

O'REILLY®



Data Science

ИНСАЙДЕРСКАЯ ИНФОРМАЦИЯ ДЛЯ НОВИЧКОВ.
ВКЛЮЧАЯ ЯЗЫК R

Кэти О'Нил, Рэйчел Шатт

 ПИТЕР®

Кэти О'Нил, Рэйчел Шатт

Data Science

ИНСАЙДЕРСКАЯ ИНФОРМАЦИЯ ДЛЯ НОВИЧКОВ.
ВКЛЮЧАЯ ЯЗЫК R



Санкт-Петербург · Москва · Екатеринбург · Воронеж
Нижний Новгород · Ростов-на-Дону · Самара · Минск

2019

ББК 32.972.2-018
УДК 004.3
О-58

О'Нил Кэти, Шатт Рэйчел

О-58 Data Science. Инсайдерская информация для новичков. Включая язык R. — СПб.: Питер, 2019. — 368 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-0622-6

Data Science (исследование данных) — одна из самых востребованных специализаций нашего времени. Изучение данных позволяет преобразить любую традиционную или инновационную бизнес-модель. Эта книга основана на вводном курсе по Data Science из Колумбийского университета, и начинающему специалисту-аналитику она совершенно необходима.

Эта книга увлекательно и доступно рассказывает о:

- байесовском методе;
- статистических алгоритмах;
- финансовом моделировании;
- рекомендательных движках;
- визуализации данных;
- MapReduce.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.972.2-018
УДК 004.3

Права на издание получены по соглашению с O'Reilly.

Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги.

Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-1449358655 англ.

Authorized Russian translation of the English edition of Doing Data Science
ISBN9781449358655 © 2013 by Cathy O'Neil, Rachel Schutt

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

ISBN 978-5-4461-0622-6

© Перевод на русский язык ООО Издательство «Питер», 2019

© Издание на русском языке, оформление ООО Издательство «Питер», 2019

© Серия «Библиотека программиста», 2019

Оглавление

Предисловие Рэйчел Шатт.....	14
Мотивация	14
Происхождение курса.....	15
Как появилась эта книга	17
Чего следует ждать от книги	17
Структура издания.....	18
Как читать книгу.....	18
Как в книге используется код	19
Для кого это издание.....	19
Что вы уже должны знать.....	20
Дополнительная литература	20
О тех, кто внес вклад в книгу	22
Условные обозначения.....	22
Использование примеров кода	23
Благодарности.....	24
Глава 1. Введение: что такое наука о данных	26
Большие данные и наука о данных	26
За пологом шумихи	27
Почему именно сейчас.....	29
Сегодняшняя картина (и немного истории).....	31
Профиль науки о данных.....	34

Мысленный эксперимент: метаопределение	36
Итак, кто же такой исследователь данных	37
В академических кругах	37
В промышленности	39
Глава 2. Статистический анализ, разведочный анализ данных и процесс их научного исследования	41
Статистическое мышление в век больших данных	41
Статистический анализ.....	42
Генеральные совокупности и выборки.....	43
Генеральные совокупности и выборки больших данных	44
Большие данные могут означать большие допущения.....	47
Моделирование.....	49
Разведочный анализ данных	57
Философия разведочного анализа данных.....	58
Упражнение: РАД.....	60
Процесс научных исследований данных.....	63
Мысленный эксперимент: как бы вы имитировали хаос?.....	66
Практический пример: RealDirect.....	68
Как RealDirect зарабатывает деньги	68
Упражнение. Стратегия по данным RealDirect.....	69
Глава 3. Алгоритмы	73
Алгоритмы машинного обучения	74
Три основных алгоритма	75
Линейная регрессия	77
k-ближайшие соседи	91
k-средние.....	101
Упражнение. Основные алгоритмы машинного обучения.....	105
Решения.....	105
Резюмируя вышесказанное.....	109
Мысленный эксперимент: автоматизированный статистик	110
Глава 4. Фильтры спама, наивный классификатор Байеса и перебор данных	111
Мысленный эксперимент: обучение на примере	111
Почему линейная регрессия не работает для фильтрации спама.....	113
Что насчет k-ближайших соседей	114
Наивный классификатор Байеса	116
Закон Байеса.....	116
Спам-фильтр для отдельных слов	117
Спам-фильтр, комбинирующий слова: наивный классификатор Байеса.....	119

Пофантазируем: сглаживание Лапласа	121
Сравнение наивного классификатора Байеса с k-БС	122
Пример кода в оболочке bash	123
Веб-агрегация: API и другие инструменты	124
Упражнение от Джейка: использование наивного классификатора Байеса для классификации статей	126
Пример кода на языке R для работы с NYT API	127
Глава 5. Логистическая регрессия	130
Мысленные эксперименты	131
Классификаторы	132
Время выполнения	133
Интерпретируемость	134
Масштабируемость	134
Тематическое исследование логистической регрессии M6D	134
Модели переходов	135
Математическая основа	136
Оценка α и β	138
Метод Ньютона	140
Стохастический градиентный спуск	140
Реализация	141
Оценка	141
Упражнение от компании Media 6 Degrees	144
Пример кода на R	144
Глава 6. Метки времени и финансовое моделирование	149
Кайл Тиг и GetGlue	149
Метки времени	151
Разведочный анализ данных (РАД)	152
Метрики и новые переменные или признаки	156
Что дальше?	156
Кэти О'Нил	157
Мысленный эксперимент	158
Финансовое моделирование	159
В пределах выборки, за пределами выборки, причинная зависимость	159
Подготовка финансовых данных	161
Логарифмическая доходность	163
Пример: индекс S&P	164
Разработка измерения волатильности	165
Экспоненциальное понижающее взвешивание	168
Финансовое моделирование петли обратной связи	169

Почему регрессия?	171
Добавление гипотез.....	171
Детская модель	172
Упражнение: GetGlue и данные о событиях с метками даты/времени	174
Упражнение: финансовые данные	176
Глава 7. Извлечение смысла из данных	177
Уильям Кукерски	177
Общая информация: соревнования по анализу данных	178
Общая информация: краудсорсинг	179
Модель Kaggle	181
Единственный участник	182
Их клиенты	182
Мысленный эксперимент: каковы этические последствия использования робота-оценщика?	185
Выбор признаков	187
Пример: привлечение пользователей	188
Фильтры.....	191
Обертки	192
Встроенные методы: деревья решений.....	194
Энтропия.....	195
Алгоритм дерева решений	198
Обработка непрерывных переменных в деревьях решений	198
Случайные леса	200
Удержание пользователей: интерпретируемость и прогнозирующая способность	202
Дэвид Хаффакер: гибридный подход к проведению социологических исследований Google.....	203
Переход от описаний к прогнозам	204
Социальность в Google	205
Конфиденциальность	206
Мысленный эксперимент: что является наилучшим способом снизить беспокойство и повысить понимание и контроль?	207
Глава 8. Рекомендательные механизмы: создание ориентированных на пользователя масштабируемых информационных продуктов	208
Реальный рекомендательный механизм	210
Обзор метода k-ближайших соседей.....	211
Некоторые проблемы, связанные с методом k-БС	211
За рамками метода k-БС: классификация машинного обучения	213
Проблема размерности.....	215
Сингулярное разложение (SVD)	216

Важные свойства SVD	217
Метод главных компонент (PCA).....	218
Вариант метода наименьших квадратов	219
Фиксируйте V и скорректируйте U	220
Последние размышления о данных алгоритмах	221
Мысленный эксперимент: фильтр для пузырей	221
Упражнение: постройте собственную рекомендательную систему.....	222
Пример кода на Python	222
Глава 9. Визуализация данных и выявление мошенничества	224
История визуализации данных	224
Габриэль Тард	225
Мысленный эксперимент Марка	226
Что такое возрожденная наука о данных	227
Processing	228
Франко Моретти	228
Примеры проектов визуализации данных.....	229
Проекты визуализации данных от Марка	233
Фойе The New York Times: «Наборный шрифт»	233
Проект «Каскад»: жизнь на экране	235
Кронкайт Плаза	236
Транзакции eBay и Books.....	236
Общественный театр «Машина для Шекспира»	239
Цели этих экспозиций.....	240
Наука о данных и риски	240
O Square.....	241
Проблема рисков.....	242
Проблема оценки эффективности	245
Советы по построению моделей	248
Визуализация данных в Square	252
Мысленный эксперимент Яна	254
Визуализация данных для остальной части	254
Глава 10. Социальные сети и журналистика данных	257
Анализ социальных сетей в Morning Analytics.....	257
Анализ социальных сетей.....	259
Терминология из социальных сетей.....	260
Показатели центральности.....	261
Индустрия показателей центральности.....	262
Мысленный эксперимент	263
Morningside Analytics.....	264

Дополнительные сведения об анализе социальных сетей с точки зрения статистики	267
Представление сетей и характеристическое число центральности	267
Первый пример случайных графов: модель Эрдеша — Реньи	269
Второй пример случайных графов: экспоненциальная модель случайных графов	269
Журналистика данных	272
Немного из истории журналистики данных	273
Техническая документация в журналистике: совет профессионала	273
Глава 11. Причинность	275
Корреляция не подразумевает причинности	276
Задаем причинные вопросы	277
Искажающие факторы: на примере сайта знакомств	277
Пример с сайта знакомств ОК Cupid	278
Золотой стандарт: рандомизированные клинические испытания	281
А/В-тестирование	283
Второе место: исследования методом наблюдения	285
Парадокс Симпсона	285
Причинно-следственная модель Рубина	287
Визуализация причинности	288
Определение: причинно-следственное влияние	289
Три совета	291
Глава 12. Эпидемиология	292
О Мэдигане	292
Мысленный эксперимент	293
Современная академическая статистика	294
Медицинская литература и исследования методом наблюдения	295
Стратификация не решает проблему искажающих факторов	295
Есть ли лучший способ?	298
Экспериментальное исследование (партнерство по наблюдению за медицинскими результатами, ОМОР)	299
Завершение мысленного эксперимента	304
Глава 13. Уроки, извлеченные из соревнований по данным:	
утечка данных и оценка моделей	305
Профиль Клаудии как исследователя данных	306
Жизнь главного исследователя данных	306
О том, каково это: быть женщиной — исследователем данных	307
Соревнования по интеллектуальному анализу данных	307
Как стать хорошим моделистом	309

Утечка данных	309
Предсказания рынков.....	310
Кейс Amazon: транжиры	310
Ювелирные изделия: проблема с выборкой.....	311
Таргетинг клиентов IBM.....	312
Выявление рака груди.....	313
Предсказание пневмонии.....	314
Как избежать утечки	315
Оценка моделей	315
Точность: фи	316
Вероятности имеют значение, а не 0 и 1	317
Выбор алгоритма	320
Последний пример.....	321
Финальные мысли	321
Глава 14. Проектирование данных: MapReduce, Pregel и Hadoop	323
О Дэвиде Кроушоу.....	324
Мысленный эксперимент	325
MapReduce	326
Задача подсчета частот слов	327
Другие примеры использования MapReduce	331
Pregel.....	333
О Джоше Уиллсе	333
Еще один мысленный эксперимент	333
Что значит быть исследователем данных	334
Избыток и нехватка данных.....	334
Проектирование моделей.....	334
Экономическая интерлюдия: Hadoop	335
Краткое введение в Hadoop.....	336
Cloudera.....	336
Возвращаемся к Джошу: последовательность выполняемых действий	337
Как же начать работать с Hadoop.....	337
Глава 15. Мнения студентов	339
Мыслительный процесс	339
Более не наивный	341
Неоценимая помощь	342
Длина пройденного пути может варьироваться.....	344
Строим мосты	346
Некоторые из наших работ	347

Глава 16. Исследователи данных нового поколения,	
завышенная самооценка и этика	349
Что вы обрели	349
И все-таки что такое наука о данных.....	350
Кто такие исследователи данных нового поколения.....	352
Умение решать проблемы	352
Развитие личных качеств.....	353
Умение задавать вопросы	354
Моральные принципы исследователей данных	355
Советы по профессиональному развитию.....	361
Об авторах	363
Об иллюстрации на обложке	364