

*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
<i>Кризис субстандартного ипотечного кредитования</i>	20
<i>Всеобщие выборы в США 2016 года</i>	22
<i>Наша гипотеза</i>	23
Данные на рабочем месте	24
<i>Сцена в зале заседаний</i>	24
Вы можете понять общую картину	26
<i>Классификация ресторанов</i>	26
<i>Что дальше?</i>	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
<i>Почему эта проблема важна?</i>	41
<i>Кого затрагивает эта проблема?</i>	42
<i>Что, если у нас нет нужных данных?</i>	43
<i>Когда проект будет завершен?</i>	44
<i>Что, если нам не понравятся результаты?</i>	44
Причины провала проектов по работе с данными	45
<i>Клиентское восприятие</i>	45
<i>Обсуждение</i>	47
Работа над значимыми проблемами	48
Подведение итогов	49

ГЛАВА 2		
Что такое данные?		51
Данные и информация		51
<i>Пример набора данных</i>		52
Типы данных		53
Сбор и структурирование данных		55
<i>Данные наблюдений и экспериментальные данные</i>		55
<i>Структурированные и неструктурированные данные</i>		56
Основы сводной статистики		57
Подведение итогов		58
ГЛАВА 3		
Готовьтесь мыслить статистически		60
Задавайте вопросы		61
Во всем есть вариации		63
<i>Сценарий: Клиентское восприятие (продолжение)</i>		64
<i>Анализ реальной ситуации: показатели заболеваемости раком почки</i>		67
Вероятности и статистика		69
<i>Вероятность и интуиция</i>		71
<i>Открытия с помощью статистики</i>		73
Подведение итогов		75
ЧАСТЬ II		
Говорите как главный по данным		
ГЛАВА 4		
Сомневайтесь в данных		79
Что бы вы сделали?		80
<i>Катастрофа, вызванная недостатком данных</i>		82
Расскажите мне историю происхождения данных		87
<i>Кто собирал данные?</i>		87
<i>Как собирались эти данные?</i>		88
Являются ли данные репрезентативными?		89
<i>Имеет ли место предвзятость выборки?</i>		90
<i>Что вы сделали с выбросами?</i>		90
Какие данные я не вижу?		91
<i>Как вы поступили с отсутствующими значениями?</i>		91
<i>Позволяют ли данные измерить то, что вас интересует?</i>		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
<i>Направляющие вопросы</i>	96
<i>Сценарий</i>	97
Позволяют ли данные ответить на поставленный вопрос?	97
<i>Определитесь с ожиданиями и руководствуйтесь здравым смыслом</i>	97
<i>Имеют ли данные интуитивный смысл?</i>	98
<i>Осторожно: выбросы и отсутствующие значения</i>	102
Обнаружили ли вы какие-либо взаимосвязи?	103
<i>Корреляция</i>	104
<i>Осторожно: неверная интерпретация корреляции</i>	105
<i>Осторожно: корреляция не означает причинность</i>	107
Обнаружили ли вы новые возможности в данных?	108
Подведение итогов	109
ГЛАВА 6	
Изучайте вероятности	110
Попробуйте угадать	111
Правила игры	112
<i>Нотация</i>	112
<i>Условная вероятность и независимые события</i>	114
<i>Вероятность наступления множества событий</i>	115
<i>Одновременное наступление двух событий</i>	115
<i>Наступление одного или другого события</i>	117
Мысленное упражнение на определение вероятности	119
<i>Дальнейшие шаги</i>	120
Будьте осторожны, делая предположения о независимости событий	121
<i>Не допускайте ошибку игрока</i>	122
Все вероятности являются условными	123
<i>Не меняйте зависимости местами</i>	123
<i>Теорема Байеса</i>	125
Убедитесь, что вероятности имеют смысл	128
<i>Калибровка</i>	128
<i>Редкие события могут случаться и случаются</i>	129
Подведение итогов	130

ГЛАВА 7

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

ЧАСТЬ III**Освойте набор инструментов дата-сайентиста****ГЛАВА 8**

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

ГЛАВА 9	
Освойте модели регрессии	171
Обучение с учителем	171
Линейная регрессия: что она делает	174
<i>Регрессия методом наименьших квадратов: больше, чем умное название</i>	175
Линейная регрессия: что она дает	179
<i>Включение множества признаков</i>	180
Линейная регрессия: какую путаницу она вызывает	181
<i>Пропущенные переменные</i>	182
<i>Мультиколлинеарность</i>	183
<i>Утечка данных</i>	184
<i>Ошибки экстраполяции</i>	185
<i>Многие взаимосвязи не являются линейными</i>	186
<i>Вы объясняете или предсказываете?</i>	186
<i>Производительность регрессионной модели</i>	187
Прочие модели регрессии	189
Подведение итогов	189
ГЛАВА 10	
Освойте модели классификации	191
Введение в классификацию	191
<i>Чему вы научитесь</i>	192
<i>Постановка задачи классификации</i>	193
Логистическая регрессия	194
<i>Логистическая регрессия: что дальше?</i>	197
Деревья решений	199
Ансамблевые методы	203
<i>Случайные леса</i>	203
<i>Деревья решений с градиентным усилением</i>	204
<i>Интерпретируемость ансамблевых моделей</i>	206
Остерегайтесь ловушек	206
<i>Неправильное определение типа задачи</i>	207
<i>Утечка данных</i>	207
<i>Отсутствие разделения данных</i>	208
<i>Выбор неправильного порогового значения для принятия решения</i>	208
<i>Неправильное понимание точности</i>	209
<i>Матрицы ошибок</i>	210
Подведение итогов	212

ГЛАВА 11	
Освойте текстовую аналитику	214
Ожидания от текстовой аналитики	214
Как текст превращается в числа	216
<i>Большой мешок слов</i>	216
<i>N-граммы</i>	221
<i>Векторное представление слов</i>	222
Тематическое моделирование	225
Классификация текстов	227
<i>Наивный байесовский алгоритм</i>	229
<i>Анализ настроений</i>	232
Практические соображения при работе с текстом	233
<i>Преимущества технологических гигантов</i>	234
Подведение итогов	235
ГЛАВА 12	
Концептуализируйте глубокое обучение	237
Нейронные сети	238
<i>Чем нейронные сети похожи на мозг?</i>	238
<i>Простая нейронная сеть</i>	239
<i>Как учится нейронная сеть</i>	241
<i>Чуть более сложная нейронная сеть</i>	242
Применение глубокого обучения	245
<i>Преимущества глубокого обучения</i>	247
<i>Как компьютеры «видят» изображения</i>	249
<i>Сверточные нейронные сети</i>	250
<i>Глубокое обучение для обработки языка</i> <i>и последовательностей</i>	252
Глубокое обучение на практике	254
<i>Есть ли у вас данные?</i>	254
<i>Являются ли ваши данные структурированными?</i>	256
<i>Как будет выглядеть сеть?</i>	256
Искусственный интеллект и вы	257
<i>Преимущества технологических гигантов</i>	258
<i>Этический аспект глубокого обучения</i>	259
Подведение итогов	261

ЧАСТЬ IV**Гарантируйте успех****ГЛАВА 13****Остерегайтесь ловушек 265**

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

ГЛАВА 14**Найдите людей и типы личностей 278**

Семь сцен коммуникативного сбоя	279
<i>Постмортем</i>	280
<i>Время историй</i>	280
<i>Игра «Телефон»</i>	281
<i>В дебри</i>	282
<i>Проверка реальности</i>	282
<i>Захват власти</i>	283
<i>Хвастун</i>	283
Отношение к данным	284
<i>Энтузиасты</i>	284
<i>Циники</i>	285
<i>Скептики</i>	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 будущего¹. 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 Сауда 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 регулярностью².

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