые ответы?»[149] Другие зрители, обладающие некоторыми техническими знаниями, недоумевают, почему были проигнорированы технические аспекты проекта.

Обдумывая свою презентацию, студентка понимает, что многие нюансы были потеряны, — и чувствует, что в некотором смысле она предала свою исходную работу.

Игра «Телефон»

Во время случайной встречи за чашкой кофе дата-сайентист рассказывает своему коллеге об интересном факте, обнаруженном в данных компании. У нее еще не было возможности разобраться во всем как следует, но беглый обзор показал, что 75% участников опроса заявили о своем намерении стать постоянными клиентами.

После встречи дата-сайентист возвращается к своему столу и снова просматривает результаты анализа. Она еще раз видит показатель 75% и понимает, что в опросе приняли участие всего 8 человек из нескольких сотен. Затем она выясняет, что соответствующий вопрос был добавлен в анкету совсем недавно, так что ни один из тех людей, которые высказали намерение стать постоянными клиентами, еще не стал таковым.

---

[149] Подробнее об этом мнении можно узнать в статье Джеффа Лика “Data science done well looks easy — and that is a big problem for data scientists” на сайте: **https://simplystatistics.org/posts/2015-03-17-data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/.**

Месяц спустя на общем собрании компании руководители хвастаются своими успехами в деле удержания клиентов. Они говорят о том, что 3 из 4 клиентов стали постоянными, судя по результатам опроса сотен людей.

Дата-сайентист понимает, что этот факт был озвучен во время случайной встречи за чашкой кофе и никогда не должен был распространяться без проверки. К этому моменту он повторялся в компании так много раз, что стал восприниматься как нечто самоочевидное. Дата-сайентист задается вопросом о том, можно ли как-то остановить его использование в компании — и стоит ли вообще это делать.

В дебри

Дата-сайентист применяет подходящие методы для решения сложной проблемы и, по общему мнению, отвечает на поставленные бизнес-вопросы. Но его итоговая презентация для проектной группы оказывается слишком технической. Он не предпринял практически никаких усилий для того, чтобы значимым образом связать полученные результаты с ценностью для бизнеса.

Пытаясь добиться признания в качестве уважаемого технического эксперта, он излишне увлекается технической терминологией, и хотя заинтересованные стороны считают результаты его работы впечатляющими, они покидают зал без четкого представления о том, что делать дальше. По их мнению, проект нельзя считать завершенным из-за того, что он не был представлен понятно.

Проблема превращается в порочный круг: специалиста по работе с данными просят разработать лучшее решение для завершения проекта. Дата-сайентист углубляется в дебри…

Проверка реальности

Дата-сайентист проводит анализ рынка, но разработанное им решение не может быть реализовано в качестве рыночной стратегии, так как оно оторвано от того, как работает бизнес. Если бы в распоряжении компании было бесконечное количество качественных данных, средств и времени, это было бы отличным решением! Однако в действительности это отличное решение идеальной проблемы, а не той, которая на самом деле стоит перед бизнесом.

Но дата-сайентист непреклонен. Он хочет реализовать это решение «правильным» (то есть своим) способом. Он высокомерно заявляет своим

коллегам по бизнесу, что они должны придумать, как это сделать. Наконец вмешивается старший партнер и говорит, что проект будет свернут, если они не смогут проложить путь по текущей траектории.

«Что еще мы можем сделать?» — спрашивает старший партнер. (До сих пор никто не задавал этот вопрос.)

После этого команда находит способ, выигрышный для всех.

Захват власти

После многолетнего взаимодействия с клиентами в страховой отрасли команда проекта привлекает дата-сайентиста для анализа данных о клиентах, накопившихся за многие годы. Этот специалист недавно был повышен до старшего дата-сайентиста, и новое звание вскружило ему голову.

Команда усердно работала над выстраиванием доверительных отношений с каждым клиентом. Но старший дата-сайентист, которому не терпится решить проблемы и доказать свою ценность, настаивает на встрече с клиентом, утверждая, что в противном случае он ничего не сможет сделать. Вместо того чтобы рассматривать себя как часть команды, он считает себя консультантом, которому предстоит спасти компанию, введя ее в пространство данных.

Хотя нет ничего лучше, чем задать вопрос и услышать мнение клиента из первых уст, команда проекта чувствует неуважение со стороны дата-сайентиста. Во-первых, это сигнализирует об отсутствии веры в способности команды определять правильный контекст, выявлять реальную проблему и налаживать связи для оказания воздействия.

Во-вторых, это умаляет значимость сложной работы по выстраиванию доверительных отношений, побуждающих клиента говорить о своих потребностях. Это свидетельствует о (сознательном или неосознанном) пренебрежении риском, связанным с ненужной встречей и потенциальным оскорблением клиента.

Хвастун

Статистик одержим предоставлением максимально технических объяснений.

Хотя другие участники команды — столь же образованные и компетентные люди, хвастун тратит много времени на споры о том, какая методология является самой лучшей, а также на разбор примеров из Интернета

и учебников. Он открыто критикует бизнес-коллег как недостаточно умных, чтобы понять результаты проделанной работы (хотя они работают в компании намного дольше, чем он).

Все знают, насколько он умен, и он буквально упивается своим статусом софиста. Но он не производит результатов. Процесс создания презентации для него поистине мучителен. А за его тягой к спорам скрывается аналитический паралич. Делая каждый слайд, он будто идет на компромисс со своими убеждениями.

В результате к нему обращаются лишь тогда, когда кто-то должен сыграть роль адвоката дьявола в рамках проекта. Но даже в этом случае людям приходится выслушивать его наполненные жаргоном тирады и сравнения с предыдущими проектами, при реализации которых его мнения игнорировались.

В повседневной работе его мнения скорее мешают, чем помогают.

ОТНОШЕНИЕ К ДАННЫМ

В конечном итоге каждая из описанных сцен демонстрирует отсутствие эмпатии и уважения к вкладу каждого участника. Задумайтесь об этом на минутку.

Очень многие советы по подготовке бизнесов к работе с данными сводятся к инвестициям в технологии и обучение сотрудников. При этом множество неудач возникает на уровне коммуникации. Проблемы, лежащие в основе вышеописанных сцен, связаны не с технологиями или данными как таковыми, а скорее с конфликтующими людьми, не желающими слушать друг друга. Хорошая новость: все может быть иначе. Есть люди (мы надеемся, что читатели этой книги относятся именно к этой категории), которые хотят слушать и понимать других.

Далее мы поговорим о разных типах личности, с которыми вы можете столкнуться при работе с данными.

Энтузиасты

Первое соприкосновение с данными может показаться чем-то невероятно интересным и увлекательным. А возможности для бизнеса, связанные с использованием данных, вызывают у людей желание это попробовать. По своей сути это неплохо. Однако некоторые люди слишком сильно увлекаются.

Для них каждая новая вещь кажется чем-то потрясающим. По их мнению, данные позволяют решить любую проблему. А любое тематическое исследование, результат или диаграмма — доказательство проведения тщательного научного анализа. Энтузиасты часто просят показать им данные, но не задают дополнительных сложных вопросов, чтобы отделить шумиху от реальности.

Работая с энтузиастами, вам следует поощрять их любовь к данным, но одновременно напоминать им о том, что данные не могут сделать невозможное. Взращивая здоровый скептицизм, вы можете помочь им однажды стать главными по данным.

Циники

Для циников личный опыт важнее, чем наука о данных, статистика или машинное обучение. Такие люди часто насмехаются над вкладом дата-сайентистов. Они рассматривают данные в лучшем случае как раздражающую необходимость, предпочитая следовать своей интуиции. Когда им не нравятся результаты, циники выискивают недостатки, выходящие за рамки конструктивной критики.

Остановитесь на мгновение и подумайте о том, чем обусловлено такое отношение. Некоторый цинизм может быть вполне оправдан. Может быть, это объясняется тем, что они не росли в эпоху данных? Может быть, они наблюдали за провалом проектов по работе с данными? Не ожидайте того, что они будут ценить данные так же, как и вы. При работе с ними проявляйте эмпатию и попытайтесь понять, что ценят они. Обращайтесь к этим ценностям в ходе ваших коммуникаций и покажите им, что вы учитываете их экспертные знания в разрабатываемых решениях.

Со временем, когда циники научатся доверять данным, они могут стать главными по данным, но вы должны позволить им прийти к этому в их собственном темпе.

Скептики

В глубине души главные по данным являются скептиками. Они не пытаются вызвать этим раздражение, а просто используют свои навыки критического мышления. Как и энтузиасты, они выступают за использование данных там, где это уместно. Подобно циникам они ставят под вопрос то, что следует

подвергнуть сомнению. Их здоровый скептицизм основывается на технических знаниях и понимании предметной области и выражается с сочувствием.

Как главному по данным вам следует прислушиваться ко всей команде. В конце концов, каждый хочет быть услышанным и оцененным. Поэтому вам необходимо знать о препятствиях, с которыми сталкиваются ваши сотрудники.

ПОДВЕДЕНИЕ ИТОГОВ

В этой главе мы рассмотрели семь сцен, разыгрывающихся при взаимодействии разных участников проекта. Каждая сцена продемонстрировала различные пробелы в общении, выражающиеся в следующем:

— Специалисты по бизнесу не способны оценить результаты работы или вникнуть в проблемы специалистов по работе с данными. Эта тема раскрывается в сценах «Постмортем», «Время историй» и «Игра “Телефон”».
— Специалисты по работе с данными не могут оценить результаты работы или вникнуть в проблемы специалистов по бизнесу. Эта тема раскрывается в сценах «Проверка реальности» и «Захват власти».
— Специалисты по работе с данными отказываются отклониться от своей чисто технической роли из-за либо неосознанности (сцена «В дебри»), либо высокомерия (сцена «Хвастун»).

Мы также поговорили о том, что главный по данным должен взаимодействовать с разными личностями исходя из их особенностей, одновременно способствуя лучшему пониманию данных. Для этого необходимо проявлять эмпатию при общении. В следующей главе мы поговорим о том, что могут сделать главные по данным для создания в своих организациях среды, способствующей лучшему пониманию данных.

ГЛАВА 15

Что дальше?

«Из книг и примеров можно узнать лишь о том, что в принципе возможно. Настоящее обучение требует фактических действий»

_ *Фрэнк Герберт, американский писатель*

В этой краткой главе перечислены те вещи, которые помогут вам добиться успеха на пути становления главным по данным.

Главный по данным — это тот, кто:

— Думает статистически и понимает, какую роль вариации играют в жизни и процессе принятия решений.
— Разбирается в данных — разумно говорит и задает правильные вопросы о статистике и результатах, с которыми сталкивается на рабочем месте.
— Осознает истинное положение вещей в сфере машинного обучения, текстовой аналитики, глубокого обучения и искусственного интеллекта.
— Избегает распространенных ловушек при работе с данными и их интерпретации.

Другими словами, главный по данным — это кто-то вроде вас.

Чтобы добиться успеха на этом поприще, вы должны стать человеком, использующим данные для управления изменениями в вашей организации. Мы надеемся, что эта книга предоставила вам достаточно сценариев для обдумывания. Однако помните, что мы испытали на своем опыте далеко

не все, поэтому в эту книгу были включены некоторые намеренно тривиальные примеры. Реальный мир намного сложнее. Идея использования данных для изменения мира хорошо выглядит в книге, но на практике сделать это совсем не просто.

Если эта книга вас вдохновила, мы очень рады. Однако настоящая работа только начинается. Эти важные идеи не смогут распространиться без вашей помощи.

Вот несколько вещей, которые вы можете сделать:

— Создайте рабочую группу главных по данным в своей компании.
— Организуйте регулярные встречи для углубленного рассмотрения тем, которые мы обсудили, а также тех тем, которые мы не затрагивали.
— Возьмите на себя обязательство делиться своими знаниями и помогать другим.

Когда-то изучение новых концепций в области данных происходило в рамках конференций, саммитов и мастерских. В деле обучения компании и сотрудники полагались на мероприятия. И все же, заканчивая работу над этой книгой в январе 2021 года, спустя 10 месяцев после начала глобальной пандемии, мы не можем не думать о том, что эта модель уже не такая жизнеспособная, какой когда-то была. Более того, компании могут на бумаге заявлять о том, что инвестируют в обучение своих сотрудников, но в реальности бюджеты на обучение постоянно сокращаются.

Из этого следует, что компании сняли с себя бремя обучения. Раньше они считали новые идеи, связанные с данными, чем-то внешним, что нужно было донести до сотрудников. Теперь они ожидают, что новые сотрудники уже обладают некоторым опытом работы с данными. А умение быстро обучаться в настоящее время само по себе является необходимым навыком.

Быть главным по данным значит быть готовым к этой новой реальности, то есть к тому, что большая часть вашего обучения будет происходить вне работы (а не на работе). Вы будете учиться с помощью книг вроде этой, онлайн-курсов и программ сертификации. Наш мир сделал выбор в пользу более дешевого обучения, а это означает, что за свое образование ответственность несете вы. Независимо от того, где именно вы находитесь в иерархии компании, вы не можете ограничивать свое личное развитие событиями, которые происходят два раза в год. Не стоит рассчитывать, что вас вдохновит какой-нибудь доклад. Данные не будут ждать, пока вы захотите подумать

о них критически. Вы должны продолжать учиться и нести ответственность за траекторию своего развития.

Теперь у вас есть правильные инструменты и образ мышления для того, чтобы стать главным по данным. Люди, умеющие понимать, думать и говорить на языке данных, способны пробиться сквозь шум, ажиотаж и заблуждения. Вам не обязательно быть титаном отраслевых технологий, чтобы использовать машинное обучение и искусственный интеллект. Хотя многие концепции, представленные в этой книге, отражают новые технологии, проблемы, которые они представляют для бизнеса, существуют уже на протяжении нескольких десятилетий и сводятся к низкому качеству данных, ошибочным предположения и нереалистичным ожиданиям.

В то же время шумиха и завышенные ожидания от данных часто отвлекают внимание от этих основополагающих проблем. В начале книги мы обсудили несколько катастроф, которые произошли из-за того, что сотрудники организации не думали как главные по данным. По мере роста объема данных риск таких ошибок возрастает.

В лучшем случае такие ошибки бывает легко исправить. В худшем случае они приводят к пустой трате денег, подвергают опасности жизни людей и укрепляют заложенные в данных стереотипы. Как главный по данным вы должны задавать правильные вопросы, спорить с данными и вести неудобные разговоры. Фундамент, который вы заложили прочтением этой книги, поможет вам справиться с этой задачей.

Об авторах

Алекс Дж. Гутман — дата-сайентист, корпоративный тренер, получатель гранта Фулбрайта и аккредитованный профессиональный статистик, который с удовольствием преподает широкий спектр тем, связанных с наукой о данных, слушателям с разным уровнем технической подготовки. Он получил степень доктора философии по прикладной математике в Технологическом институте ВВС США, где в настоящее время работает адъюнкт-профессором.

Джордан Голдмайер — всемирно признанный профессионал в области аналитики и эксперт по визуализации данных, автор и спикер. В прошлом он был операционным директором Excel.TV и много лет обучал людей работе с данными. Он написал книги «Advanced Excel Essentials» и «Dashboards for Excel». Его работы цитировались в Associated Press, Bloomberg BusinessWeek и American Express OPEN Forum. В настоящее время он — семикратный обладатель награды Excel MVP Award, что дает ему право предоставлять обратную связь и давать рекомендации группам разработчиков продуктов Microsoft. Однажды с помощью программы Excel он сэкономил для ВВС США 60 миллионов долларов. Также Джордан работает техником скорой медицинской помощи в качестве волонтера.

О технических редакторах

Уильям А. Бреннеман — научный сотрудник и ведущий специалист по глобальной статистике в отделе моделирования и науки о данных компании Procter & Gamble, а также адъюнкт-профессор практики в Школе промышленной и системной инженерии им. Стюарта при Технологическом институте Джорджии. После прихода в P&G он работал над широким спектром проектов, связанных с применением статистических методов в таких областях своей компетенции, как проектирование и анализ экспериментов, робастное параметрическое проектирование, проектирование надежности, статистическое управление процессами, компьютерные эксперименты, машинное обучение и статистическое мышление. Он также сыграл важную роль в разработке корпоративной учебной программы по статистике. Он получил степень доктора философии в области статистики в Мичиганском университете, степень магистра математики в Университете Айовы и степень бакалавра математики с правом преподавания в старших классах средней школы в Таборском колледже. Уильям является членом Американской статистической ассоциации (ASA) и Американского общества качества (ASQ). Он был председателем отдела статистики ASQ, председателем секции качества и производительности ASA, а также заместителем редактора журнала *Technometrics*. Уильям также семь лет работал преподавателем в средней школе и колледже.

Дженнифер Стиррап — основатель и генеральный директор Data Relish, ведущей британской консалтинговой компании, работающей в области искусственного интеллекта и бизнес-аналитики и разрабатывающей стратегии работы с данными и бизнес-ориентированные решения. Джен — признанный авторитет в области ИИ и бизнес-аналитики, всемирно известный спикер из списка Fortune 100. Она входит в число 50 лучших мировых провидцев в области науки о данных, в число лучших дата-сайентистов, на которых следует подписаться в Twitter, а также в число 50 самых влиятельных женщин мира в сфере технологий.

Джен консультирует клиентов из 24 стран, расположенных на пяти континентах, и имеет ученые степени в области искусственного интеллекта

и когнитивных наук. Джен пишет книги, посвященные науке о данных и искусственному интеллекту. Она принимала участие в передачах, выходивших на CBS Interactive и BBC, а также в создании таких известных подкастов, как Digital Disrupted, Run As Radio и собственной серии вебинаров Make Your Data Work.

Кроме того, Джен выступала с программными докладами в колледжах и университетах, а также делилась своим опытом с благотворительными и некоммерческими организациями, выступая в качестве неисполнительного директора. Все доклады Джен основаны на ее более чем 20-летнем опыте самоотверженной и усердной работы по всему миру.

Благодарности

Я заметил, что при написании раздела «Благодарности» авторы книг обычно упоминают своих супругов в самом конце. Возможно, это объясняется их желанием оставить лучшее напоследок. Однако я пообещал своей жене, что если когда-нибудь напишу книгу, то первым делом упомяну ее, чтобы ясно показать, чей вклад для меня является наиболее важным. Итак, я благодарю свою жену Эрин за ее любовь, поддержку и улыбку. Прямо сейчас она берет наших троих маленьких детей на велосипедную прогулку, предоставляя мне возможность дописать последнюю страницу. (Я уверяю читателей, что этот поступок в полной мере отражает тот образ жизни, которого мы придерживались на протяжении прошедшего года.)

Я хотел бы поблагодарить своих родителей, Эда и Нэнси, за то, что они поддерживали меня во всех начинаниях и показали мне пример хорошего родительства, а также моих братьев Райана и Росса и сестру Эрин за их поддержку.

Эта книга — кульминация множества дискуссий, проведенных с друзьями и коллегами, с которыми мы обсуждали всевозможные вопросы, начиная с целесообразности написания книги об овладении языком науки о данных и заканчивая выбором тем, которые стоит в нее включить. Я выражаю особую благодарность Алтынбеку Исмаилову, Энди Ноймайеру, Брэдли Бёмке, Брэндону Гринвеллу, Бренту Расселу, Кейду Сайе, Калебу Гудро, Карлу Парсону, Дэниэлу Уппенкампу, Дугласу Кларку, Грегу Андерсону, Джейсону Фрилсу, Джоэлу Чейни, Джозефу Келлеру, Джастину Мауреру, Нэйтану Свигарту, Филу Хартке, Сэмюэлу Риду, Шону Шнайдеру, Стивену Ферро и Закари Аллену.

Я также в долгу перед сотнями инженеров, бизнес-профессионалов и специалистов в области науки о данных, с которыми я общался лично или через Интернет, и которые помогли мне стать более эффективным дата-сайентистом и коммуникатором. Я также хочу сказать спасибо своим «студентам»

(коллегам), которые предоставили честные отзывы о курсах, которые я преподавал. Я услышал вас и благодарен вам.

Мне посчастливилось иметь множество академических и профессиональных наставников, которые помогли мне обрести собственный голос и уверенность в качестве статистика, дата-сайентиста и тренера. Я выражаю благодарность Джеффри Вейру, Джону Тудоровичу, К. Т. Арасу, Рэймонду Хиллу, Робу Бейкеру, Скотту Кроуфорду, Стивену Чэмбалу, Тони Уайту и Уильяму Бреннеману (который любезно согласился стать техническим редактором этой книги). Общаясь с такими людьми, просто невозможно не стать мудрее.

Я также хочу сказать спасибо команде издательства Wiley: Джиму Минателу за веру в проект и предоставленный нам шанс, Питу Гогану и Джону Слива, которые направляли нас на протяжении всего процесса написания книги, а также производственному персоналу Wiley за тщательную вычитку глав. Также выражаю благодарность нашим техническим редакторам Уильяму Бреннеману и Джен Стиррап за ценные предложения и опыт, благодаря которым книга стала гораздо лучше.

Отдельно хочу поблагодарить своего соавтора Джордана Голдмайера и не только за книгу, которую вы держите в руках. В начале своей карьеры я пожаловался Джордану на то, что люди не разделяют моего интереса к статистике и статистическому образу мышления. На это он сказал, что раз меня это так беспокоит, то я должен это изменить. С тех пор я выполняю это обязательство.

Наконец, я хотел бы снова сказать спасибо своей жене Эрин (потому что лучшее действительно следует оставлять напоследок).

— Алекс

Я хотел бы поблагодарить всех тех людей, благодаря которым эта книга вышла в свет.

Прежде всего я выражаю благодарность моему соавтору Алексу Гутману. В течение многих лет мы обсуждали идею совместного написания книги. Когда подходящий момент настал, мы это сделали. О лучшем соавторе я не мог бы и мечтать.

Спасибо замечательным сотрудникам Wiley, в том числе рецензенту издательства Джиму Минателу и руководителю проекта Джону Слива. Кроме

того, я хотел бы выразить признательность нашим техническим редакторам, Уильяму Бреннеману и Джен Стиррап, за их усердную работу по рецензированию книги. Мы учли все ваши комментарии.

И последнее, но не менее важное: я хочу сказать спасибо моему партнеру Кэти Грей, которая всегда верила в этот проект — и в меня.

— Джордан

Предметный указатель

Amazon, компания 234
  Alexa, система 215
Apple, компания 234
  Siri, система 215
CART, алгоритм 200
Datasaurus, набор данных 106
FiveThirtyEight, блог 22
Google, компания 222, 234
  Assistant, система 215
  Gmail, сервис 254
  News, сервис 222
  Translate, система 215
GPT-3, модель 234
IBM, компания
  Watson, компьютер 215
Keras, библиотека 257
Microsoft, компания 234
N-грамма 221
OpenAI, компания 234
P-значение 136
  механизм вычисления 139
P-хакинг 273
R-квадрат 178
R, язык программирования 148
Smart Compose, функция 254
Word2vec, модель 224
XGBoost, алгоритм 209
Zestimate, инструмент 97
Zillow, компания 97

А/Б-тестирование 56
Алгоритм 28, 172
  обучение 29
Анализ главных компонент (АГК) 157
  ловушки 163
Анализ настроений 46, 214, 232
Ансамблевые методы 203
Атрибут 52
Байеса, теорема 125
Беринато, Скотт 279
Биграмма 221
Биномиальная регрессия 84
Булева логика 114
Бэггинг 204
Вальд, Абрахам 267
Вариация 63, 110
  измерений 64
  недооценка 72
  случайная 64
Вектор 222
Вероятность 110, 112
  и интуиция 71
  калибровка 128
  ловушки 120
  наступления множества событий 115
  независимость событий 121
  нотация 112
  суммарная 114

условная 114, 123
Вес 241
Визуализация данных 98
Выборка 132
небольшая 74
нерепрезентативная 273
предвзятая 74
размер 74, 133
Выборы в США 2016 года 22
Выброс 90, 102
Гальтон, Фрэнсис 267
Гарбер, Алан 63
Генеративно-состязательные сети 260
Генерация понятного человеку текста 234
Гипотеза
альтернативная 135, 143
нулевая 134
проверка 134
Гистограмма 98
Главные компоненты 158
Главный по данным 39, 287
вопросы 40
Глубокое обучение 237
наличие данных 254
обработка языка и последовательностей 252
преимущества 247
применение 245
структурированные данные 256
этический аспект 259
Градиент 206
Графические процессоры 247
Данные 19, 52
большие 15, 20
визуализация 98
возможности 92
входные 97, 171
выходные 97, 171
главный по 287
избыточность 155
категориальные 53
ловушки при работе с 265
масштаб 168
наблюдений 55, 88
набор 52
на рабочем месте 24
наука о 20
необработанные 79
непрерывные 53
нерепрезентативные обучающие 274
несбалансированные 210
неструктурированные 56, 214
неупорядоченные (или номинальные) 54
обучающие 172
отношение людей к 279
отсутствие разделения 208
отсутствующие 91
оценка качества 79
перцептивные 256
предвзятость 266
происхождение 87
разведочный анализ 94
размерность 155
репрезентативность 89
сбор 87, 88
спор с 79, 93
странности в 266
структурированные 56, 214, 256
счетные (или дискретные) 54
текстовые 215

типы 53
упорядоченные (или порядковые) 54
утечка 184, 207, 274
числовые 53
экспериментальные 55, 88
Дата-сайентист 15
Дерево решений 199
обрезка 203
переобучение 202
с градиентным усилением 204
случайный лес 203
Джеймс, Леброн 58, 135
Диаграмма
Венна 116
древовидная 127
линейная 101
размаха 99
рассеяния 101
Дипфейк 260
Дисперсия 58, 162
Доверительный интервал 132, 146
Доказательство обратного 135
Документ 216
Допущение эквивалентности 144
Дэвенпорт, Томас Х. 18
Европейская организация по ядерным исследованиям (ЦЕРН) 144
Зависимости
перестановка 123
Заключение
ложноотрицательное 137
ложноположительное 137
Закон
малых чисел 72
средних чисел 122
Запись 52
Иллюзия квантификации 67
Индуктивная статистика 73, 89
Интерпретируемость 183
ансамблевых моделей 206
Информация 51
кодирование 52
Исключение голосов 142
Искусственный интеллект
общий 257
узкого назначения (слабый) 257
Испытание 52
Исследовательский образ мышления 96
Каиро, Альберто 106
Квартет Энскомба, набор данных 105
Класс большинства 209
Классификация 29
бинарная 191
изображений 249
многоклассовая 192
постановка задачи 193
Кластеризация 164
иерархическая 169
методом k-средних 165
Клиентское восприятие 45, 64
Клинтон, Хиллари 22
Количество проводимых тестов 145
Коммуникативный сбой 279
Конструирование признаков 247
Корреляция 104, 162
выявление неправильной 268
неверная интерпретация 105
не означает причинность 107, 268, 272
отрицательная 104

положительная 104
Кортеж 52
Коэффициент 241
корреляции Пирсона 104
Кризис ипотечного кредитования 20
Латентное размещение Дирихле (ЛРД) 226
Латентно-семантический анализ (ЛСА) 226
ЛеКун, Янн 208, 246, 249
Лемматизация 219
Линейная комбинация 157
Линейная регрессия 173
варианты 189
множественная 180
подводные камни 181
производительность модели 187
простая 180
что дает 179
что делает 174
Ловушка
машинного обучения 272
невозвратных затрат 270
статистики 272
Логистическая регрессия 194, 229
преимущества 197
проблемы 198
решающее правило 197
точка отсечения 197
Логистическая функция 196
потерь 197
Ложноположительный результат 127
Матрица \«документ — термин\» 219
Матрица ошибок 210
путаница в терминах 212
Машинное обучение
ловушки 272
Медиана 58
Ментальная модель 33
Мера
вариации 58
линейной зависимости 105
разброса 58
центральной тенденции 58
Метка 29, 172, 191
Метод
k-ближайших соседей 29, 189
ансамблевый 203
выявления причинности 88
наименьших квадратов 173, 175
Мешок слов 216
недостаток 221
Мода 58
Модель 97
классификационная 173, 191
переобучение 186
прогностическая 172
регрессионная 173
черный ящик 206, 245
Мощность 137
Мультиколлинеарность 183, 199
Наблюдение 52, 88, 164
Набор данных 52
обучающий 187
тестовый 187
Наилучшее соответствие 175
НАСА 82
Нг, Эндрю 254
Независимые события 115
Нейрон 240
Нейронная сеть 238

веса 241
настройка 256
обратное распространение ошибки 242
память 253
принцип работы 239
процесс обучения 241
рекуррентная 252
сверточная 250
с долгой кратковременной памятью 253
функция активации 240
Нелинейные взаимосвязи 186
Неопределенность 110
Неправильное
определение типа задачи 207
понимание точности 209
пороговое значение 208
Нерепрезентативная выборка 273
Обама, Барак 22
Облако слов 218
Обработка естественного языка 233
Обратное распространение ошибки 242
Обучение 288
глубокое 237, 257
контролируемое 29, 171
машинное 29, 257
с учителем 171
трансферное 255
Общий искусственный интеллект (ОИИ) 257
Описательная статистика 73
Опрос 132
погрешность результатов 132
Основополагающие допущения 74
Отзывчивость 211
Отсутствующие значения 91
Ошибка 265
второго рода 137
выжившего 266
первого рода 137
при принятии решения 137
экстраполяции 185
Переменная 52
категориальная 92, 191
отклика 29
пропущенная 182, 199
смешивающаяся 108
Переобучение 186, 274
Перепись 89, 142
Погрешность 176
Поле 52
Поправка на множественную проверку гипотез 146
Популяция 132
Правило умножения 116
Практическое значение результатов теста 147
Предвзятость 74, 266
выборки 90
подтверждения 270
Предиктор 52
Представление 243
Преобразование
речи в текст 234
текста в речь 234
текста в текст 234
Признак 52, 155
вес 159
включение множества 180
конструирование 247
смешение 56
составной 155

Причинно-следственная связь 148
  принятие корреляции за 268, 272
Проблема
  важность 41
  определение 39
  отсутствие нужных данных 43
Проект по работе с данными
  длительность 44
  неудовлетворительные результаты 44
  причины провала 45
Проклятье Madden 268
Разведочный анализ данных 94
Размах 58
Размерность 155
  снижение 155
Размер эффекта 147
Регрессия
  биномиальная 84
  к среднему 267
  линейная 173
  логистическая 194
  методом наименьших квадратов 175
  множественная 181
Рейган, Рональд 83
Рекуррентная нейронная сеть 252
Свертка 251
Сверточная нейронная сеть 250
Сезонность 101
Сильвер, Нейт 22
Симпсона, парадокс 106, 268
Систематическая ошибка выжившего 266
Скрытые группы 154
Словарь 219
Случайный лес 203
Спам-фильтр 228
Специфичность 211
Среднее значение 57
Стандартное отклонение 58
Статистика
  индуктивная 73, 89, 131
  контекст 140
  ловушки 272
  описательная 73
  сводная 57
Статистический вывод 131
Статистический тест 139
Статистическое мышление
  основной принцип 61
Стемминг 219
Столбиковый график 100
Стоп-слово 219
Структурированные данные 214
Таблица
  разреженная 219
Текст
  майнинг 215
  практические соображения при работе с 233
  преобразование в числа 216
  технологии анализа 215
Текстовая аналитика
  ожидания от 214
Тематическое моделирование 225
Технологические гиганты
  преимущества 258
Токен 219
Точечная оценка 132
Точка данных 53
Точность 209
  неправильное понимание 209
Трамп, Дональд Дж. 22

Трансферное обучение 255
Тьюки, Джон 94
Уровень
  доверия 146
  значимости 136, 144, 146
Утечка данных 184, 207, 274
Фишер, Рональд Э. 108
Фокус
  на конечном результате 41
  на методологии 41
Функция активации 240
Хиггса, бозон 144
Цветовые каналы 250
Центроид 167
Цепное правило 242
Чат-бот 234
Челленджер, шаттл
  катастрофа 82
Человеческий мозг 238
Чувствительность 211
Шолле, Франсуа 257
Эйлера, постоянная 196
Эксперимент 88
Экстраполяция 185, 199
Ящик с усами 99

Научно-популярное издание

МИРОВОЙ КОМПЬЮТЕРНЫЙ БЕСТСЕЛЛЕР

Гатман Алекс Дж., Голдмейер Джордан

РАЗБЕРИСЬ В DATA SCIENCE

Как освоить науку о данных и научиться думать как эксперт

Главный редактор *Р. Фасхутдинов*
Руководитель направления *В. Обручев*
Ответственный редактор *Д. Калачева*
Младший редактор *Д. Данилова*
Художественный редактор *А. Шуклин*
Компьютерная верстка *Э. Брегис*
Корректоры *А. Баскакова*, *Л. Макарова*

В оформлении обложки использована иллюстрация:
INGARA / Shutterstock.com
Используется по лицензии от Shutterstock.com

Страна происхождения: Российская Федерация
Шығарылған елі: Ресей Федерациясы

ООО «Издательство «Эксмо»
123308, Россия, город Москва, улица Зорге, дом 1, строение 1, этаж 20, каб. 2013.
Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru
Өндіруші: «ЭКСМО» АҚБ Баспасы,
123308, Ресей, қала Мәскеу, Зорге көшесі, 1 үй, 1 гимарат, 20 қабат, офис 2013 ж.
Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru.
Тауар белгісі: «Эксмо»
Интернет-магазин : www.book24.ru

Интернет-магазин : www.book24.kz
Интернет-дүкен : www.book24.kz
Импортёр в Республику Казахстан ТОО «РДЦ-Алматы».
Қазақстан Республикасындағы импорттаушы «РДЦ-Алматы» ЖШС.
Дистрибьютор и представитель по приему претензий на продукцию,
в Республике Казахстан: ТОО «РДЦ-Алматы»
Қазақстан Республикасында дистрибьютор және өнім бойынша арыз-талаптарды
қабылдаушының өкілі «РДЦ-Алматы» ЖШС,
Алматы қ., Домбровский көш., 3«а», литер Б, офис 1.
Тел.: 8 (727) 251-59-90/91/92; E-mail: RDC-Almaty@eksmo.kz
Өнімнің жарамдылық мерзімі шектелмеген.
Сертификация туралы ақпарат сайтта: www.eksmo.ru/certification

Сведения о подтверждении соответствия издания согласно законодательству РФ
о техническом регулировании можно получить на сайте Издательства «Эксмо»
www.eksmo.ru/certification
Өндірген мемлекет: Ресей. Сертификация қарастырылмаған

Дата изготовления / Подписано в печать 03.02.2023. Формат 70x100 $^{1}/_{16}$.
Печать офсетная. Усл. печ. л. 24,63.
Тираж экз. Заказ

ISBN 978-5-04-174810-4

12+

ИСЧЕРПЫВАЮЩЕЕ **РУКОВОДСТВО** ПО ОСНОВАМ **DATA SCIENCE**

Что мешает раскрытию истинного потенциала науки о данных? Очевидно, проблема не в медленных алгоритмах, не в недостатке данных и уж точно не в нехватке вычислительной мощности или дата-сайентистов. Дело в распространенном заблуждении, что Data Science – это сложно и заниматься наукой о данных могут только опытные программисты. На самом деле это не так.

ЭТА КНИГА РАЗВЕЕТ ВСЕ МИФЫ И НАУЧИТ ВАС:

- **МЫСЛИТЬ СТАТИСТИЧЕСКИ** И ПОНИМАТЬ, КАКУЮ РОЛЬ В ВАШЕЙ РАБОТЕ ИГРАЕТ АНАЛИТИКА.
- **ПОЛЬЗОВАТЬСЯ ЯЗЫКОМ НАУКИ О ДАННЫХ,** ТО ЕСТЬ ОСМЫСЛЕННО ГОВОРИТЬ И ЗАДАВАТЬ ПРАВИЛЬНЫЕ ВОПРОСЫ ОТНОСИТЕЛЬНО СТАТИСТИКИ.
- **ПОНИМАТЬ РЕАЛЬНОЕ ПОЛОЖЕНИЕ ВЕЩЕЙ** В ТАКИХ ОБЛАСТЯХ, КАК МАШИННОЕ ОБУЧЕНИЕ, ТЕКСТОВАЯ АНАЛИТИКА, ГЛУБОКОЕ ОБУЧЕНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ.
- **ИЗБЕГАТЬ РАСПРОСТРАНЕННЫХ ОШИБОК** ПРИ РАБОТЕ С ДАННЫМИ И ИХ ИНТЕРПРЕТАЦИЕЙ.

Руководство будет полезным каждому желающему научиться ориентироваться в грядущем будущем, неразрывно связанном с Data Science.

ISBN 978-5-04-174810-4
9 785041 748104 >

БОМБОРА
ИЗДАТЕЛЬСТВО