УДК 004.43
ББК 32.372.1
Б98

Бюиссон Ф.
Б98 Анализ поведенческих данных на R и Python / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2022. – 368 с.: ил.

ISBN 978-5-97060-992-7

Задействуйте всю мощь поведенческих данных в своей компании, используя инструменты, специально разработанные для их анализа. Автор, эксперт в области экономики и бихевиористики, показывает, как повысить ценность и результаты аналитических проектов за счет понимания того, что движет поведением людей. Практическая часть книги содержит полные примеры и упражнения на языках R и Python, которые помогут вам получать более глубокую информацию о данных.

Издание предназначено для бизнес-аналитиков и других специалистов, исследующих данные и владеющих программированием на R или Python. Для чтения требуется минимальное знакомство с линейной и логистической регрессией.

УДК 004.43
ББК 32.372.1

ISBN 978-1-492-06137-3 (англ.)
ISBN 978-5-97060-992-7 (рус.)

Положительные отзывы на книгу «Анализ поведенческих данных на R и Python»

«В отличие от некоторых книг по науке о данных, в которых авторы стремятся научить своих читателей новым техническим приемам, цель Флорана – иная и более глубокая. Он стремится научить нас мудрости, ориентированной на данные: как строить подробное и тонкое понимание данных, содержащих следы человеческого поведения».

– *Стив Вендель*,
руководитель отдела бихевиористики, Morningstar

«Книга “Анализ поведенческих данных” поможет вам разбираться в данных, даже если вы не можете проводить контролируемые эксперименты».

– *Колин Макфарланд*,
директор платформы экспериментирования, Netflix

«Мы переполнены данными, и эта книга является давно востребованным ресурсом, который направляет практиков в том, как использовать эти данные для строительства достоверных причинно-следственных моделей, которые предсказывают и объясняют поведения в реальном мире».

– *Дэвид Льюис*, президент научно-иссследовательского
института BEworks в компании BEworks

«Для всех, кто хочет применять бихевиористику в качестве проводника в принятии деловых решений, эта книга представляет собой ценное подробное введение в принципы эффективного использования причинно-следственных диаграмм во время экспериментирования и в поведенческом анализе».

– *Мэтт Райт*, директор по бихевиористике, WiderFunnel

«Часть того, что делает бихевиористику работоспособной, заключена в бесшовном сочетании количественных и качественных выводов в поддержку нашего понимания причин, почему люди делают то, что они делают. Эта книга поможет любому человеку, обладающему несколькими базовыми навыками работы с данными, принимать осмысленное участие в этом процессе бихевиористики».

– *Мэтт Уоллерт*,
руководитель отдела бихевиористики в frog,
автор книги «Начало в конце: как создавать продукты,
которые создают изменения»

Содержание

Предисловие

Статистика является предметом удивительно многих применений и инструментом удивительно немногих эффективных практиков.

– Брэдли Эфрон и Р. Дж. Тибширани, «Введение в бутстрап» (1993)

Добро пожаловать в «Анализ поведенческих данных на R и Python»! Высказывание о том, что мы живем в век данных, уже стало банальным. Инженеры теперь регулярно используют сенсорные данные на машинах и турбинах, чтобы предсказывать время, когда они выйдут из строя, и проводят превентивное техническое обслуживание. Аналогичным образом маркетологи используют массивы данных, начиная с вашей демографической информации и заканчивая вашими прошлыми покупками, чтобы определять вид объявления, которое вам следует показывать, и время его показа. Как говорится, «данные – это новая нефть», а алгоритмы – это новый двигатель внутреннего сгорания, двигающий нашу экономику вперед.

В большинстве книг по аналитике, машинному обучению и науке о данных авторы неявно предполагают, что задачи, которые пытаются решать инженеры и маркетологи, могут решаться с помощью одних и тех же подходов и инструментов. Разумеется, переменные имеют разные имена, и необходимо приобретать некоторые знания, относящиеся к конкретной области, но кластеризация *k*-средних – это кластеризация *k*-средних, независимо от того, кластеризуете вы данные о турбинах или сообщения в социальных сетях. Принимая на вооружение инструменты машинного обучения в таком ключе, компании нередко могли точно предсказывать поведения, но ценой более глубокого и богатого понимания того, что на самом деле происходит. Это привело к критике моделей науки о данных как «черных ящиков».

Вместо того чтобы стремиться к точным, но непрозрачным предсказаниям, эта книга стремится ответить на вопрос «Что движет поведением?». Если мы решим отправить электронное письмо потенциальным клиентам, то купят ли они подписку на нашу службу в результате отправки этого электронного письма? И какие группы клиентов должны получать это электронное письмо? Склонны ли пожилые клиенты покупать разные товары, потому что они старше? Как влияет опыт клиентов на лояльность и удержание клиентов? Изменив нашу точку зрения с предсказания поведений на их объяснение и измерение причин, мы сможем снять проклятие «корреляция не есть каузация», которое мешало поколениям аналитиков быть уверенными в результатах своих моделей.

Этот сдвиг не будет связан с введением новых аналитических инструментов: мы будем использовать только два инструмента анализа данных: старую добрую линейную регрессию и ее логистического собрата. Указанные две

модели по своей сути читаются намного легче, чем другие типы моделей. Определенно, это нередко происходит ценой более низкой предсказательной точности (т. е. они допускают все больше и больше ошибок в предсказании), но здесь для нашей цели измерения взаимосвязей между переменными это не имеет значения.

Вместо этого мы потратим много времени на то, чтобы научиться разбираться в данных. В своей роли специалиста, проводящего собеседование по науке о данных, я повидал немало кандидатов, которые были способны использовать сложные алгоритмы машинного обучения, но не развили в себе сильное чувство данных: у них мало интуиции относительно того, что, собственно, происходит в их данных, кроме того что им говорят их алгоритмы.

Я твердо убежден, что вы можете развить эту интуицию и попутно повысить ценность и результаты ваших аналитических проектов – нередко значительно, – приняв следующие меры:

- бихевиористский менталитет, который взирает на данные не как на самоцель, а как на линзу для изучения психологии и поведений людей;
- инструментарий причинно-следственной (каузальной) аналитики, который позволяет нам уверенно утверждать, что одна вещь обусловливает другую, и определять силу этой взаимосвязи.

Хотя каждая из них может приносить большие выгоды сама по себе, я считаю, что они являются естественными дополнениями, которые лучше всего использовать вместе. Учитывая, что словосочетание «бихевиористский менталитет с использованием инструментария причинно-следственной аналитики» трудно выговорить, вместо него я будут называть его причинно-поведенческим подходом, или каркасом. Указанный каркас имеет дополнительную выгоду: он в равной степени применим к экспериментальным и историческим данным, используя при этом различия между ними. Это контрастирует с традиционной аналитикой, которая манипулирует ими с помощью совершенно других инструментов (например, ANOVA и T-тест для экспериментальных данных), и наукой о данных, которая не трактует экспериментальные данные отлично от исторических данных.

Для кого эта книга предназначена

Если вы анализируете данные в бизнесе на R или Python, то эта книга для вас. Я использую слово «бизнес» в широком смысле для обозначения любой коммерческой, некоммерческой или правительственной организации, где важны правильные идеи и практические выводы, которые движут действиями.

С точки зрения математики и статистики, не имеет значения, кем вы являетесь: деловым аналитиком, строящим ежемесячные прогнозы, исследователем опыта пользователей (UX), изучающим поведения на основе кликабельности, или исследователем данных, строящим модели машинного обучения. У этой книги есть одно фундаментальное условие: вы должны быть хотя бы немного знакомы с линейной и логистической регрессией. Если вы понимаете регрессию, то вы сможете проследить за аргументами этой кни-

ги и извлечь из нее большую пользу. С другой стороны, я убежден, что даже опытные исследователи данных с докторскими степенями в области статистики или компьютерных наук найдут этот материал новым и полезным, при условии что они еще не являются специалистами в области поведенческой или причинно-следственной аналитики.

С точки зрения подготовленности в качестве программиста, вы должны уметь читать и писать исходный код на R или Python, в идеале на том и другом. Я не буду показывать вам, как определять функцию или как манипулировать структурами данных, такими как кадры данных в pandas. Уже есть отличные книги, которые справляются с этим лучше, чем я, например «Python для анализа данных» Уэса Маккинни (*Python for Data Analysis*, Wes McKinney, O'Reilly)[1] и «R для науки о данных» Гарретта Гролемунда и Хэдли Уикхэма (*R for Data Science*, Garrett Grolemund and Hadley Wickham, O'Reilly)[2]. Если вы читали какую-либо из этих книг, посещали вводные занятия или использовали хотя бы один из двух языков на работе, то здесь вы будете подготовлены к излагаемому материалу. Точно так же я обычно не буду представлять и обсуждать исходный код, используемый для создания многочисленных рисунков в книге, хотя он будет размещен в репозитории книги на GitHub[3].

Для кого эта книга не предназначена

Если вы работаете в академических кругах или в области, которая требует от вас соблюдения академических норм (например, фармацевтические испытания), то эта книга все еще может представлять для вас интерес, но рецепты, которые я описываю, могут вызывать у вас проблемы с вашим консультантом/редактором/менеджером.

Эта книга не является обзором традиционных методов анализа поведенческих данных, таких как T-тест или ANOVA. Мне еще не приходилось сталкиваться с ситуацией, когда регрессия была менее эффективной, чем эти методы для предоставления ответа на деловой вопрос, поэтому я намеренно ограничиваю эту книгу линейной и логистической регрессией. Если вы хотите изучать другие методы, то вам придется поискать в другом месте (например, в книге «Практическое машинное обучение с помощью Scikit-Learn, Keras и TensorFlow» Орельена Жерона (*Hands-On Machine Learning with Scikit-Learn*, Keras, and TensorFlow, Aurélien Géron, O'Reilly)[4] в отношении алгоритмов машинного обучения).

Понимание и изменение поведений в прикладных условиях требует как анализа данных, так и качественных навыков. В этой книге основное внимание уделяется первому, в первую очередь по соображениям пространства. В дополнение к этому уже есть отличные книги, которые охватывают послед-

---

[1] См. https://www.oreilly.com/library/view/python-for-data/9781491957653/.

[2] См. https://www.oreilly.com/library/view/r-for-data/9781491910382/.

[3] См. https://oreil.ly/BehavioralDataAnalysisCh8.

[4] См. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

нее, такие как «Толчок в верном направлении: совершенствование решений о здоровье, богатстве и счастье» Ричарда Талера и Касса Санштейна (*Nudge: Improving Decisions About Health, Wealth, and Happiness*, Richard Thaler and Cass Sunstein, Penguin) и «Дизайн для изменения поведения: применение психологии и поведенческой экономики» Стивена Венделя (*Designing for Behavior Change: Applying Psychology and Behavioral Economics*, Stephen Wendel, O'Reilly)[1]. Тем не менее я дам введение в концепции бихевиористики, чтобы вы могли применять инструменты из этой книги, даже если вы – новичок в данной области.

Наконец, если вы – абсолютный новичок в анализе данных на R или Python, то эта книга не для вас. Я рекомендую начать с нескольких отличных введений, таких как те, которые упомянуты в этом разделе.

Исходный код на R и Python

Почему именно R и Python? Почему бы не выбрать один язык из перечисленных? Дебаты по теме «R против Python» все еще оживленны и продолжаются в интернете. Этот вопрос, по моему скромному мнению, в сущности, тоже не имеет значения. Реальность такова, что вам придется применять любой язык, который используется в вашей организации, и точка. Однажды я работал в медицинской компании, где по техническим и нормативным причинам доминирующим языком был SAS. Я регулярно использовал R и Python для своих собственных аналитических расчетов, но так как я не мог избежать работы с унаследованным исходным кодом SAS, в течение первого месяца работы я заставил себя усвоить SAS настолько, насколько мне было нужно. Если вы не проведете всю свою карьеру в компании, в которой не используется R или Python, то вы, скорее всего, в любом случае подхватите некоторые основы и того, и другого, так что с таким же успехом вы могли бы постичь двуязычие. Я еще не встречал никого, кто заявил бы, что «обучение чтению исходного кода на [другом языке] было пустой тратой моего времени».

Если исходить из допущения, что вам повезло работать в организации, в которой используется и то, и другое, с каким языком вам следует работать? Я думаю, что это на самом деле зависит от вашего контекста и задач, которые вам приходится выполнять. Например, я лично предпочитаю выполнять разведывательный анализ данных (EDA) на R, но нахожу, что Python намного проще использовать для создания веб-страниц. Советую выбирать, исходя из специфики вашей работы и опираясь на актуальную информацию: оба языка постоянно совершенствуются, и то, что было верно для предыдущей версии R или Python, может оказаться неверным для текущей версии. Например, Python становится гораздо более дружественной средой для EDA, чем когда-либо. Лучше потратить свою энергию на изучение обоих языков, чем на изучение форумов, посвященных выбору лучшего из двух.

---

[1] См. https://www.oreilly.com/library/view/designing-for-behavior/9781492056027/.

Среды исходного кода

В начале каждой главы я буду называть пакеты R и Python, которые необходимо загружать специально для каждой отдельной главы. В дополнение к этому я также буду использовать несколько стандартных пакетов по всей книге; во избежание повторов они называются только здесь (они уже включены во все скрипты в репозитории на GitHub). Вы всегда должны начинать свой исходный код с них, а также с нескольких параметрических настроек:

```
## R
library(tidyverse)
library(boot)       #Требуется для бутстрап-симуляций
library(rstudioapi) #Для загрузки данных из локальной папки
library(ggpubr)     #Для генерирования мультиграфиков

# Задание начального значения случайного числа
# будет обеспечивать воспроизводимость случайных чисел
set.seed(1234)
# Я лично нахожу используемую по умолчанию научную числовую нотацию
# (т. е. с экспонентами) менее удобной для чтения в распечатках, поэтому я ее отменяю
options(scipen=10)

## Python
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt # Для графики
import seaborn as sns           # Для графики
```

Условные обозначения в исходном коде

Я использую R в RStudio. R 4.0 был запущен, когда я писал эту книгу, и я принял эту версию за основу, чтобы сделать книгу как можно более актуальной.

Исходный код R пишется шрифтом, специально предназначенным для исходного кода, с комментарием, указывающим используемый язык, вот так:

```
## R
> x <- 3
> x
[1] 3
```

Я использую Python в среде интерактивной разработки Spyder дистрибутива Anaconda. Обсуждение темы «Python 2.0 против 3.0», надеюсь, уже позади (по меньшей мере, в отношении нового исходного кода; унаследованный исходный код – это уже другая история), и я буду использовать Python 3.7. Условные обозначения, принятые для исходного кода Python, несколько похожи на условные обозначения для R:

```
## Python
In [1]: x = 3
In [2]: x
Out[2]: 3
```

Мы часто будем смотреть на результаты регрессий. Они бывают довольно многословными, с большим объемом диагностики, которая не имеет отношения к аргументам этой книги. Вы не должны пренебрегать ими в реальной жизни, но данный вопрос лучше освещен в других книгах. Поэтому я буду сокращать результат следующим образом:

```
## R
> model1 <- lm(icecream_sales ~ temps, data=stand_dat)
> summary(model1)

...
Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4519.055    454.566  -9.941   <2e-16 ***
temps        1145.320      7.826 146.348   <2e-16 ***
...
```

```
## Python
model1 = ols("icecream_sales ~ temps", data=stand_data_df)
print(model1.fit().summary())

...
                 coef   std err         t     P>|t|      [0.025         0.975]
------------------------------------------------------------------------------
Intercept  -4519.0554   454.566    -9.941     0.000    -5410.439     -3627.672
Temps       1145.3197     7.826   146.348     0.000     1129.973      1160.666
...
```

Программирование в функциональном стиле

Один из шагов перехода от начинающего программиста к программисту среднего уровня состоит в том, чтобы перестать писать скрипты, в которых ваш исходный код представляет собой просто длинную последовательность инструкций, и вместо этого структурировать свой исходный код в функции. В этой книге мы будем писать и многократно использовать функции в разных главах, наподобие приведенных ниже, для строительства бутстраповских[1] интервалов уверенности[2]:

---

[1] Термин «бутстрап» (bootstrap) дословно означает «вытягивание себя за шнурки ботинок». Неплохой аналогией является история барона Мюнхгаузена, который вытянул себя вместе с лошадью из болота за волосы. – *Прим. перев.*

[2] Указанный термин (confidence interval), обозначающий вычисляемый из наблюдаемых данных диапазон, ограниченный нижним и верхним пределами, переведен в книге именно как интервал уверенности, поскольку речь идет об уверенности (confidence) исследователя в своих данных, а не о доверии к ним (trust), а это, как говорят в Одессе, две большие разницы. – *Прим. перев.*

```
## R
boot_CI_fun <- function(dat, metric_fun, B=20, conf.level=0.9){

  boot_vec <- sapply(1:B, function(x){
    cat("итерация бутстрапа ", x, "\n")
    metric_fun(slice_sample(dat, n = nrow(dat), replace = TRUE))})
  boot_vec <- sort(boot_vec, decreasing = FALSE)
  offset = round(B * (1 - conf.level) / 2)
  CI <- c(boot_vec[offset], boot_vec[B+1-offset])
  return(CI)
}
```

```
## Python
def boot_CI_fun(dat_df, metric_fun, B = 20, conf_level = 9/10):

  coeff_boot = []
  # Вычислить коэффициент, представляющий интерес для симуляции
  for b in range(B):
      print("Номер итерации " + str(b) + "\n")
      boot_df = dat_df.groupby("rep_ID").sample(n=1200, replace=True)
      coeff = metric_fun(boot_df)
      coeff_boot.append(coeff)

  # Извлечь интервал уверенности
  coeff_boot.sort()
  offset = round(B * (1 - conf_level) / 2)
  CI = [coeff_boot[offset], coeff_boot[-(offset+1)]]

  return CI
```

Функции также имеют добавочное преимущество в лимитировании остатков непонимания: даже если вы не понимаете, как работают приведенные выше функции, вы все равно можете считать само собой разумеющимся, что они возвращают интервалы уверенности, и следовать остальным рассуждениям, откладывая более глубокое погружение в их исходный код на потом.

Использование примеров исходного кода

Дополнительные материалы (примеры исходного кода и т. д.) доступны для скачивания по адресу https://oreil.ly/BehavioralDataAnalysis.

Адаптированный вариант примеров в виде электронного архива вы можете скачать со страницы книги на веб-сайте https://dmkpress.com/.

Навигация по книге

Стержневая интуитивная мысль книги состоит в том, что эффективный анализ данных основывается на постоянном взаимодействии между тремя компонентами:

- фактическими поведениями в реальном мире и связанными с ними психологическими явлениями, такими как намерения, мысли и эмоции;
- причинно-следственной аналитикой и в особенности причинно-следственными диаграммами;
- данными.

Книга разделена на пять частей:

часть I «Понимание поведений».
Эта часть закладывает основу для причинно-поведенческого каркаса и взаимосвязей между поведениями, причинно-следственным рассуждением и данными;

часть II «Причинно-следственные диаграммы и распутывание».
В этой части вводится понятие спутывания и объясняется, каким образом причинно-следственные диаграммы позволяют нам распутывать наши аналитические расчеты на данных;

часть III «Устойчивый анализ данных».
Здесь мы занимаемся разведкой инструментов для работы с пропущенными данными и знакомим с бутстраповскими симуляциями, поскольку в остальной части книги мы будем широко опираться на бутстраповские интервалы уверенности.
Данные, которые малы по объему, неполные или имеют неправильную форму (например, с несколькими пиками или выбросами), не являются новой проблемой, но она бывает особенно острой с поведенческими данными;

часть IV «Дизайн и анализ экспериментов».
В этой части мы обсудим вопросы дизайна и анализа экспериментов;

часть V «Расширенные инструменты анализа поведенческих данных».
Наконец, мы сводим все вместе, чтобы разведать модерацию, опосредование и инструментальные переменные.

Различные части книги в некоторой степени основаны друг на друге, и поэтому я рекомендую читать их по порядку, по меньшей мере при вашем первом подходе к книге.

Условные обозначения в книге

В книге используются следующие типографические условные обозначения.

Курсивный шрифт
Обозначает новые термины, URL-адреса, адреса электронной почты, имена файлов и расширения файлов.

`Моноширинный шрифт`
Используется для листингов программ, а также внутри абзацев для ссылки на элементы программ, такие как переменные или имена функций, базы данных, типы данных, переменные среды, инструкции и ключевые слова.

`Жирный моноширинный шрифт`
Показывает команды либо другой текст, который должен быть набран пользователем.

`Моноширинный шрифт курсивом`
Показывает текст, который должен быть заменен значениями, передаваемыми пользователем, либо значениями, определяемыми по контексту.


Этот элемент обозначает общее замечание.


Данный элемент обозначает предупреждение или предостережение.

Об авторе

Флоран Бюиссон – поведенческий экономист с 10-летним опытом работы в бизнесе, аналитике и бихевиористике. Еще недавно он основал и в течение четырех лет возглавлял научную группу по бихевиористике в страховой компании Allstate.

Ранее он работал во французской консалтинговой фирме по стратегиям, где использовал экономику и анализ данных для ответа на сложные вопросы эконометрии, например для построения индекса, измеряющего стабильность сельскохозяйственной политики в развивающихся странах от имени Продовольственной и сельскохозяйственной организации ООН. Он также работал в области специализированной медицинской аналитики, анализируя поведение пациентов с тяжелыми заболеваниями.

Флоран публикует научные статьи в таких журналах, как рецензируемый журнал Journal of Real Estate Research, посвященный исследованиям в сфере недвижимости. Он имеет степень магистра эконометрии, а также степень доктора философии в области поведенческой экономики в Университете Сорбонны в Париже.

Часть I

ПОНИМАНИЕ ПОВЕДЕНИЙ

В этой первой части книги дается объяснение причины, почему анализ поведенческих данных требует нового подхода.

В главе 1 будет описан этот новый подход – причинно-поведенческий каркас анализа данных. Мы рассмотрим конкретный пример, показывающий, как даже самые простые аналитические расчеты на данных бывают сорваны присутствием спутывающего фактора. Решение этой проблемы в лучшем случае осложнено, а в худшем – невозможно при использовании традиционных подходов, но новый каркас обеспечивает простой процесс.

В главе 2 будет продолжено изучение особенностей поведенческих данных, обеспечивая при этом осторожное введение в бихевиористику и процесс обеспечения того, чтобы наши данные адекватно отражали соответствующие реально существующие поведения.

Глава 1

Причинно-поведенческий каркас для анализа данных

Как мы обсуждали в предисловии, понимание того, что именно движет поведениями, для того чтобы их изменять, является одной из ключевых целей прикладной аналитики, будь то в коммерческой, некоммерческой или общественной организации. Мы хотим докопаться, почему кто-то купил ту или иную вещь и почему кто-то другой ее не купил. Мы хотим понять, почему кто-то продлил подписку, связался с кол-центром, вместо того чтобы оплатить онлайн, зарегистрировался в качестве донора органа или пожертвовал его некоммерческой организации. Обладание этими знаниями позволяет нам предсказывать, что конкретно люди будут делать в разных сценариях, и помогает нам определять, что именно наша организация может сделать, чтобы побуждать (или не побуждать) их делать это снова. Я считаю, что указанная цель лучше всего достигается путем комбинирования анализа данных с бихевиористским менталитетом и инструментарием причинно-следственной аналитики для создания интегрированного подхода, который я назвал «причинно-поведенческим каркасом». В этом каркасе *поведения* находятся на вершине, потому что их понимание является нашей конечной целью. Это понимание достигается посредством *причинно-следственных диаграмм* и *данных*, которые образуют два опорных столпа треугольника (рис. 1.1).

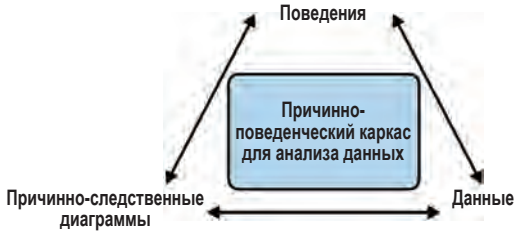

Рис. 1.1 ❖ Причинно-поведенческий каркас для анализа данных

По ходу изложения в этой книге мы разведаем каждую часть треугольника и посмотрим, как они соединяются друг с другом. В заключительной главе мы увидим, как вся наша работа сходится вместе, достигая с помощью одной строки кода того, что при традиционных подходах было бы неимоверно сложной задачей: измерения степени, в которой удовлетворенность клиентов увеличивает будущие расходования клиентов. В дополнение к выполнению таких экстраординарных задач этот новый каркас также позволит вам эффективнее проводить распространенные аналитические расчеты, такие как определение эффекта проводимой по электронной почте рекламной кампании или свойства продукта на поведение клиентов при осуществлении покупок.

Прежде чем перейти к этой теме, читатели, знакомые с предсказательной аналитикой, возможно, зададутся вопросом, почему я вместо нее выступаю за причинно-следственную аналитику. Ответ кроется в том, что, несмотря на то что предсказательная аналитика была (и останется) очень успешной в рамках бизнеса, она бывает недостаточной, когда ваши аналитические расчеты касаются поведений людей. В частности, применение причинно-следственного подхода помогает нам выявлять и устранять «спутывание», очень распространенную проблему с поведенческими данными. Я подробно остановлюсь на этих моментах в остальной части первой главы.

Почему для объяснения человеческого поведения нужна причинно-следственная аналитика

Понимание того, как причинно-следственная аналитика вписывается в аналитический ландшафт, поможет нам лучше уяснить, почему она необходима в рамках бизнеса. Как мы увидим, эта потребность проистекает из сложности человеческого поведения.

Различные типы аналитики

Существует три разных типа аналитики: описательная, предсказательная и причинно-следственная. Описательная аналитика обеспечивает *описание* данных. Проще говоря, я думаю о ней как об аналитике «каким является то или это» или «что именно мы измерили». Под этот зонтик подпадает деловая отчетность. Сколько клиентов отменили свои подписки в прошлом месяце? Сколько прибыли мы получили в прошлом году? Всякий раз, когда мы вычисляем среднее значение или другие простые метрики, мы неявно используем описательную аналитику. Описательная аналитика – это самая простая форма аналитики, но ее также недооценивают. Многие организации на самом деле изо всех сил пытаются получать четкое и единое представление о своей деятельности. Для того чтобы увидеть масштабы этой проблемы

в организации, просто задайте один и тот же вопрос финансовому отделу и операционному отделу и измерьте степень, с которой ответы будут различаться[1].

Предсказательная аналитика обеспечивает *предсказание*. Я думаю о ней как об аналитике «что будет, если сохранятся текущие условия» или «что именно мы еще не измерили». Большинство методов машинного обучения (например, нейронные сети и модели градиентного бустинга) относятся к этому типу аналитики и помогают нам отвечать на такие вопросы, как «Сколько клиентов отменят подписку в следующем месяце?» и «Является ли этот заказ мошенническим?». За последние несколько десятилетий предсказательная аналитика изменила мир; легионы занятых в бизнесе исследователей данных являются свидетельством ее успеха.

Наконец, причинно-следственная аналитика предоставляет *причины* данных. Я думаю о ней как об аналитике «что будет, если?» или «что будет при других условиях». Она отвечает на такие вопросы, как «Сколько клиентов отменят свою подписку в следующем месяце, если мы не отправим им уведомление?». Наиболее известным инструментом причинно-следственной аналитики является A/B-тест, т. н. рандомизированный эксперимент, или рандомизированное контролируемое испытание (randomized controlled trial, аббр. RCT). Это связано с тем, что самый простой и эффективный способ ответить на предыдущий вопрос состоит в том, чтобы отправить купон заранее отобранной группе клиентов и посмотреть, сколько из них отменят подписку по сравнению с контрольной группой.

Мы охватим эксперименты в части IV книги, но перед этим, в части II, мы рассмотрим еще один инструмент из этого инструментария, а именно причинно-следственные диаграммы, которые можно использовать даже тогда, когда мы не можем экспериментировать. И действительно, одна из моих целей состоит в том чтобы побудить вас думать о причинно-следственной аналитике шире, а не просто приравнивать ее к экспериментированию.

Хотя эти ярлыки, возможно, создают впечатление четкой категоризации, на самом деле между этими тремя категориями существует больший градиент, и вопросы и методы между ними размыты. Вы также можете столкнуться с другими терминами, такими как *предписывающая аналитика*, которые еще больше размывают границы и добавляют другие нюансы, не меняя общую картину кардинально.

Люди – сложные существа

Если предсказательная аналитика была настолько успешной, а причинно-следственная аналитика использует те же инструменты анализа данных, что и регрессия, почему бы не придерживаться предсказательной анали-

---

[1] Справедливости ради во многих обстоятельствах они просто *должны* быть разными, потому что данные используются для разных целей и подчиняются разным правилам. Но даже вопросы, на которые вы ожидали бы получить единственно верный ответ (например, «Сколько у нас сейчас сотрудников?»), как правило, обнаруживают расхождения.

тики? Если коротко, потому что люди сложнее, чем ветряные турбины. Поведение человека:

имеет несколько причин.
Поведение турбины не зависит от ее личности, социальных норм сообщества турбин или обстоятельств ее воспитания, в то время как предсказательная сила любой отдельной переменной на поведение человека почти всегда разочаровывает из-за этих факторов;

зависит от контекста.
Незначительные или косметические изменения в окружающей среде, такие как изменение принятого по умолчанию варианта выбора, могут оказывать большое влияние на поведение. Эта ситуация является благословением с точки зрения поведенческого *дизайна*, поскольку она позволяет нам подстегивать изменения в поведениях, но является проклятием с точки зрения поведенческой *аналитики*, потому что это означает, что каждая ситуация уникальна настолько, что становится трудной для предсказания;

является переменным (ученые сказали бы, что поведение недетерминированно).
Один и тот же человек может вести себя совершенно по-другому, когда его неоднократно помещают в одну и ту же ситуацию, которая внешне выглядит совершенно одинаковой, даже после учета косметических факторов. Это может быть связано с преходящими эффектами, такими как настроение, либо долговременными эффектами, такими как скука от приема одного и того же ежедневного завтрака. Оба этих фактора могут радикально менять поведение, но их трудно улавливать;

является инновационным.
Когда условия в окружающей среде меняются, человек может переключаться на поведение, которого он буквально никогда раньше не демонстрировал, и это происходит даже при самых обыденных обстоятельствах. Например, впереди на вашем обычном пути следования происходит автомобильная авария, и поэтому вы в последнюю минуту решаете повернуть направо;

является стратегическим.
Люди делают выводы и реагируют на поведения и намерения других людей. В некоторых случаях это может означать «восстановление» сотрудничества, которое было нарушено внешними обстоятельствами, что делает его устойчиво предсказуемым. Но в других случаях это может предусматривать добровольное запутывание своего поведения, чтобы сделать его непредсказуемым во время соревновательной игры, такой как шахматы (или мошенничество!).

Все эти аспекты человеческого поведения делают его менее предсказуемым, чем поведение физических объектов. В целях отыскания регулярностей, более надежных для анализа, мы должны спуститься на один уровень глубже, чтобы понять и измерить причины поведения. Тот факт, что кто-то

съел овсянку на завтрак и выбрал определенный маршрут в понедельник, не означает, что он сделает то же самое во вторник, но вы можете быть более уверены в том, что он хоть как-то позавтракает и отправится по какому-то маршруту на работу.

Экстраполяция в аналитике, проклятие размерности и критика Лукаса

Читатели с квантитативным опытом, возможно, будут не совсем удовлетворены моим высказыванием о том, что «поведение человека трудно предсказывать, потому что оно сложное», поэтому вот математическая версия этого аргумента. Я начну с описания разницы между интерполяцией и экстраполяцией. На рис. 1.2 показано немного симулированных данных с линейной взаимосвязью между двумя переменными.

Линия на рисунке – это регрессионная линия наилучшей подгонки, т. е. линия, соответствующая линейной регрессии между двумя переменными, с наклоном, приближенно равным 3. Мы можем использовать ее для предсказания неизвестных значений Y на основе известного значения X (и наоборот). Например, имея значение X = 50, мы бы предсказали, что Y равно 150. Слева от этого значения есть наблюдаемые точки, то есть точки, для которых X < 50, а также точки справа от этого значения, для которых X > 50. Этот процесс предсказывания называется интерполяцией, потому что наша точка находится между наблюдаемыми точками (приставка «интер» означает «между»; например, international = «межнациональный»). И наоборот, если бы мы использовали линию регрессии с X = 0, чтобы предсказать, что Y = 0, это было бы названо экстраполяцией, поскольку точка, которую мы пытаемся предсказать, находится за пределами облака наблюдаемых точек («экстра» означает «снаружи»; например, экстраординарное = «вне обычного»). В статистике и в повседневной жизни экстраполировать – значит покидать область наблюдаемого и известного, чтобы делать предсказания. В то время как интерполяция обычно безопасна и надежна, экстраполяция всегда несколько умозрительна: требуется «рывок веры», чтобы допустить, что правила, применяемые внутри неких границ, по-прежнему будут соблюдаться за их пределами.

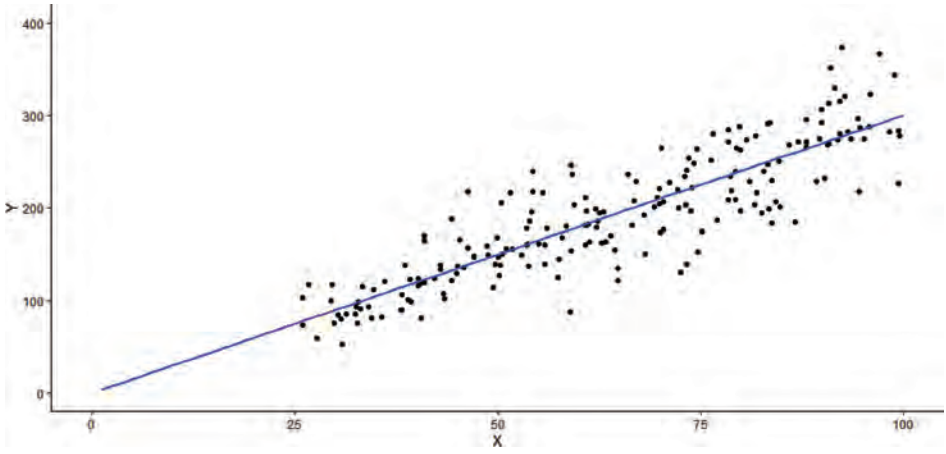

Рис. 1.2 ❖ Линейная связь между двумя переменными, с линией регрессии

Физические объекты, такие как ветряная турбина, находятся под воздействием достаточно малого и постоянного числа факторов (это не похоже на то, как некоторые законы физики берут выходные или новые дни и делают их похожими на случайные). Следовательно, у нас есть много точек данных относительно размерностей пространства задачи, а значит, мы почти всегда интерполируем. Для простоты модель может пренебрегать вторичными или редкими явлениями, такими как шторм 1 раз в 100 лет, но даже когда подобные выбросы происходят, результат остается в какой-то мере предсказуемым: бритвенное лезвие сломается и упадет в воду, а не улетит.

С другой стороны, поведение человека находится под воздействием большого числа разных факторов, которые могут быть или не быть релевантными в данный день и могут расти или затухать с течением времени. Следовательно, у нас обычно оказывается мало точек данных относительно размеров пространства задачи, а значит, мы гораздо чаще экстраполируем – наталкиваясь на проблему, известную в статистике под названием «проклятия размерности». В дополнение к этому незначительные изменения в окружающей среде могут приводить к серьезным изменениям в поведении, что делает попытку предсказывать будущее поведение человека, основываясь только на прошлом поведении, азартной игрой с плохими шансами на успех.

Для людей, интересующихся генеалогией поведенческой экономики, макроэкономист Роберт Лукас сформулировал этот аргумент в 1970-х годах («критика Лукасом» кейнсианских моделей). Вместо этого он рекомендовал выявлять более глубокие параметры человеческих поведений, такие как потребительские предпочтения, – еще одна версия аргумента, который я приводил ранее.

Чтоб ей пусто было! Скрытые опасности, когда разбирательства отданы на усмотрение регрессии

В предыдущем разделе я упомянул, что причинно-следственная аналитика часто использует те же инструменты, что и предсказательная аналитика. Однако, поскольку у них разные цели, инструменты используются по-разному. Так как регрессия является одним из главных инструментов для обоих типов аналитики, она бывает отличным средством иллюстрирования разницы между предсказательной и причинно-следственной аналитикой. Регрессия, подходящая для предсказательной аналитики, зачастую приводила бы к ужасной регрессии для целей причинно-следственной аналитики, и наоборот.

Регрессия в предсказательной аналитике используется для оценивания неизвестного значения (часто, но не всегда, в будущем). Она делает это, беря известную информацию и используя различные факторы для триангулирования наилучшей догадки в отношении данной переменной. Важным является предсказанное значение и его точность, а не то, почему или как оно было предсказано.

В причинно-следственной аналитике регрессия тоже используется, но фокус внимания лежит не на оценивании значения целевой переменной. Вместо этого основное внимание уделяется причине этого значения. В терминах

регрессии нас больше интересует не сама зависимая переменная, а ее связь с данной независимой переменной. При правильно структурированной регрессии коэффициент корреляции может быть переносимой мерой причинно-следственного эффекта независимой переменной на зависимую переменную.

Но что значит иметь правильно структурированную регрессию для этой цели? Почему мы не можем просто взять регрессии, которые мы уже используем для предсказательной аналитики, и рассматривать предоставленные коэффициенты как меры причинно-следственной связи? Мы не можем этого сделать, потому что каждая переменная в регрессии может модифицировать коэффициенты других переменных. Следовательно, наша смесь переменных должна быть искусно изготовлена не для создания наиболее точного предсказания, а для создания наиболее точных коэффициентов. Два набора переменных, как правило, различаются, потому что переменная может быть сильно коррелирована с нашей целевой переменной (и, следовательно, быть очень предсказуемой), фактически не влияя на эту переменную.

В этом разделе мы проведем разведку вопросов, почему эта разница имеет важность в перспективе и почему отбор переменных – это более чем половина битвы в поведенческой аналитике. Мы сделаем это на конкретном примере C-Mart, вымышленной сети супермаркетов с магазинами по всей территории Соединенных Штатов. Первая из двух вымышленных компаний, используемых на протяжении всей книги, C-Mart, поможет нам понять возможности и трудности анализа данных для традиционных компаний в цифровую эпоху.

Данные

Папка этой главы в репозитории на GitHub[1] содержит два CSV-файла, *chap1-stand_data.csv* и *chap1-survey_data.csv*, с наборами данных для двух примеров этой главы.

В табл. 1.1 показана информация, содержащаяся в CSV-файле *chap1-stand_data.csv*, о продажах мороженого и холодного кофе на ежедневном уровне в киосках C-Mart.

Таблица 1.1. Информация о продажах в файле chap1-stand_data.csv

Имя переменной	Описание переменной
IceCreamSales (ПродажиМороженого)	Ежедневные продажи мороженого в киосках C-Mart
IcedCoffeeSales (ПродажиХолодногоКофе)	Ежедневные продажи холодного кофе в киосках C-Mart
SummerMonth (ЛетнийМесяц)	Двоичная переменная, которая относит день к летним месяцам
Temp (Температура)	Средняя температура за этот день и у этого киоска

В табл. 1.2 показана информация, содержащаяся в CSV-файле *chap1-survey_data.csv*, полученная в результате опроса прохожих за пределами киосков C-Mart.

[1] См. https://oreil.ly/BehavioralDataAnalysisCh1.

Таблица 1.2. Информация об опросе в файле chap1-survey_data.csv

Имя переменной	Описание переменной
VanillaTaste (ПредпочтениеВанильногоВкуса)	Пристрастие опрашиваемого к ванильному вкусу, 0–25
ChocTaste (ПредпочтениеШоколадногоВкуса)	Пристрастие опрашиваемого к шоколадному вкусу, 0–25
Shopped (Покупал)	Двоичная переменная, которая указывает на то, что опрашиваемый когда-либо совершал покупки в местном киоске C-Mart

Почему корреляция не есть каузация: спутывающий фактор в действии

В каждом магазине C-Mart есть киоск с мороженым. Компания считает, что на ежедневные продажи мороженого влияет погода – или, выражаясь на жаргоне причинно-следственных связей, что погода является причиной продаж. Другими словами, при прочих равных условиях мы исходим из допущения, что люди с большей вероятностью будут покупать мороженое в более жаркие дни, что имеет интуитивно понятный смысл. Это мнение подтверждается сильной корреляцией исторических данных между температурой и продажами, как показано на рис. 1.3 (соответствующие данные и исходный код находятся в репозитории книги на GitHub).

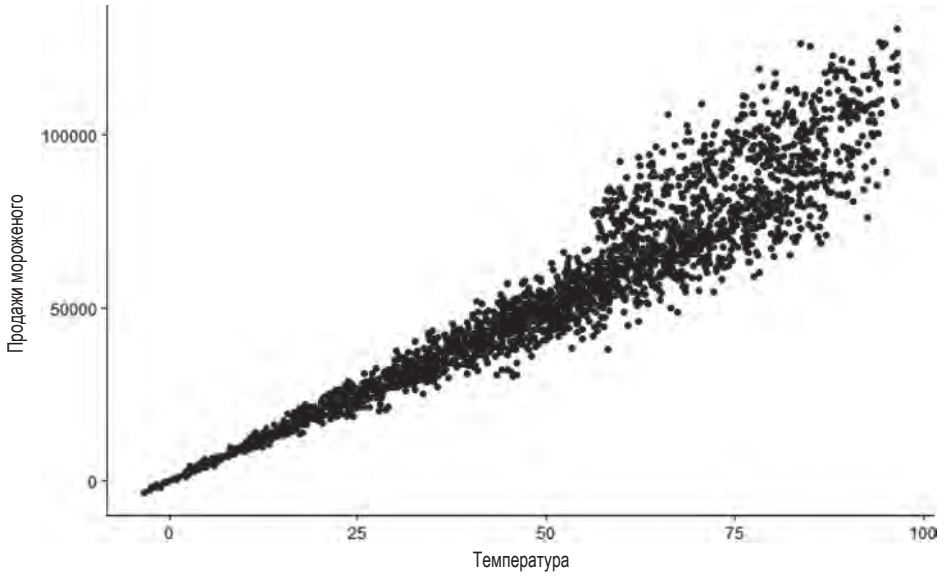

Рис. 1.3 ❖ Продажи мороженого как функция от наблюдаемой температуры

Как указано в предисловии, мы будем использовать регрессию в качестве нашего главного инструмента анализа данных. Выполнение линейной регрессии продаж мороженого на температуре занимает одну строку исходного кода:

```
## Python (результат не показан)
print(ols("icecream_sales ~ temps", data=stand_data_df).fit().summary())

## R
> summary(lm(icecream_sales ~ temps, data=stand_dat))
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4519.055    454.566  -9.941   <2e-16 ***
Temps        1145.320      7.826 146.348   <2e-16 ***
...
```

Для наших целей в этой книге наиболее важной частью результата на выходе является раздел *коэффициентов*, в котором говорится, что оценочное пересечение (коэффициент сдвига) – теоретическое среднее значение продаж мороженого при температуре ноль градусов – составляет –4519, что, очевидно, является бессмысленной экстраполяцией. Он также говорит нам о том, что оценочный коэффициент для температуры составляет 1145, а значит, каждый дополнительный градус температуры, как ожидается, будет увеличивать продажи мороженого на 1145 долларов.

Теперь давайте вообразим, что мы находимся в конце особенно теплой недели октября и, основываясь на предсказаниях модели, компания заранее увеличила запасы в киосках с мороженым. Тем не менее еженедельные продажи, хотя и были выше, чем обычно на этой неделе октября, сильно не дотянули до объема, предсказанного моделью. Та-ак! И что же случилось? Следует ли уволить аналитика данных?

Произошло то, что модель не учитывает важнейший факт: бóльшая часть продаж мороженого приходится на летние месяцы, когда дети не ходят в школу. Регрессионная модель сделала свое лучшее предсказание на основе имеющихся данных, но часть причины увеличения продаж мороженого (летние каникулы у учеников) была ошибочно отнесена к температуре, так как летние месяцы положительно коррелируют с температурой. Поскольку повышение температуры в октябре не привело к внезапным летним каникулам (вы уж извините, детки!), мы увидели более низкие продажи, чем в летние дни при такой температуре.

С технической точки зрения, месяц года – это спутывающий фактор в нашей взаимосвязи между температурой и продажами. *Спутывающий фактор*[1] – это переменная, которая вносит систематическое смещение в регрессию; когда в ситуации, которую вы анализируете, присутствует спутывание, это означает, что интерпретация коэффициента регрессии как причинно-следственного будет приводить к ненадлежащим выводам.

---

[1] Для справки вот выжимка определения указанного термина в переводе с нескольких языков. Спутывающий фактор (confounder), или повреждающий, помеховый, мешающий, смешивающий фактор, – это переменная, которая влияет как на зависимую переменную, так и на независимую переменную, вызывая ложную ассоциацию. Являясь причинно-следственным понятием, указанная переменная имеет связь как с интересующей причиной, так и с интересующим следствием. – *Прим. перев.*

Давайте подумаем о таком месте, как Чикаго, где присутствует континентальный климат: зима – очень холодная, а лето – очень жаркое. Сравнивая продажи в случайный жаркий день с продажами в случайный холодный день без учета соответствующего месяца года, вы, скорее всего, будете сравнивать продажи в жаркий летний день, когда дети не ходят в школу, с продажами в холодный зимний день, когда дети учатся в школе; эта ситуация раздувает очевидную взаимосвязь между температурой и продажами.

В приведенном примере мы также можем ожидать неуклонного занижения предсказания продаж в более холодную погоду. По правде говоря, в летние месяцы происходит сдвиг парадигмы, и когда этим сдвигом приходится управлять исключительно с помощью температуры в линейной регрессии, коэффициент регрессии для температуры неизменно будет слишком высоким для более теплых температур и слишком низким для более холодных температур.

Слишком много переменных может испортить всю обедню

Потенциальным решением проблемы спутывающих факторов было бы добавление в регрессию всех переменных, которые можно добавить. Менталитет в стиле «все, что есть, и кухонная раковина в придачу» все еще имеет сторонников среди статистиков. В книге «Книга вопросов почему» Джуди Перл и Дана Макензи (*The Book of Why*, Judea Pearl and Dana Mackenzy) упоминают, что «даже ведущий статистик недавно написал, что “избегание обусловленности на некоторых наблюдаемых ковариатах[1]… является ненаучной сиюминутной эквилибристикой”» (Pearl & Mackenzie 2018, стр. 160)[2]. Это также довольно распространено среди исследователей данных. Справедливости ради, если ваша цель состоит только в том, чтобы предсказывать переменную, и у вас есть модель, обстоятельно продуманная для обобщения за пределы ваших тестовых данных, и вас не волнует, почему предсказанная переменная принимает некоторое значение, то это совершенно правильная позиция. Но это не сработает, если ваша цель состоит в том, чтобы понять причинно-следственные связи, дабы действовать в соответствии с ними. По этой причине простое добавление как можно большего числа переменных в вашу модель не только является неэффективным, но и может стать совершенно контрпродуктивным и вводить в заблуждение.

Давайте продемонстрируем это на нашем примере, добавив переменную, которую мы могли бы включить, но которая будет систематически смещать нашу регрессию. Я создал переменную *ПродажиХолодногоКофе*, чтобы она коррелировала с *Температурой*, но не с *ЛетнимМесяцем*. Давайте посмотрим,

---

[1] Ковариат (covariate) – это объясняющая переменная, естественным образом существующая в исследуемой модели и могущая являться предсказательной. – *Прим. перев.*

[2] На всякий случай, если вам интересно, вышеупомянутого статистика зовут Дональдом Рубином (Donald Rubin).

что произойдет с нашей регрессией, если мы добавим эту переменную в дополнение к *Температуре* и *ЛетнимМесяцам* (двоичной переменной, обозначающей месяц июль или август как 1 и любой другой месяц как 0):

```
## R (результат не показан)
> summary(lm(icecream_sales ~ iced_coffee_sales + temps + summer_months))

## Python
print(ols("icecream_sales ~ temps + summer_months + iced_coffee_sales",
          data=stand_data_df).fit().summary())
...
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         24.5560    308.872     0.080      0.937    -581.127     630.239
Temps          -1651.3728   1994.826    -0.828      0.408   -5563.136    2260.391
summer_months   1.976e+04    351.717    56.179      0.000    1.91e+04    2.04e+04
iced_coffee_sales 2.6500       1.995     1.328      0.184      -1.262       6.562
...
```

Мы видим, что коэффициент *Температуры* резко изменился по сравнению с нашим предыдущим примером до такой степени, что теперь он – отрицательный. Высокие *p*-значения для *Температуры* и *ПродажХолодногоКофе* обычно воспринимаются как признаки того, что что-то не так, но поскольку *p*-значение для *Температуры* «хуже», аналитик может прийти к выводу, что он должен удалить его из регрессии. Как это возможно?

Истина, лежащая в основе данных (которая по понятным причинам известна, так как я сфабриковал связи вручную и рандомизировал данные вокруг этих связей), заключается в том, что, когда становится жарко, люди с большей вероятностью покупают холодный кофе. В жаркие дни люди также с большей вероятностью покупают больше мороженого. Но покупка холодного кофе сама по себе не делает покупателей более или менее склонными покупать мороженое. Летние месяцы также не коррелируют с покупками холодного кофе, поскольку школьники не являются существенным фактором спроса на холодный кофе (подробную информацию о применяемой математике см. во врезке).

Техническое более глубокое погружение: что здесь произошло?

Уравнение для продаж мороженого, которое я использовал для генерирования симулированных данных, выглядит следующим образом:

$$\textit{ПродажиМороженого} := 1000 \cdot \textit{Температура} + 20\,000 \cdot \textit{ЛетнийМесяц} + \varepsilon_1,$$

где ε_1 представляет некий случайный шум со средним значением, равным нулю, а знак «:=» обозначает, что это уравнение представляет то, как переменная слева, *ПродажиМороженого*, определяется или строится.

Однако уравнение, которое мы оцениваем в нашей линейной регрессии, таково:

$$\textit{ПродажиМороженого} = \beta_T \cdot \textit{Температура} + \beta_S \cdot \textit{ЛетнийМесяц} + \beta_C \cdot \textit{ПродажиХолодногоКофе}.$$

Истинное уравнение, которое использовалось для генерирования продаж холодного кофе, таково:

$$ПродажиХолодногоКофе := 1000 \cdot Температура + \varepsilon_2.$$

Приведенное выше означает, что мы можем переписать предыдущее уравнение следующим образом:

$$ПродажиМороженого = \beta_T \cdot Температура + \beta_S \cdot ЛетнийМесяц + \beta_C \cdot (1000 \cdot Температура + \varepsilon_2) = (\beta_T + 1000\ \beta_C) \cdot Температура + \beta_S \cdot ЛетнийМесяц.$$

За исключением некого случайного везения, наш коэффициент β_S должен быть близок к истинному значению, равному 20 000. Но в случае температуры наша программа попытается решить уравнение, которое приведено ниже:

$$\beta_T + 1000 \cdot \beta_C = 1000.$$

Это одно уравнение с двумя неизвестными, поэтому оно имеет бесконечное число решений. Будут работать $\beta_T = 0$ и $\beta_C = 1$, но будут работать и $\beta_T = 500$ и $\beta_C = 0.5$ либо $\beta_T = 5000$ и $\beta_C = -4$. Алгоритм наименьших квадратов выберет комбинацию, которая обеспечивает наибольшее значение R^2, но она не будет надежной (хотя на практике ненадежность, как правило, будет намного меньше, чем в этом симулированном примере). В техническом плане мы ввели мультиколлинеарность.

На рис. 1.4 показана положительная корреляция между продажами холодного кофе и продажами мороженого, поскольку и то, и другое увеличивается, когда становится теплее, однако любое увеличение продаж холодного кофе в летние месяцы можно объяснить совместной корреляцией с переменной температуры. Когда регрессионный алгоритм пытается объяснить продажи мороженого с использованием трех имеющихся переменных, объяснительная сила температуры на продажах холодного кофе была добавлена в переменную температуры, тогда как холодный кофе был вынужден компенсировать придание избыточной силы температуре. Несмотря на то что продажи холодного кофе статистически не значимы и коэффициент этой переменной относительно невелик, величина продаж в долларах намного выше, чем величина в градусах температуры, поэтому в конечном счете продажи холодного кофе нивелируют инфляцию коэффициента температуры.

В предыдущем примере добавление переменной *ПродажиХолодногоКофе* в регрессию запутало взаимосвязь между температурой и продажами мороженого. К сожалению, верно и обратное: включение в регрессию неправильной переменной может создавать иллюзию взаимосвязи, когда ее нет.

Придерживаясь нашего примера с мороженым в C-Mart, предположим, что категорийный менеджер заинтересован в понимании вкусов покупателей, поэтому он просит сотрудника встать у входа в магазин и опрашивать проходящих мимо людей о том, насколько им нравится ванильное мороженое и насколько им нравится шоколадное мороженое (оба по шкале от 0 до 25), а также о том, покупали ли они когда-либо мороженое в киоске. В целях обеспечения простоты мы будем исходить из того, что в киоске продается только шоколадное и ванильное мороженое.

Будем считать для примера, что вкус ванильного мороженого и вкус шоколадного мороженого совершенно не связаны. Некоторым людям нравится

одно, но не другое, некоторым одинаково нравится и то, и другое, некоторым одно нравится больше, чем другое, и так далее. Но все эти предпочтения влияют на то, покупает человек в киоске или нет, т. е. на двоичную переменную (Да/Нет).

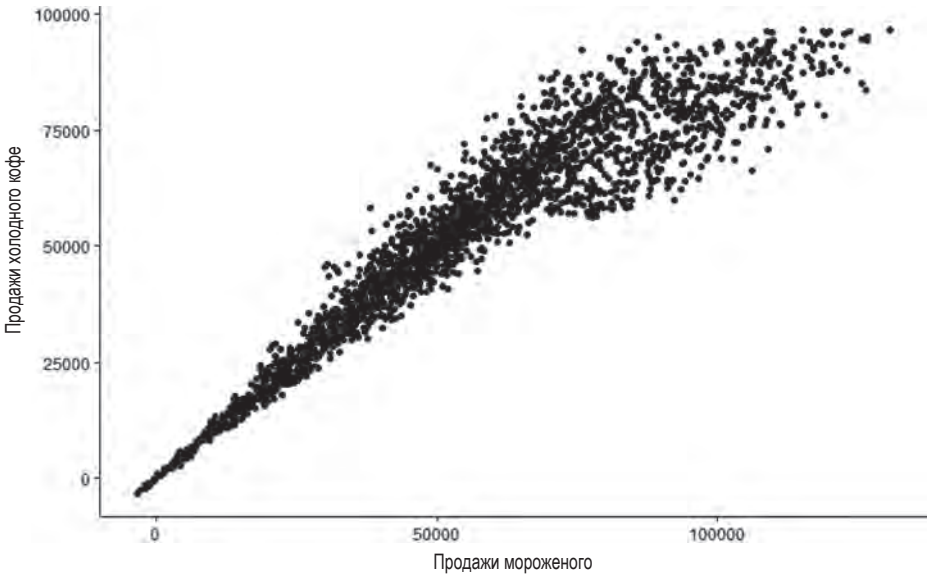

Рис. 1.4 ❖ График продаж холодного кофе в сопоставлении с продажами мороженого

Поскольку переменная *Покупал* является двоичной, мы бы использовали логистическую регрессию, если бы хотели измерить влияние любой из переменных *Вкуса* на поведение при осуществлении покупок. Поскольку две переменные *Вкуса* некоррелированы, мы бы увидели обычное облако без видимой корреляции, если бы сопоставили их друг с другом; однако каждая из них влияет на вероятность покупок в киоске с мороженым (рис. 1.5).

На первом графике я добавил линию наилучшей подгонки, которая является почти идеально горизонтальной, что отражает отсутствие корреляции между переменными (коэффициент корреляции равен 0.004, который отражает ошибку отбора). На втором и третьем графиках мы видим, что предпочтение ванильного вкуса и шоколадного вкуса в среднем выше у покупателей (*Покупал* = 1), чем у непокупателей, что имеет смысл.

Пока что все идет хорошо. Допустим, после того как вы получаете данные опроса, ваш деловой партнер сообщает вам о том, что он подумывает о введении поощрительного купона для киоска с мороженым: когда вы покупаете мороженое, вы получаете купон на случай будущих посещений. Этот фактор лояльности не повлияет на респондентов, которые никогда не делали покупки в киоске, поэтому релевантной популяцией являются те, кто делал покупки в магазине. Деловой партнер рассматривает возможность использования вкусовых ограничений в купонах для балансировки запасов, но не знает,

насколько можно повлиять на выбор того или иного вкуса покупателем. Если бы кто-то, купивший ванильное мороженое, получил купон на скидку 50 % на шоколадное мороженое, то сделает это что-то, кроме добавления большего количества бумаги в корзину для мусора? В любом случае, насколько благосклонно люди, которые любят ванильное мороженое, смотрят на шоколадное мороженое?

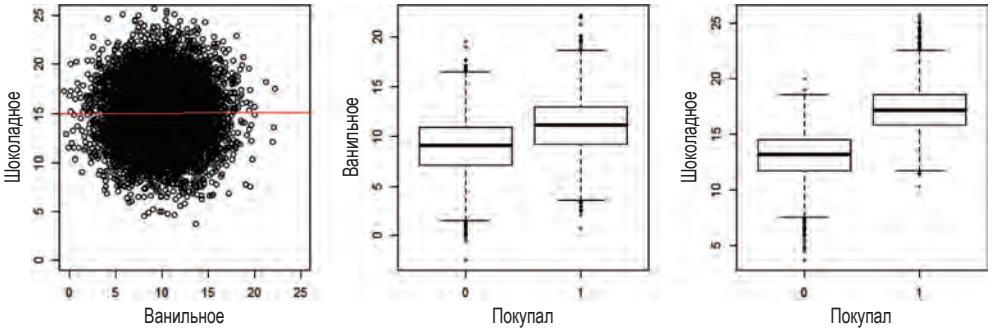

Рис. 1.5 ❖ Левая панель: пристрастия к ванильному и шоколадному вкусам не коррелируют в совокупной популяции; средняя панель: пристрастия к ванильному вкусу выше у людей, которые покупают в киоске с мороженым, чем для людей, которые этого не делают; правая панель: тот же результат для пристрастия к шоколадному вкусу

Вы снова строите тот же график, на этот раз ограничивая данные людьми, которые на вопрос о покупках ответили «Да» (рис. 1.6).

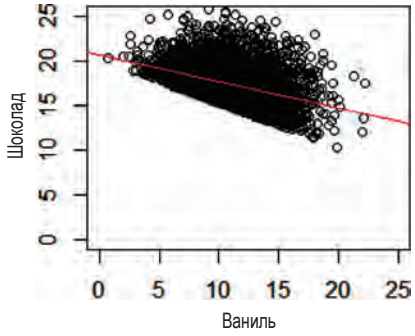

Рис. 1.6 ❖ Предпочтение ванильного вкуса и шоколадного вкуса среди покупателей

В настоящее время между этими двумя переменными существует сильная отрицательная корреляция (коэффициент корреляции равен –0.39). Что случилось? Неужели любители ванили, которые подходят к вашему киоску, превращаются в ненавистников шоколада, и наоборот? Конечно же нет. Эта корреляция была создана искусственно, когда вы сдерживали себя по отношению к покупателям.

Давайте вернемся к нашим истинным причинно-следственным связям: чем сильнее у кого-то пристрастие к ванильному вкусу, тем больше вероятность того, что он будет делать покупки в вашем киоске, и то же самое для шоколадного вкуса. Это означает, что существует кумулятивный эффект этих двух переменных. Если у кого-то слабое вкусовое пристрастие и к ванильному, и к шоколадному мороженому, то он вряд ли будет делать покупки в вашем киоске; другими словами, большинство людей со слабым пристрастием к ванильному вкусу среди ваших покупателей имеют сильное пристрастие к шоколадному вкусу. С другой стороны, если у кого-то есть сильное пристрастие к ванильному вкусу, то он, возможно, будет делать покупки в вашем киоске, даже если у него нет сильного пристрастия к шоколадному вкусу. Отражение этого факта можно увидеть на предыдущем графике: для высоких значений пристрастия к ванильному вкусу (скажем, выше 15) существуют точки данных с более низкими значениями пристрастия к шоколадному вкусу (ниже 15), тогда как для низких значений пристрастия к ванильному вкусу (ниже 5) единственные точки данных на графике имеют высокое значение пристрастия к шоколадному вкусу (выше 17). Ничьи пристрастия не изменились, но люди со слабым пристрастием и к ванильному вкусу, и к шоколадному вкусу исключены из вашего набора данных.

Указанное явление имеет свой технический термин – *парадокс Берксона*[1], но Джуди Перл и Дана Маккензи дают ему более интуитивное название: «эффект оправдания» (explain-away effect, или эффект отмазки). Если у одного из ваших покупателей есть сильное пристрастие к ванильному вкусу, то это полностью объясняет, почему он делает покупки в вашем киоске, и ему «не нужно» иметь сильное пристрастие к шоколадному вкусу. С другой стороны, если у одного из ваших покупателей есть слабое пристрастие к ванильному вкусу, то это не может объяснить, почему он делает покупки в вашем киоске, и у него должно быть более сильное, чем в среднем, пристрастие к шоколадному вкусу.

Парадокс Берксона противоречит здравому смыслу, и поначалу его трудно понять. Он может приводить к систематическому смещению в ваших данных, в зависимости от того, как они были собраны, даже до того, как вы начнете какой-либо анализ. Классическим примером того, как эта ситуация может создавать искусственные корреляции, является то, что некоторые заболевания демонстрируют более высокую степень корреляции, если рассматривать популяцию пациентов больниц в сопоставлении с общей популяцией. На самом деле, конечно же, происходит то, что и той, и другой болезни недостаточно для того, чтобы попасть в больницу; чье-то состояние здоровья становится настолько неважнецким, что оправдывает госпитализацию только тогда, когда обе присутствуют[2].

---

[1] См. https://oreil.ly/KwJ1R.

[2] С технической точки зрения, это несколько иная ситуация, поскольку вместо двух линейных (или логистических) связей существует пороговый эффект, но основополагающий принцип, согласно которому включение неправильной переменной может приводить к искусственным корреляциям, по-прежнему применим.

Выводы

Предсказательная аналитика была чрезвычайно успешной в течение последних нескольких десятилетий и останется таковой. С другой стороны, при попытке понять и – что важнее – изменить поведение человека причинно-следственная аналитика предлагает убедительную альтернативу.

Причинно-следственная аналитика, однако, требует иного подхода, чем тот, к которому мы привыкли с предсказательной аналитикой. Надеюсь, примеры в этой главе убедили вас в том, что вы не можете просто вбросить кучу переменных в линейную или логистическую регрессию и надеяться на лучшее (что мы могли бы рассматривать как подход «включи все, что есть, и Бог подскажет свой вариант»). Однако вы все еще, возможно, задаетесь вопросом о других типах моделей и алгоритмов. Являются ли модели градиентного бустинга или глубокого обучения каким-то образом невосприимчивыми к спутывающим факторам, мультиколлинеарности и ложным корреляциям? К сожалению, ответ отрицательный. Во всяком случае, из «черно-ящичной» природы этих моделей вытекает, что отлавливать спутывающие факторы становится труднее.

В следующей главе мы разведаем вопрос о том, как думать о самих поведенческих данных.