Data Science

Лучшие практики

Достижение успеха
в науке о данных

Дэниел Воган

УДК 004.42
ББК 32.973
В61

Data Science: The Hard Parts
Daniel Vaughan

Воган, Дэниел.

В61 Data Science. Лучшие практики / Дэниел Воган: [перевод с английского Е. Жевлаковой]. — Алматы: Астана иностранная пресса, 2026. — 288 с. — (O’Reilly. Книги по программированию).

ISBN 978-601-12-7679-5

Не только владение алгоритмами, но и умение применять их там, где это действительно имеет значение отличает хорошего специалиста по данным от выдающегося. Книга для тех, кто хочет выйти за рамки учебных примеров и научиться строить проекты, приносящие ощутимую ценность бизнесу.

Автор делится проверенными подходами, которые редко встречаются в классических курсах. В центре внимания не только техника, но и практическое мастерство: как убедительно презентовать результаты, как формулировать сильные аргументы, как превращать аналитику в реальные решения.

Настольный справочник для тех, кто стремится не просто «работать с данными», а создавать проекты, которые влияют на решения и меняют процессы. Она будет одинаково полезна как амбициозным новичкам, так и опытным дата-сайентистам, желающим прокачать навыки и повысить эффективность своей работы.

УДК 004.42
ББК 32.973

ISBN 978-601-12-7679-5

Посвящаю эту книгу своему брату Николасу,
которого я очень люблю и уважаю.

Оглавление

ЧАСТЬ II
Машинное обучение

Предисловие

Позвольте мне исходить из того, что изучать науку о данных и применять ее на практике действительно *сложно*. От вас ожидают отличных навыков программиста, который не только разбирается в тонкостях структур данных и их вычислительной сложности, но и хорошо знаком с Python и SQL. Статистика и новейшие методы прогнозирования с помощью машинного обучения должны быть для вас вторым языком, и, конечно, вы должны уметь применять эти знания для решения реальных бизнес-задач. Однако эта работа сложна еще и потому, что вы должны уметь убедительно общаться со стейкхолдерами, которые, возможно, не имеют технического образования и не привыкли принимать решения на основе данных.

Будем честны: теория и практика в науке о данных *сложны*. И любая книга, посвященная *сложным аспектам* в этой области, либо является энциклопедичной и исчерпывающей, либо в ней необходимо отобрать нужные нам темы.

Прежде всего я должен вас предупредить, что такой отбор я бы провел по *своим* критериям, которые указывают на *сложность* изучения в области науки о данных, и что такое определение субъективно по своей природе. Чтобы смягчить его, давайте проясним: дело не в том, что эти темы замысловаты и их *труднее* изучать, а скорее в том, что на данный момент профессия придает им достаточно низкий вес как *отправным* темам для карьеры в науке о данных. И на практике их труднее освоить, потому что по ним трудно найти материал.

В учебной программе по науке о данных обычно делается упор на изучение программирования и машинного обучения — темам, которые я называю *основополагающими*. Ожидается, что почти всему остальному вы научитесь на рабочем месте, и, к сожалению, это зависит от того, повезет ли вам найти наставника на месте своего первого или второго трудоустройства. Крупные технологические компании хороши тем, что в них довольно часто встречаются талантливые специалисты, поэтому многие из этих *неосвещенных* тем становятся частью местной корпоративной субкультуры, недоступной многим практикам.

Эта книга рассказывает о методах, которые помогут вам стать более продуктивным дата-сайентистом. Я выделил здесь две части: в первой рассматриваются вопросы анализа данных и более простые аспекты этой науки, а во второй — все о машинном обучении (machine learning, ML).

Вы можете спокойно читать книгу в любом порядке; в некоторых главах содержатся ссылки на предыдущие разделы. В большинстве случаев можно пропустить эти ссылки, материал все равно останется понятным для вас. Ссылки в основном используются для того, чтобы показать связь между, казалось бы, независимыми темами.

В части I рассматриваются следующие темы.

Глава 1. Что дальше? Создание ценности с помощью Data science

Какова роль науки о данных в создании ценности для организации и как вы ее оцениваете?

Глава 2. Разработка метрик

Я считаю, что дата-сайентисты лучше всех могут справиться с усовершенствованием разработки *действенных метрик*. И здесь я расскажу вам, как это сделать.

Глава 3. Декомпозиции роста: попутные и встречные ветры

Разобраться с тем, что происходит с бизнесом, и составить убедительный нарратив — это обычная задача для дата-сайентистов. В этой главе я рассказал о некоторых методах *декомпозиции роста*, которые можно использовать для частичной автоматизации этого рабочего процесса.

Глава 4. Матрицы 2×2

Изучение техник, помогающих упростить мир, может занять у вас много времени. *Матрица 2×2* поможет вам сделать это, а также повысить качество взаимодействия со стейкхолдерами.

Глава 5. Создание бизнес-кейсов

Перед началом проекта у вас должен быть *бизнес-кейс*. В этой главе показано, как это сделать.

Глава 6. Что показывает прирост?

Какими бы простыми они ни были, *лифты* могут ускорить анализ, который вы предполагаете сделать с помощью машинного обучения. Я расскажу о лифтах в этой главе.

Глава 7. Нарративы

Дата-сайентисту необходимо развивать свой навык сторителлинга и составлять убедительные *нарративы*. Здесь я расскажу, как это делается.

Глава 8. Datavis: выбираем диаграмму для передачи своей идеи

Уделяя достаточно времени *визуализации данных*, вы сможете составить более убедительный нарратив. В этой главе мы обсудим несколько лучших практик.

Часть II посвящена ML и охватывает следующие темы.

Глава 9. Моделирование и бутстрэп

Методы *моделирования* помогут вам лучше понять различные алгоритмы прогнозирования. Я покажу вам, как это работает, но с некоторыми предостережениями относительно использования ваших любимых методов регрессии и классификации. Мы также рассмотрим *бутстрэп*, который можно применять для определения доверительных интервалов по некоторым трудновычислимым оценкам.

Глава 10. Линейная регрессия: возвращаемся к основам

Наличие глубоких знаний о *линейной регрессии* имеет решающее значение для понимания некоторых более сложных тем. В этой главе я возвращаюсь к основам, чтобы сформировать более ясное понимание алгоритмов машинного обучения.

Глава 11. Утечка данных

Что такое *утечка данных* и как ее заметить и предотвратить?

Глава 12. Вывод модели в продакшн

Модель хороша только в том случае, если она дошла до *стадии продакшна*. К счастью, это понятная и структурированная задача, и я показываю наиболее важные ее шаги.

Глава 13. Сторителлинг в машинном обучении

Есть несколько отличных приемов, которые можно использовать, чтобы понять логику *сторителлинга* в контексте ML и преуспеть в нем.

Глава 14. От прогнозирования к принятию решений

Мы создаем ценность, оттачивая наши способности по принятию решений на основе данных и ML. Здесь я покажу вам примеры перехода *от прогнозирования к принятию решения*.

Глава 15. Инкрементальность: святой Грааль науки о данных?

Техники *выявления причинно-следственной связи* набрали некоторую популярность в науке о данных, однако по-прежнему считаются в некотором роде нишевыми. В этой главе я расскажу об основах и приведу несколько примеров и фрагменты кода, которые можно легко применить в вашей организации.

Глава 16. A/B-тесты

A/B-тесты — типичный пример того, как можно оценить инкрементальность альтернативных вариантов действий. Но эксперименты требуют определенных базовых знаний в области статистики (и бизнеса).

Последняя глава (глава 17) довольно уникальна, поскольку только в ней я не рассказываю о каком-либо методе. Я размышляю о будущем науки о данных после

появления генеративного искусственного интеллекта (ИИ). Мой основной вывод — в ближайшие несколько лет список должностных обязанностей кардинально изменится, а дата-сайентисты должны быть готовы к этой революции.

Эта книга предназначена для дата-сайентистов любого уровня и опыта работы. У вас получится извлечь максимум пользы из книги, если вы владеете средними или продвинутыми знаниями об алгоритмах машинного обучения, поскольку я здесь не трачу время на знакомство с линейной регрессией, классификацией и регрессионными деревьями, а также ансамблевыми методами, такими как случайный лес или градиентный бустинг.

Условные обозначения, используемые в этой книге

В этой книге используются следующие типографские условные обозначения.

Курсив
: Указывает на новые термины, URL-адреса, адреса электронной почты, имена файлов и расширения файлов.

`Моноширинный шрифт`
: Используется в списках программ, а также в тексте при упоминании программных элементов, таких как имена переменных или функций, базы данных, типы данных, переменные окружения, операторы и ключевые слова.

Этот элемент означает подсказку или предложение.

Этот элемент обозначает общее замечание.

Этот элемент указывает на предупреждение.

Использование примеров кода

Дополнительные материалы (примеры кода, упражнения и т. д.) доступны для скачивания по адресу https://oreil.ly/dshp-repo.

Если у вас возникнут технические вопросы или проблемы с использованием примеров кода, пожалуйста, отправьте электронное письмо по адресу bookquestions@oreilly.com.

Эта книга призвана помогать вам в выполнении вашей работы. Если на ее страницах приведены примеры кода, то, как правило, вы можете использовать их в своих программах и документации. Вам не нужно обращаться к нам за разрешением, если только вы не воспроизводите значительную часть кода. Например, для написания программы, использующей несколько фрагментов кода из этой книги, разрешение не требуется. Для продажи или распространения примеров из книг O'Reilly требуется разрешение. Если вы хотите ответить на вопрос, сославшись на эту книгу и приведя пример кода, то разрешение не требуется. Для включения значительного количества примеров кода из этой книги в документацию к вашему продукту требуется разрешение.

Мы ценим указание авторства, но, как правило, не требуем этого. Для указания авторства обычно прописывают название книги, автора, издателя и ISBN. Например: Дэниел Воган, «*Наука о данных: сложные аспекты*» (O'Reilly). Авторское право 2024 Дэниел Воган, 978-1-098-14647-4.

Если вы сомневаетесь и считаете применение в вашем проекте примеров кода выходит за рамки добросовестного использования или вышеуказанного разрешения, смело пишите нам по адресу permissions@oreilly.com.

Онлайн-обучение в O'Reilly

Вот уже более 40 лет *O'Reilly Media* проводит обучение в сфере технологий и бизнеса, делится знаниями и открытиями, которые помогают компаниям добиться успеха.

Мы привлекаем уникальных экспертов и новаторов, которые делятся своими знаниями и опытом с помощью книг, статей и курсов на нашей платформе онлайн-обучения. Платформа онлайн-обучения O'Reilly предоставляет вам доступ по запросу к учебным курсам с преподавателями в прямом эфире, методам углубленного обучения, интерактивным средам программирования и обширной коллекции текстов и видео от O'Reilly и более 200 других издателей. Для получения дополнительной информации посетите сайт *https://oreilly.com*.

Как с нами связаться

Пожалуйста, направляйте комментарии и вопросы, касающиеся этой книги, издателю:

O'Reilly Media, Inc.
95472, Калифорния, Себастопол, Северное Гравенштайнское шоссе, 1005
800-998-9938 (в Соединенных Штатах или Канаде) 707-829-0515 (международные или местные звонки)
707-829-0104 (факс)
support@oreilly.com
https://www.oreilly.com/about/contact.html

Для этой книги у нас есть веб-сайт, где мы приводим список исправлений, примеры и другую дополнительную информацию. Вы можете найти на эту страницу по адресу: *https://oreil.ly/data-science-the-hard-parts*.

Для получения новостей и информации о наших книгах и курсах посетите сайт *https://oreilly.com*.

Благодарности

Многие темы, затронутые в этой книге, я объяснял на внутренних технических семинарах системы CLIP. Поэтому я в долгу перед потрясающей командой датасайентистов, которой я имел честь руководить, где был наставником и учился. Их опыт и знания сыграли важную роль в формировании содержания и структуры этой книги.

Я также глубоко признателен моему редактору Корбину Коллинзу, который терпеливо и внимательно вычитывал рукопись, находил ошибки и упущения, вносил предложения, которые радикально улучшили изложение материала во многих отношениях. Я также хотел бы выразить свою искреннюю признательность Джонатону Оуэну (выпускающему редактору) и Соне Саруба (литературному редактору) за их внимательный взгляд, исключительные навыки и преданность делу. Их совместные усилия значительно повысили качество этой книги, и я бесконечно благодарен им за это.

Большое спасибо техническим рецензентам, которые нашли ошибки и опечатки в тексте и сопроводительном коде, а также внесли предложения по улучшению содержания книги. Особая благодарность Навину Кришнараджу, Бретту Холлеману и Чандре Шукле за подробную обратную связь. Мы часто не соглашались друг с другом, но их конструктивная критика была как усмиряющей пыл,

так и аргументированной. Излишне говорить, что все остальные ошибки — мои собственные.

Они никогда не прочтут этого, но я бесконечно благодарен своим собакам, Матильде и Доминго, за их безграничную способность дарить любовь, смех, нежность и дружеское общение.

Я также благодарен своим друзьям и семье за их безоговорочную поддержку и одобрение. Особая благодарность Клаудии: невозможно переоценить твое любящее терпение, когда я снова и снова обсуждал с тобой некоторые идеи, даже когда они почти ничего не значили для тебя.

Наконец, я хотел бы выразить признательность многочисленным исследователям и практикам в области науки о данных, чьи работы вдохновили меня и оказали влияние на мой труд. Эта книга не появилась бы без их самоотверженности и вклада, и я горд быть частью этого оживленного сообщества.

Спасибо вам всем за поддержку.