Spis treści

Część II. Diagramy przyczynowe i usuwanie czynników zakłócających

Wstęp

> Statystykę można wykorzystać w zdumiewająco wielu zastosowaniach, lecz niewiele osób stosuje ją w sposób praktyczny.
>
> – Bradley Efron i R. J. Tibshirani, *An Introduction to the Bootstrap* (1993)

Witamy w książce *Analiza danych behawioralnych przy użyciu języków R i Python*! Stwierdzenie, że żyjemy w epoce danych, jest już banałem. Inżynierowie rutynowo wykorzystują dane z czujników zainstalowanych w maszynach i turbinach, aby przewidzieć, kiedy pojawi się awaria. Dzięki temu mogą przeprowadzać zapobiegawcze konserwacje sprzętu. Podobnie też marketingowcy wykorzystują wiele różnych informacji, od danych demograficznych aż po wiedzę o wcześniejszych zakupach, aby ustalić, jakie reklamy i kiedy należy wyświetlać. Istnieje takie powiedzenie: „Dane są paliwem". Jeśli tak jest, wówczas algorytmy są silnikiem napędzającym gospodarkę.

W większości książek o analityce, uczeniu maszynowym i danetyce domyślnie zakłada się, że problemy, przed którymi stoją inżynierowie i marketingowcy, można rozwiązać za pomocą takich samych metod i narzędzi. Oczywiście zmienne mają różne nazwy, a poza tym wymagana jest pewna wiedza specyficzna dla danej dziedziny, jednak algorytm grupowania k-średnich jest zawsze taki sam niezależnie od tego, czy grupuje się dane dotyczące turbin, czy też wpisy w mediach społecznościowych. Firmy wykorzystując w taki bezkrytyczny sposób narzędzia do uczenia maszynowego często były w stanie dokładnie przewidywać zachowania, ale odbywało się to kosztem głębszego i lepszego zrozumienia zachodzących procesów. Powodowało to, że modele stosowane w danetyce były krytykowane i uważane za „czarne skrzynki".

W niniejszej książce nie zostaną zaprezentowane metody pozwalające na uzyskanie dokładnych, ale nieprzejrzystych prognoz, ale zamiast tego można będzie znaleźć odpowiedź na pytanie „Co kieruje ludzkim zachowaniem?". Jeśli zostanie podjęta decyzja o wysłaniu maili do potencjalnych klientów, czy *w wyniku tego* zasubskrybują oni usługę? Do jakich grup klientów powinny być skierowane te maile? Czy starsi klienci wolą kupować różne produkty, *ponieważ* po prostu są starsi? W jaki sposób doświadczenie klienta wpływa na poziom jego lojalności i przywiązanie do firmy? Jeśli zamiast przewidywania zachowań

zrozumiemy je i będziemy potrafić określać ich przyczyny, przełamiemy klątwę „korelacja to nie przyczynowość", która sprawia, że kolejne pokolenia analityków nie są pewne wyników wygenerowanych przez ich modele.

Ta zmiana perspektywy nie wynika z wprowadzenia nowych narzędzi analitycznych. Do analizy danych zostaną użyte tylko dwie metody – stara, dobra regresja liniowa i jej logistyczny odpowiednik. Wspomniane rozwiązania są z natury bardziej czytelne niż pozostałe typy modeli. Oczywiście często odbywa się to kosztem mniejszej dokładności predykcyjnej (oznacza to, że modele popełniają większe błędy podczas prognozowania), ale nie ma to znaczenia w przypadku oceniania relacji zachodzących między zmiennymi.

W książce poświęcono dużo miejsca na wyjaśnienie, w jaki sposób należy rozumieć dane. Będąc rekruterem zajmującym się danetyką miałem do czynienia z wieloma kandydatami, którzy potrafili korzystać z zaawansowanych algorytmów uczenia maszynowego, ale nie rozwinęli w sobie odpowiedniego podejścia do danych – posiadali niewielką intuicję związaną z tym, co się w nich działo, a poza tym wykorzystywali przede wszystkim informacje zwracane przez algorytmy.

Wierzę, że poprzez przyjęcie następujących zasad można rozwinąć tę intuicję, a przy okazji – często radykalnie – poprawić poziom i wyniki projektów analitycznych:

- Wykorzystywać sposób myślenia związanego z naukami behawioralnymi, który pozwala na traktowanie danych nie jako celu samego w sobie, ale wpływa na zrozumienie psychologii i sposobu zachowania ludzi.
- Używać zestawu narzędzi do analizy przyczynowej, który pozwala bez wahania stwierdzić, że dana przyczyna powoduje określony skutek i określić, jak silny jest związek między nimi.

Chociaż zastosowanie w praktyce tylko jednej z powyższych zasad może od razu zapewnić ogromne korzyści, wydaje się, że uzupełniają się one w sposób naturalny, dlatego najlepiej je stosować razem. Biorąc pod uwagę, że sformułowanie „zmiana sposobu myślenia związanego z naukami behawioralnymi za pomocą zestawu narzędzi do analizy przyczynowej" jest trochę kiepskie, można przyjąć określenie „podejście lub koncepcja przyczynowo-behawioralna". Taka metoda ma dodatkową zaletę – znajduje zastosowanie zarówno w przypadku danych eksperymentalnych, jak i historycznych poprzez wykorzystywanie różnic między nimi. Kontrastuje to z tradycyjną metodą, która stosuje zupełnie inne narzędzia (na przykład analizę wariancji i test t-Studenta dla danych eksperymentalnych), a także z danetyką, która traktuje dane eksperymentalne podobnie, jak dane historyczne.

Kto może skorzystać z tej książki?

Ta książka przyda się Czytelnikom, którzy analizują dane biznesowe za pomocą języków R lub Python. Słowo „biznesowy" zostało tu użyte luźno, dlatego może ono oznaczać każdą organizację nastawioną na zysk, organizację non-profit lub organizację rządową, dla których ważne są prawidłowe spostrzeżenia i praktyczne wnioski prowadzące do określonych działań.

Jeśli chodzi o podstawy matematyki i statystyki, nie ma znaczenia, czy Czytelnik jest analitykiem biznesowym tworzącym prognozy miesięczne, badaczem UX analizującym zachowania klienta po kliknięciu jakiejś opcji, czy też naukowcem zajmującym się danetyką i projektującym modele uczenia maszynowego. Ta książka wymaga spełnienia jednego, podstawowego warunku wstępnego – należy znać przynajmniej podstawy regresji liniowej i logistycznej. Jeśli rozumie się działanie regresji, można podążać za treścią tej książki i czerpać z niej wielkie korzyści. Z drugiej strony wydaje się, że nawet eksperci zajmujący się danetyką, z doktoratem ze statystyki lub informatyki, mogą uznać materiał zaprezentowany w książce za nowy i użyteczny pod warunkiem, że nie są już specjalistami w dziedzinie analityki behawioralnej lub przyczynowej.

Jeśli chodzi o kwestię programowania, Czytelnik powinien umieć czytać i tworzyć kod w języku R lub Python, a najlepiej w obu. W książce nie zostanie wyjaśnione, jak można zdefiniować funkcję ani jak obsługiwać struktury danych takie jak typ `DataFrame` czy bibliotekę *pandas*. Istnieją doskonałe książki, z których należy skorzystać (na przykład *Python for Data Analysis* autorstwa Wesa McKinney'a (wydawnictwo O'Reilly) i *R for Data Science* autorstwa Garretta Grolemunda i Hadley'a Wickhama (również wydawnictwo O'Reilly)). Jeśli Czytelnik przeczytał którąkolwiek z nich, ukończył kurs dla początkujących programistów lub używał w pracy co najmniej jednego z dwóch wymienionych języków programowania, może bez problemu studiować tę książkę. Również nie zostanie zaprezentowany i przeanalizowany kod używany w celu uzyskania różnych wykresów, jednak jest on dostępny na portalu GitHub.

Dla kogo nie jest przeznaczona ta książka?

Jeśli Czytelnik pracuje w środowisku akademickim lub zajmuje się dziedziną, która wymaga przestrzegania norm akademickich (na przykład przeprowadza badania farmaceutyczne), ta książka może nadal być dla niego przydatna. Gdy jednak wykorzysta przedstawione w niej metody, może się okazać, że będzie mieć kłopoty ze swoim doradcą/redaktorem/szefem.

W tej książce *nie* zaprezentowano konwencjonalnych metod analizy danych behawioralnych, takich jak test t-Studenta lub analiza wariancji. Nie spotkałem się jeszcze z sytuacją, w której zastosowanie regresji byłoby mniej skuteczne od tych metod udzielania odpowiedzi na zapytanie biznesowe, dlatego celowo używałem w książce jedynie regresji liniowej i logistycznej. Jeśli Czytelnik chce jednak poznać inne metody, musi zdobyć wiedzę gdzie indziej (na przykład przeczytać książkę *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (wydawnictwo O'Reilly) Auréliena Gérona, w której zaprezentowano algorytmy uczenia maszynowego).

Zrozumienie i zmiana zachowań w określonych warunkach wymaga zarówno przeanalizowania danych, jak i podejścia jakościowego. W tej książce skoncentrowano się na pierwszym wymaganiu – przede wszystkim ze względu na dostępność miejsca. Ponadto istnieją już książki, w których doskonale wyjaśniono drugie wymaganie, takie jak *Nudge: Improving Decisions About Health, Wealth, and Happiness* (wydawnictwo

Penguin) Richarda Thalera i Cass Sunsteina oraz *Designing for Behavior Change: Applying Psychology and Behavioral Economics* (wydawnictwo O'Reilly) Stephena Wendla. Niemniej jednak zostanie zaprezentowane wprowadzenie wyjaśniające podstawowe koncepcje nauk behawioralnych, dzięki czemu nawet niedoświadczony Czytelnik będzie mógł zastosować narzędzia z tej książki.

Jeśli Czytelnik podczas analizy danych nie wykorzystywał wcześniej języka R ani Pythona, ta książka nie jest przeznaczona dla niego. Zamiast tego powinien się zainteresować doskonałymi pozycjami książkowymi przeznaczonymi dla początkujących, takimi jak wymienione w tym podrozdziale.

Kod języków R i Python

Dlaczego R *i* Python? Dlaczego nie wybrano tego, który jest lepszy? Dyskusja „R kontra Python" wciąż budzi emocje i zainteresowanie w internecie. Moim skromnym zdaniem jest ona jednak raczej nieistotna. W rzeczywistości trzeba po prostu używać języka wykorzystywanego w danej organizacji. Kiedyś pracowałem w firmie z branży medycznej, w której ze względów historycznych i regulacyjnych dominującym językiem był SAS. Regularnie używałem też języków R i Python do własnych analiz, ale ponieważ nie mogłem uniknąć wykorzystywania starszego kodu SAS, poznałem go na tyle, na ile potrzebowałem podczas mojego pierwszego miesiąca pracy. Jeśli Czytelnik nie spędzi całej swojej kariery zawodowej w firmie, która nie używa R ani Pythona, najprawdopodobniej i tak sam nauczy się podstaw obu tych języków, więc równie dobrze można założyć, że powinien je znać. Nie spotkałem jeszcze nikogo, kto stwierdziłby, że „nauka czytania kodu [jakiegoś języka] była stratą czasu".

Jeśli Czytelnik ma szczęście pracować w organizacji, która używa obu z nich, być może zadaje sobie teraz pytanie, jaki język powinien przede wszystkim wykorzystywać? Myślę, że to naprawdę zależy od kontekstu i zadań, które ma do wykonania. Osobiście wolę przeprowadzać eksploracyjną analizę danych (EDA) w R, ale uważam, że Python jest znacznie łatwiejszy podczas wydobywania danych ze stron internetowych (ang. *web scraping*). Radzę wybierać język w zależności od specyfiki pracy i opierać się na bieżących informacjach – oba języki są stale ulepszane, a to, co było prawdą w przypadku poprzedniej wersji R lub Pythona, może już być przestarzałe w najnowszej. Na przykład Python staje się coraz bardziej przyjaznym środowiskiem dla eksploracyjnej analizy danych. Energię lepiej spożytkować na naukę obu języków niż na przeszukiwanie forów w celu podjęcia decyzji, który z nich wybrać.

Środowisko programistyczne

Na początku każdego z rozdziałów zostanie określone, jakie pakiety języków R i Python należy wczytać, by wykonać zaprezentowane przykłady. Ponadto w całej książce będzie używanych kilka standardowych pakietów. Aby uniknąć powtórzeń, odpowiedni kod zostanie zaprezentowany jedynie poniżej (wywołania znajdują się we wszystkich skryptach

dostępnych na GitHubie). Podane pakiety powinny być zawsze wczytywane na samym początku programu. Należy również ustawić kilka parametrów:

```
## R
library(tidyverse)
library(boot) # Symulacje bootstrapowe
library(rstudioapi) # Wczytywanie danych z lokalnego folderu
library(ggpubr) # Generowanie wykresów wielokrotnych

# Ustawienie ziarna generatora losowego zapewni powtarzalność liczb losowych
set.seed(1234)
# Osobiście uważam, że domyślny zapis za pomocą notacji naukowej
# (czyli z użyciem wykładnika)
# jest mniej czytelny w przypadku wyników, więc go nie używam

## 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 # Wykresy
import seaborn as sns # Wykresy
```

Konwencje używane w kodzie

W niniejszej książce użyto języka R dostępnego w środowisku RStudio. Już podczas jej pisania pojawiła się wersja 4.0 języka R, więc wprowadzono odpowiednie zmiany, aby wydanie było jak najbardziej aktualne. Kod R wykorzystuje czcionkę stosowaną do prezentowania programów oraz komentarz wskazujący używany język, na przykład:

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

Jeśli chodzi o drugi język, wykorzystano Pythona ze środowiska Spyder dostępnego w dystrybucji Anacondy. Mam nadzieję, że dyskusja „Python 2.0 kontra 3.0" jest już przeszłością (przynajmniej w przypadku nowego kodu – starsze programy to zupełnie inna historia). W książce użyto Pythona 3.7. Konwencja dla kodu Pythona jest dość podobna do tej, którą zaprezentowano powyżej w przypadku R:

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

W książce często będą się pojawiać wyniki regresji. Mogą one zajmować dużo miejsca i zawierać wiele wierszy z informacjami diagnostycznymi, które nie są związane

z poruszanymi zagadnieniami. W przypadku realnych projektów nie należy ich jednak lekceważyć, ale o tym można się dokładniej dowiedzieć z innych książek. Dane wyjściowe zostaną więc uproszczone w następujący sposób:

```
## 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
...
```

Podstawy stylu programowania funkcyjnego

Jednym z etapów przejścia od poziomu początkującego do średniozaawansowanego programisty jest zaprzestanie pisania skryptów, w których kod jest jednym długim ciągiem instrukcji, a zamiast tego opakowanie go w funkcje. W różnych rozdziałach tej książki będziemy tworzyć i ponownie wykorzystywać funkcje takie jak poniższe, aby utworzyć bootstrapowe przedziały ufności:

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

  boot_vec <- sapply(1:B, function(x){
    cat("bootstrap iteration ", 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 = []

# Oblicz współczynnik zainteresowania dla każdej symulacji
for b in range(B):
    print("beginning iteration number " + str(b) + "\n")
    boot_df = dat_df.groupby("rep_ID").sample(n=1200, replace=True)
    coeff = metric_fun(boot_df)
    coeff_boot.append(coeff)

# Wyodrębnij przedział ufności
coeff_boot.sort()
offset = round(B * (1 - conf_level) / 2)
CI = [coeff_boot[offset], coeff_boot[-(offset+1)]]

return CI
```

Funkcje mają również tę dodatkową zaletę, że ograniczają efekty uboczne braku ich zrozumienia – nawet jeśli nie wie się dokładnie, jak działają, nadal można przyjąć za pewnik, że zwracają one przedziały ufności i realizują właściwy algorytm, dzięki czemu dokładniejszą analizę kodu można przeprowadzić w odpowiedniejszym czasie.

Użycie przykładów kodu

Dodatkowe materiały (przykłady kodów itd.) można pobrać ze strony *https:// oreil.ly/ BehavioralDataAnalysis*.

W przypadku pytań technicznych lub problemów z wykorzystaniem przykładów kodów należy przesłać maila na adres *bookquestions@oreilly.com*.

Celem tej książki jest wsparcie wykonywanej pracy. Ogólnie rzecz ujmując, przykładowe kody dołączone do książki można używać w programach i dokumentacji. Nie trzeba się kontaktować z wydawnictwem w celu uzyskania pozwolenia, chyba że jest wykorzystywana znaczna część kodu. Na przykład napisanie programu, który używa kilka fragmentów kodu z tej książki, nie wymaga pozwolenia. Sprzedaż lub dystrybucja przykładów z książek wydawnictwa O'Reilly wymaga już pozwolenia. Odpowiedź na pytanie poprzez zacytowanie tej książki i wykorzystanie przykładowego kodu nie wymaga pozwolenia. Włączenie znacznej ilości przykładowego kodu z tej książki do dokumentacji produktu wymaga pozwolenia.

Doceniamy uznanie autorstwa, ale nie wymagamy go. Zazwyczaj obejmuje ono tytuł, autora, wydawcę i numer ISBN. Na przykład: „*Behavioral Data Analysis with R and Python*, autor: Florent Buisson (O'Reilly). Copyright 2021 Florent Buisson, 978-1-492-06137-3".

W przypadku, gdy wykorzystanie przykładów kodu może wykraczać poza dozwolony użytek lub powyżej podane zezwolenie, należy się skontaktować z wydawnictwem poprzez wysłanie maila na adres *permissions@oreilly.com*.

Korzystanie z tej książki

Podstawowa zasada, na której opiera się ta książka, polega na tym, że efektywna analiza danych wymaga ciągłego wykorzystywania trzech elementów:

- faktycznych zachowań w świecie rzeczywistym i związanych z nimi zjawisk psychologicznych, takich jak intencje, myśli i emocje,
- analizy przyczynowej, a zwłaszcza diagramów przyczynowych,
- danych.

Książka została podzielona na pięć części:

Część I. Zrozumienie zachowań.
: W tej części zostaje wyjaśniona koncepcja przyczynowo-behawioralna, a także powiązania między zachowaniami, rozumowaniem przyczynowym i danymi.

Część II. Diagramy przyczynowe i usuwanie czynników zakłócających.
: Ta część wprowadza pojęcie czynników zakłócających i wyjaśnia, w jaki sposób diagramy przyczynowe umożliwiają usuwanie tych czynników z procesu analizy danych.

Część III. Profesjonalna analiza danych.
: W tej części zostają zaprezentowane narzędzia stosowane w przypadku brakujących danych. Zostaje także objaśniona symulacja bootstrapowa, ponieważ pozostała część książki będzie w dużej mierze wykorzystywać bootstrapowe przedziały ufności.

  Dane nieliczne czy niekompletne, a także takie, które mają nieregularny kształt (na przykład zawierają wiele wartości szczytowych lub odstających), nie są czymś nowym, ale mogą być szczególnie uciążliwe w przypadku analizy behawioralnej.

Część IV. Projektowanie i analizowanie eksperymentów.
: W tej części zostanie przedstawione, jak należy projektować i analizować eksperymenty.

Część V. Użycie zaawansowanych narzędzi w analizie danych behawioralnych.
: Na koniec zostanie wykorzystana cała zdobyta wiedza, aby przeanalizować moderację, mediację i zmienne instrumentalne.

Poszczególne części książki są częściowo zależne od siebie, dlatego należy je przeczytać w naturalnej kolejności – przynajmniej za pierwszym razem.

Konwencje użyte w tej książce

W tej książce zostały użyte następujące konwencje typograficzne:

Kursywa
: Oznacza nowe terminy, adresy URL, adresy mailowe, nazwy plików i ich rozszerzenia.

`Czcionka o stałej szerokości`
: Używana w listingach, a także w akapitach w celu odwoływania się do elementów programu, takich jak nazwy zmiennych lub funkcji, bazy danych, typy danych, zmienne środowiskowe, instrukcje i słowa kluczowe.

`Pogrubiona czcionka o stałej szerokości`
: Stosowana do wyróżniania poleceń lub innych tekstów, które są wprowadzane przez użytkownika.

`Pochyła czcionka o stałej szerokości`
: Używana w komentarzach, a także w przypadku tekstu, który powinien zostać zastąpiony wartościami podanymi przez użytkownika lub wynikającymi z kontekstu.

W ten sposób są oznaczone sztuczki i wskazówki.

W ten sposób są oznaczone ogólne uwagi.

W ten sposób są oznaczone ostrzeżenia.

Kontakt z wydawnictwem

Komentarze i pytania dotyczące tej książki należy kierować do wydawcy:

O'Reilly Media, Inc.

1005 Gravenstein Highway North

Sebastopol, CA 95472

800-998-9938 (ze Stanów Zjednoczonych lub Kanady)

707-829-0515 (rozmowa międzynarodowa lub lokalna)

707-829-0104 (faks)

Strona internetowa książki z erratą, przykładami i dodatkowymi informacjami jest dostępna pod adresem *https://oreil.ly/Behavioral_Data_Analysis_with_R_and_Python*.

W celu skomentowania książki lub zadania pytania technicznego należy wysłać maila na adres *bookquestions@oreilly.com*.

Najnowsze wiadomości i informacje o książkach oraz szkoleniach można znaleźć na stronie *http://oreilly.com*.

Jesteśmy na Facebooku: *http://facebook.com/oreilly*.

Można nas także śledzić na Twitterze: *http://twitter.com/oreillymedia*.

Nasz kanał na YouTubie: *http://youtube.com/oreillymedia*.

Podziękowania

Autorzy książek często dziękują swoim współmałżonkom za cierpliwość i wspominają szczególnie pedantycznych recenzentów. Mam to szczęście, że te dwie osoby stanowią w moim przypadku jedną. Nie sądzę, żeby ktokolwiek inny odważył się tak często zwracać mi uwagę. Dzięki temu ta książka stała się jednak znacznie lepsza. Pierwsze podziękowania kieruję więc do mojej żony dzielącej ze mną życie i myśli.

Kilku moich kolegów i współpracowników naukowych było na tyle uprzejmych, że poświęciło cenny czas na przeczytanie i skomentowanie początkowej wersji książki, co spowodowało, że stała się ona doskonalsza. Moje podziękowania otrzymują następujące osoby (w odwrotnej kolejności alfabetycznej): Jean Utke, Jessica Jakubowski, Chinmaya Gupta i Phaedra Daipha!

Specjalne podziękowania przekazuję Bethany Winkel za wsparcie podczas pisania książki.

Z zażenowaniem zdaję sobie sprawę z tego, jak niedopracowane i niezrozumiałe były pierwsze wersje robocze. Redaktor prowadzący i recenzenci techniczni dzielili się ze mną swoją bogatą wiedzą oraz doświadczeniem, a także cierpliwie mnie wspierali, dzięki czemu byłem w stanie ukończyć tę książkę. Moje wyrazy wdzięczności otrzymują Gary O'Brien, Xuan Yin, Shannon White, Jason Stanley, Matt LeMay i Andreas Kaltenbrunner.

CZĘŚĆ I

Zrozumienie zachowań

W pierwszej części książki wyjaśnimy, dlaczego analiza danych behawioralnych wymaga zastosowania nowej metodologii.

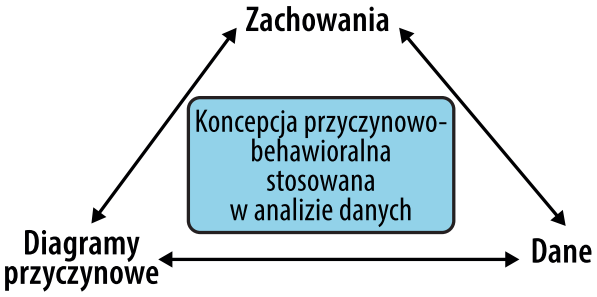
Ta nowa metoda zostanie przeanalizowana w rozdziale 1, „Koncepcja przyczynowo-behawioralna stosowana w analizie danych". W związku z tym zaprezentujemy przykład pokazujący, w jaki sposób nawet najprostsza analiza danych może zostać zniekształcona przez obecność czynnika zakłócającego. Przy zastosowaniu tradycyjnego podejścia rozwiązanie tego problemu jest w najlepszym razie skomplikowane, a w najgorszym niemożliwe. Wykorzystanie nowej metodologii umożliwia szybkie osiągnięcie sukcesu.

W rozdziale 2 dokładniej wyjaśnimy specyfikę danych behawioralnych. Przy okazji Czytelnik pozna podstawy nauk behawioralnych oraz proces, który ma zapewniać, by dane odzwierciedlały odpowiadające im zachowania z życia codziennego.

Rozdział 1

Koncepcja przyczynowo-behawioralna stosowana w analizie danych

We wstępie do książki zostało już wyjaśnione, że zrozumienie przyczyn, które powodują zmianę zachowań, jest jednym z kluczowych celów analizy stosowanej, zarówno w firmie, organizacji non-profit, jak i organizacji publicznej. Chcemy się dowiedzieć, dlaczego ktoś kupił jakiś produkt, jak również, dlaczego ktoś inny tego *nie* zrobił. Chcemy zrozumieć, dlaczego ktoś odnowił subskrypcję, skontaktował się z infolinią, zamiast wykonać opłatę przez internet, zarejestrował się jako dawca narządów lub przekazał datek organizacji non-profit. Posiadanie tej wiedzy pozwala przewidzieć, jak się zachowają ludzie w różnych sytuacjach i pomaga określić, co organizacja może zrobić, aby zachęcić (lub zniechęcić) ich do powtórzenia działań. Wydaje się, że najlepiej można osiągnąć ten cel łącząc analizę danych z odpowiednim sposobem wykorzystania nauk behawioralnych i zestawem narzędzi służących do analizy przyczynowej, aby utworzyć zintegrowaną metodę, która została nazwana „koncepcją przyczynowo-behawioralną". W tym kontekście *zachowania* znajdują się na samym szczycie, ponieważ ich zrozumienie jest celem ostatecznym. Rozwiązanie uzyskuje się za pomocą *diagramów przyczynowych* i *danych*, które tworzą dwa filary wspierające trójkąt (rysunek 1.1).



Rysunek 1.1 *Koncepcja przyczynowo-behawioralna stosowana w analizie danych*

W książce przeprowadzimy analizę każdego elementu trójkąta, by wyjaśnić, jak się łączą ze sobą. Zdobyta wiedza zostanie wykorzystana w ostatnim rozdziale. Za pomocą jednego wiersza kodu rozwiążemy zadanie, które byłoby bardzo zniechęcające przy wykorzystaniu

tradycyjnego podejścia – zmierzymy wpływ satysfakcji klienta na poziom jego przyszłych wydatków. Oprócz rozwiązywania ciekawych problemów, nowa koncepcja umożliwi także skuteczniejsze przeprowadzanie ogólnych analiz, takich jak określanie wpływu kampanii mailowej lub cech produktu na zachowania związane z zakupami.

Czytelnicy mający doświadczenie w analizie predykcyjnej mogą się zastanawiać, dlaczego została wybrana analiza przyczynowa. Oto odpowiedź: mimo że analiza predykcyjna jest (i pozostanie) bardzo skuteczna w warunkach biznesowych, może się nie sprawdzić, gdy w grę wchodzą ludzkie zachowania. W szczególności przyjęcie podejścia przyczynowego może pomóc w zidentyfikowaniu i usunięciu „zakłócenia", czyli bardzo powszechnego problemu występującego w danych behawioralnych. To zagadnienie zostanie omówione w dalszej części pierwszego rozdziału.

Dlaczego do wyjaśniania ludzkich zachowań należy zastosować analizę przyczynową?

Zrozumienie roli analizy przyczynowej pomoże w wyjaśnieniu, dlaczego należy z niej korzystać w przypadku zagadnień biznesowych. Jak zobaczymy, ta konieczność wynika ze złożoności ludzkich zachowań.

Różne rodzaje analizy

Istnieją trzy różne rodzaje analiz: opisowa, predykcyjna i przyczynowa. Analiza opisowa pozwala na uzyskanie *opisu* danych. Mówiąc prościej, opisuje ona to, „co jest" lub „co zostało zmierzone". Do tego typu analiz można zaliczyć sprawozdawczość biznesową. Ilu klientów zrezygnowało z subskrypcji w zeszłym miesiącu? Ile zarobiono w zeszłym roku? Ilekroć obliczamy średnią lub inne, proste wskaźniki, pośrednio korzystamy z analiz opisowych. Analiza opisowa jest najprostszą formą analizy. Okazuje się, że jej rola jest niedoceniana. Wiele organizacji ma trudności z uzyskaniem jasnego i jednolitego obrazu swojej działalności. Aby się dowiedzieć, jak bardzo ten problem występuje w danej firmie, wystarczy po prostu zadać to samo pytanie działowi finansowemu i operacyjnemu, a następnie porównać odpowiedzi[1].

Analiza predykcyjna umożliwia *prognozowanie*. Opisuje ona coś, „co będzie, zakładając, że obecne warunki nie ulegną zmianie", lub to, „czego jeszcze nie zmierzono". Większość metod uczenia maszynowego (na przykład sieci neuronowe i modele wzmacniania gradientu) wykorzystuje ten typ analizy i pomaga odpowiadać na pytania typu „Ilu klientów zrezygnuje z subskrypcji w przyszłym miesiącu?" oraz „Czy to zamówienie jest fałszywe?" W ciągu ostatnich kilku dekad analiza predykcyjna zmieniła świat – niezliczone rzesze analityków zatrudnionych w biznesie są świadectwem tego sukcesu.

1 Oczywiście niektóre odpowiedzi powinny się różnić od siebie, ponieważ dane są wykorzystywane do różnych celów i używane na różne sposoby. Nawet jednak proste pytania, na które można by się spodziewać jednej, prawdziwej odpowiedzi (na przykład: „Ilu pracowników jest obecnie zatrudnionych w firmie?"), generalnie spowodują pojawienie się różnych odpowiedzi.

Trzecim rodzajem jest analiza przyczynowa, która umożliwia poznanie *przyczyn* pojawienia się określonych danych. Odpowiada ona na pytania „A co, jeśli?" czy też „Co się stanie, jeśli warunki ulegną zmianie?". Na przykład: „Ilu klientów zrezygnuje z subskrypcji w przyszłym miesiącu, *jeśli nie podarujemy im kuponów*?". Najbardziej znanym narzędziem analizy przyczynowej jest test A/B, czyli eksperyment randomizowany lub inaczej mówiąc, randomizowane badanie kontrolowane (*randomized controlled trial* – RCT). Wynika to stąd, że najprostszym i najskuteczniejszym sposobem uzyskania odpowiedzi na powyższe pytanie jest wysłanie kuponów do losowo wybranej grupy klientów, a następnie sprawdzenie, ilu z nich zrezygnuje z subskrypcji w porównaniu z grupą kontrolną.

Eksperymenty zostaną przeanalizowane w IV części książki, ale wcześniej, w części II, przyjrzymy się innemu narzędziu, a mianowicie diagramom przyczynowym, z których można korzystać nawet wtedy, gdy przeprowadzanie eksperymentów jest niemożliwe. Jednym z celów tej książki jest skłonienie Czytelnika do spojrzenia na analizę przyczynową z szerszej perspektywy, by nie utożsamiał jej jedynie z eksperymentowaniem.

Chociaż podział przedstawiony powyżej może sprawiać wrażenie, że mamy do czynienia z dobrze zorganizowaną kategoryzacją, w rzeczywistości istnieją dość duże powiązania między tymi trzema rodzajami analiz, a te same pytania i metody zaczynają być stosowane w każdej z nich. Można również spotkać inne terminy, takie jak *analiza preskryptywna*, które dodatkowo zacierają granice i dodają kolejne odcienie szarości bez radykalnej zmiany ogólnego obrazu.

Istoty ludzkie są skomplikowane

Jeśli analiza predykcyjna jest rzeczywiście skuteczna, a analiza przyczynowa wykorzystuje jej narzędzia, takie jak regresja, dlaczego nie powinno się pozostać przy analizie predykcyjnej? Odpowiedź jest prosta: ponieważ ludzie są bardziej skomplikowani, niż turbiny wiatrowe. Ludzkie zachowanie:

Ma wiele przyczyn
: Na zachowanie turbiny nie ma wpływu jej osobowość, normy społeczne wymagane w społeczności turbin ani okoliczności jej wychowania. Z drugiej strony zdolność prognozowania ludzkiego zachowania przy użyciu jakiejkolwiek pojedynczej zmiennej jest z powodu tych czynników prawie zawsze poniżej oczekiwań.

Jest zależne od kontekstu
: Drobne lub kosmetyczne zmiany w środowisku, takie jak zmiana domyślnej opcji wyboru, mogą mieć duży wpływ na zachowanie. Jest to dobrodziejstwem z punktu widzenia *sterowania* behawioralnego, ponieważ pozwala zmieniać zachowania. Jednak z perspektywy *analizy* behawioralnej jest przekleństwem, ponieważ oznacza, że każda sytuacja jest wyjątkowa i trudno przewidzieć jej skutki.

Jest zmienne (naukowcy powiedzieliby, że niedeterministyczne)

Ta sama osoba w identycznej sytuacji może się zachowywać za każdym razem zupełnie inaczej, nawet po uwzględnieniu wpływu zewnętrznych przyczyn. Może to być spowodowane czynnikami przejściowymi, takimi jak nastroje, lub długotrwałymi, takimi jak znudzenie się codziennym spożywaniem tego samego śniadania. Te czynniki mogą radykalnie zmieniać zachowanie, a poza tym trudno je wykrywać.

Jest innowacyjne

Gdy warunki w środowisku ulegają zmianie, człowiek może się zacząć zachowywać w zupełnie odmienny sposób, którego nigdy wcześniej nie stosował. Dzieje się tak nawet w najbardziej prozaicznych okolicznościach. Na przykład jeśli pracownik dojeżdża do firmy samochodem, a na trasie pojawił się wypadek, podejmuje nagłą decyzję, by skręcić w prawo i pojechać inną drogą.

Jest strategiczne

Ludzie wyciągają wnioski z zachowań innych osób i reagują na nie. W niektórych przypadkach może to oznaczać „wznowienie" współpracy, która została przerwana przez okoliczności zewnętrzne, co czyni ją bardziej przewidywalną. Ale w innych sytuacjach może to obejmować dobrowolne ukrywanie zachowania, aby uczynić je nieprzewidywalnym podczas gry opartej na współzawodnictwie, takiej jak szachy (lub w celu oszustwa!).

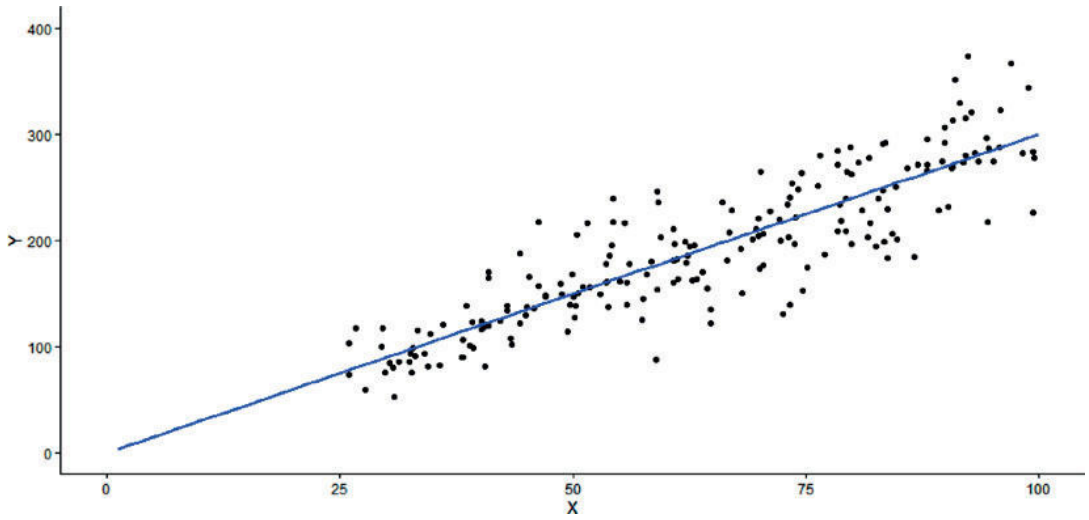
Wszystkie te aspekty ludzkiego zachowania sprawiają, że jest ono mniej przewidywalne, niż zachowanie obiektów fizycznych. Aby znaleźć prawidłowości, które mogą się przydać podczas analizy, należy zejść o jeden poziom niżej w celu zrozumienia i zmierzenia przyczyn zachowania. To, że ktoś w poniedziałek jadł na śniadanie płatki owsiane i wybrał określoną trasę, nie oznacza, że zrobi dokładnie to samo we wtorek. Można być jednak całkiem pewnym, że również zje *jakieś* śniadanie i wybierze *jakąś* trasę do pracy.

Ekstrapolacja w analizie, przekleństwo wymiarowości i krytyka Lucasa

Czytelnicy z wykształceniem ekonomicznym mogą nie być w pełni usatysfakcjonowani wcześniejszym stwierdzeniem, że „ludzkie zachowanie jest trudne do przewidzenia, ponieważ jest skomplikowane". Oto więc matematyczny dowód na poprawność argumentacji. Najpierw opiszemy różnicę między interpolacją a ekstrapolacją. Rysunek 1.2 przedstawia symulowane dane, przy czym między dwiema zmiennymi zachodzi liniowa zależność.

Linia widoczna na rysunku jest najlepiej dopasowaną linią regresji. Odpowiada ona funkcji regresji liniowej dla dwóch zmiennych i ma nachylenie równe w przybliżeniu 3. Możemy jej użyć do prognozowania nieznanych wartości Y na podstawie znanej wartości X (i odwrotnie). Na przykład dysponując wartością $X = 50$ moglibyśmy przewidzieć, że Y będzie równe 150. Na lewo od tej wartości znajdują się obserwowane punkty, dla których $X < 50$, a na prawo takie, dla których $X > 50$.

Taki proces predykcyjny nazywa się interpolacją, ponieważ wyznaczone miejsce znajduje się pomiędzy obserwowanymi punktami (przedrostek „inter" oznacza „pomiędzy", na przykład „internacjonalizm" = „współpraca pomiędzy narodami"). I odwrotnie, jeśli użyjemy linii regresji dla X = 0 do przewidzenia, że Y = 0, nazwiemy to ekstrapolacją, ponieważ miejsce, które próbujemy przewidzieć, znajduje się poza chmurą obserwowanych punktów (przedrostek „ekstra" oznacza „poza", na przykład „ekstraklasa" = „poza zwykłą klasą"). W statystyce i w życiu codziennym ekstrapolacja oznacza wyjście poza sferę świata obserwowanego i znanego w celu wyznaczenia prognozy. Interpolacja jest zwykle bezpieczna i niezawodna, natomiast ekstrapolacja zawsze nieco spekulacyjna – wymaga aktu wiary, by założyć, że reguły stosowane w pewnych granicach będą wciąż obowiązywać poza nimi.



Rysunek 1.2 *Zależność liniowa między dwiema zmiennymi, wraz z linią regresji*

Na obiekty fizyczne, takie jak turbina wiatrowa, wpływa stosunkowo niewielka i stała liczba czynników (nie jest tak, że niektóre prawa fizyki czasem nie działają lub nowe pojawiają się w sposób losowy). W wyniku tego ma się do czynienia z wieloma punktami danych w porównaniu z liczbą wymiarów przestrzeni problemowej, co oznacza, że prawie zawsze należy przeprowadzać interpolację. Dla uproszczenia w modelu można pomijać drugorzędne lub rzadkie zjawiska, takie jak burza pojawiająca się raz na sto lat. Nawet jednak w przypadku wystąpienia takich niezwykłych sytuacji wynik pozostaje dość przewidywalny – łopata wirnika może się oderwać i spaść na ziemię, ale nie odleci.

Z drugiej strony na ludzkie zachowanie wpływa wiele różnych czynników, które w danym dniu mogą być istotne (lub nie). Ich znaczenie może też z czasem wzrastać lub zanikać. Dlatego w takich przypadkach zwykle istnieje niewiele punktów danych w stosunku do liczby wymiarów przestrzeni problemowej, co oznacza, że znacznie częściej dokonuje się ekstrapolacji. Problem ten jest znany w statystyce jako „przekleństwo wymiarowości". Ponadto drobne zmiany w środowisku mogą prowadzić

do poważnych zmian w zachowaniu, co sprawia, że próba przewidzenia przyszłego zachowania człowieka na podstawie tego samego zachowania z przeszłości jest grą losową o niewielkiej szansie na wygraną.

Informacja dla Czytelników zainteresowanych genealogią ekonomii behawioralnej: makroekonomista Robert Lucas sformułował odpowiednie prawo w latach 70. XX wieku (tak zwaną „krytykę Lucasa" modeli keynesowskich). Zalecił w nim analizowanie bardziej zaawansowanych parametrów ludzkich zachowań, takich jak preferencje konsumentów, co dodatkowo wspiera zaprezentowaną wcześniej argumentację.

Zakłócenia, czyli ukryte niebezpieczeństwa rozwiązywania problemów za pomocą regresji

We wcześniejszym podrozdziale zostało wspomniane, że analiza przyczynowa często wykorzystuje te same narzędzia, co analiza predykcyjna. Ponieważ każda z tych metod kieruje się odmiennymi celami, narzędzia są używane w różny sposób. Regresja jest jednym z podstawowych narzędzi, więc może zostać skutecznie wykorzystana, by zilustrować różnicę między analizą predykcyjną a przyczynową. Regresja działająca poprawnie w przypadku analizy predykcyjnej często byłaby całkowitą porażką w analizie przyczynowej (i odwrotnie).

Regresja stosowana w analizie predykcyjnej służy do oszacowania nieznanej wartości (często, ale nie zawsze, występującej w przyszłości). Proces polega na wykorzystaniu istniejących informacji i różnych czynników, aby ustalić najlepszą, prognozowaną wartość dla danej zmiennej. Ważna jest sama przewidywana wartość i dokładność jej oszacowania, a nie sposób lub przyczyny, dla których wykonano prognozę.

Analiza przyczynowa również wykorzystuje regresję, ale nie koncentruje się na szacowaniu wartości zmiennej docelowej. Zamiast tego celem jest ustalenie przyczyny pojawienia się właśnie takiej wartości. Interesująca jest już nie sama zmienna zależna, ale jej związek z określoną zmienną niezależną. Przy prawidłowo skonstruowanej regresji współczynnik korelacji może być uproszczoną miarą przyczynowego wpływu zmiennej niezależnej na zmienną zależną.

Ale czym jest prawidłowo zdefiniowana regresja służąca temu celowi? Dlaczego nie można po prostu użyć regresji, którą wykorzystuje się w analizie predykcyjnej, a następnie potraktować dostarczone współczynniki jako miary związku przyczynowego? Nie jest to możliwe, ponieważ każda zmienna używana w regresji może modyfikować współczynniki innych zmiennych. Dlatego też odpowiednia metoda musi zostać zdefiniowana nie po to, aby tworzyć najdokładniejsze przewidywania, ale najdokładniejsze współczynniki. Te dwa zestawy zasadniczo różnią się od siebie, ponieważ dana zmienna może być mocno skorelowana ze zmienną docelową (a zatem być wysoce predykcyjna), a jednak w rzeczywistości nie wpływać na nią.

W tym podrozdziale wyjaśnimy, dlaczego różnica związana ze sposobem wykorzystania narzędzi ma znaczenie, a także dlaczego poprawny dobór zmiennych to więcej niż połowa sukcesu podczas wykonywania analizy behawioralnej. W tym celu wykorzystamy przykład wykorzystujący fikcyjną sieć supermarketów C-Mart, która zarządza sklepami na terenie całych Stanów Zjednoczonych. C-Mart, pierwsza z dwóch wymyślonych firm wykorzystywanych w tej książce, pomoże zrozumieć możliwości i wyzwania związane z analizą danych, stojące przed tradycyjnymi przedsiębiorstwami w epoce cyfrowej.

Dane

W folderze dostępnym na GitHubie (*https://oreil.ly/BehavioralDataAnalysisCh1*) znajdują się dwa pliki CSV (*chap1-stand_data.csv* i *chap1-survey_data.csv*) zawierające zbiory danych używane w dwóch przykładach z tego rozdziału.

W tabeli 1.1 przedstawiono informacje zawarte w pliku *chap1-stand_data.csv*, dotyczące dziennej sprzedaży lodów i kawy mrożonej na stoiskach sieci C-Mart.

Tabela 1.1 *Informacje o sprzedaży dostępne w pliku chap1-stand_data.csv*

Nazwa zmiennej	Opis zmiennej
IceCreamSales	Dzienna sprzedaż lodów na stoiskach sieci C-Mart
IcedCoffeeSales	Dzienna sprzedaż kawy mrożonej na stoiskach sieci C-Mart
SummerMonth	Zmienna binarna informująca, czy jest lato
Temperature	Średnia temperatura w danym dniu na określonym stoisku

W tabeli 1.2 zaprezentowano informacje zawarte w pliku *chap1-survey_data.csv*, dotyczące ankiety przeprowadzonej z przechodzącymi osobami poza obszarem stoisk C-Mart.

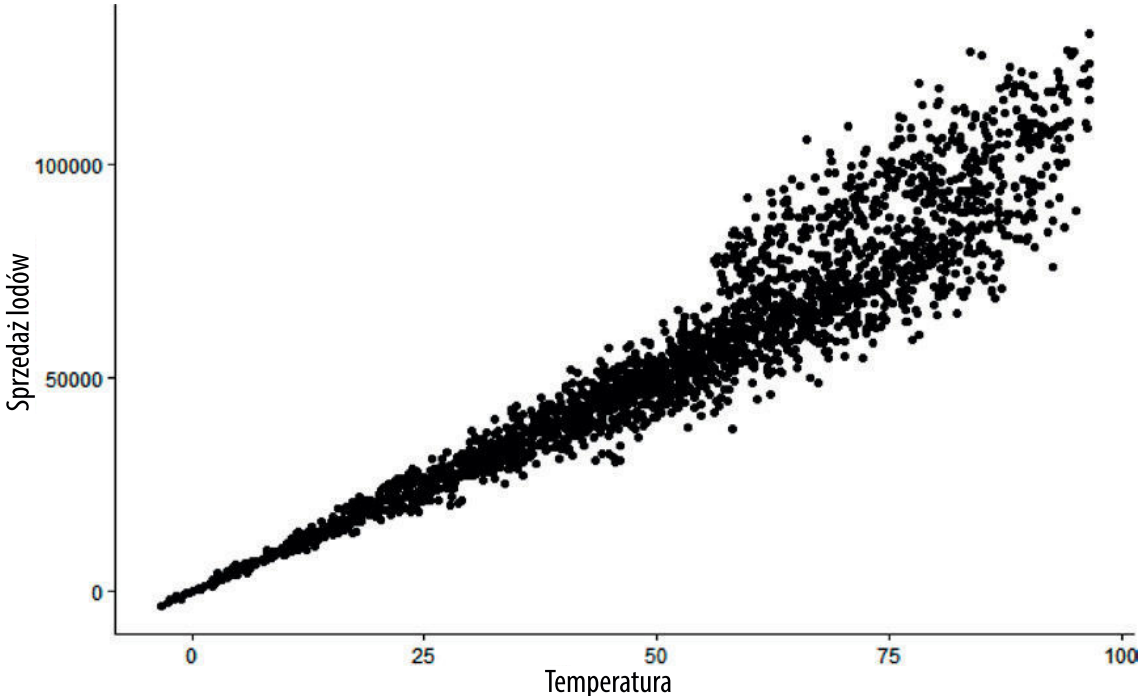
Tabela 1.2 *Ankieta z pliku chap1-survey_data.csv*

Nazwa zmiennej	Opis zmiennej
VanillaTaste	Ocena smaku lodów waniliowych (od 0 do 25)
ChocTaste	Ocena smaku lodów czekoladowych (od 0 do 25)
Shopped	Zmienna binarna oznaczająca, czy respondent kiedykolwiek robił zakupy na lokalnym stoisku C-Mart

Dlaczego korelacja nie jest związkiem przyczynowym? Rola czynnika zakłócającego

Firma C-Mart posiada w każdym sklepie stoisko z lodami. Jej zarząd wierzy, że pogoda wpływa na codzienną sprzedaż, czyli, używając żargonu przyczynowego, że jest ona przyczyną sprzedaży. Innymi słowy, gdy inne czynniki się nie zmieniają, można założyć, że ludzie chętniej kupują lody w cieplejsze dni, co ma intuicyjny sens. To przekonanie jest wspierane przez silną korelację danych historycznych między temperaturą a sprzedażą,

jak pokazano na rysunku 1.3 (odpowiednie dane i kod znajdują się w folderze książki na portalu GitHub).



Rysunek 1.3 *Sprzedaż lodów jako funkcja obserwowanej temperatury*

Jak wskazano we wstępie, regresja będzie używana w książce jako podstawowe narzędzie służące do analizy danych. Wykonanie regresji liniowej dla zależności sprzedaży lodów od temperatury wymaga użycia jednego wiersza kodu:

```
## Python (dane wyjściowe nie zostały zaprezentowane)
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 ***
...
```

Z perspektywy zagadnień omawianych w tej książce najbardziej istotnym fragmentem uzyskanego wyniku są wartości współczynników, które informują, że szacowany punkt przecięcia, czyli teoretyczna, średnia sprzedaż lodów dla temperatury równej zero wynosi −4519, co jest oczywiście ekstrapolacją niemającą sensu. Można się również dowiedzieć, że szacowany współczynnik dla temperatury wynosi 1145, co oznacza, że każdy dodatkowy stopień pozwoli zwiększyć sprzedaż lodów o 1145 dolarów.

A teraz wyobraźmy sobie, że zbliżamy się do końca wyjątkowo ciepłego tygodnia października, więc na podstawie przewidywań modelu firma z wyprzedzeniem zwiększyła zapasy w stoiskach z lodami. Jednak tygodniowa sprzedaż, choć wyższa niż zwykle w takim czasie, spadła znacznie poniżej ilości przewidywanej przez model. Oj, co się stało? Czy analityk danych powinien zostać zwolniony?

Stało się to, że model nie uwzględnia kluczowego faktu – tego, że większość sprzedaży lodów ma miejsce w miesiącach letnich, gdy dzieci nie chodzą do szkoły. Model regresji utworzył najlepszą prognozę na podstawie dostępnych danych, ale częściowy powód zwiększonej sprzedaży lodów (wakacje letnie dla uczniów) został częściowo błędnie przypisany temperaturze, ponieważ miesiące letnie są dodatnio skorelowane z temperaturą. Ponieważ wzrost temperatury w październiku nie spowodował przerwania zajęć szkolnych (przykro mi, dzieciaki!), odnotowano niższą sprzedaż, niż w przypadku letnich dni z tymi samymi temperaturami.

Z formalnego punktu widzenia czynnikiem zakłócającym relację między temperaturą a sprzedażą jest miesiąc. *Czynnik zakłócający* to zmienna, która wprowadza błąd do regresji. Gdy występuje on w analizowanej sytuacji, oznacza to, że interpretowanie współczynnika regresji jako przyczyny doprowadzi do niewłaściwych wniosków. Weźmy pod uwagę miejsce takie jak Chicago, które ma klimat kontynentalny – zima jest bardzo chłodna, a lato bardzo gorące. Podczas porównywania poziomu sprzedaży w przypadkowy, upalny dzień z poziomem sprzedaży w przypadkowy, chłodny dzień bez jednoczesnego uwzględnienia wpływu miesiąca najprawdopodobniej bierze się pod uwagę jakiś gorący, wakacyjny dzień oraz chłodny dzień zimowy, gdy dzieci są w szkole. Zawyża to pozorny związek między temperaturą a sprzedażą.

W przypadku tego przykładu można się również spodziewać konsekwentnego niedoszacowania sprzedaży w chłodniejsze dni. W rzeczywistości w miesiącach letnich następuje zmiana paradygmatu. Gdy zmiana ta zależy wyłącznie od temperatury w regresji liniowej, jej współczynnik jest zbyt wysoki dla wyższych temperatur i zbyt niski dla niższych.

Zbyt wiele zmiennych może zepsuć zabawę

Potencjalnym rozwiązaniem problemu z czynnikami zakłócającymi byłoby uwzględnienie w regresji wszystkich możliwych zmiennych. Sposób myślenia „weź wszystko, co wpadnie w ręce" wciąż ma zwolenników wśród statystyków. W książce *The Book of Why* Judea Pearl i Dana Mackenzie piszą, że „jeden z czołowych statystyków niedawno stwierdził, że «unikanie warunkowania dla niektórych obserwowalnych zmiennych towarzyszących (…) jest nienaukowym improwizowaniem»" (Pearl i Mackenzie, 2018, str. 160[2]). Jest to również zachowanie dość powszechnie spotykane wśród naukowców zajmujących się danymi. Prawdę mówiąc, jeśli celem jest tylko prognozowanie zmiennej, istnieje już model, który został starannie zaprojektowany, aby odpowiednio przewidywać wartości spoza zakresu danych testowych. Jeśli nie jesteśmy zainteresowani, dlaczego prognozowana zmienna przyjmuje określoną wartość, takie rozwiązanie jest zupełnie wystarczające. Niestety, nie

2 Tym wspomnianym statystykiem był Donald Rubin.

sprawdzi się ono, jeśli celem będzie zrozumienie związków przyczynowych. W takim przypadku dodanie do modelu jak największej liczby zmiennych będzie nie tylko nieefektywne, ale może spowodować uzyskanie wniosków przeciwnych do zamierzonych i być bardzo mylące.

Zademonstrujmy to na przykładzie poprzez dodanie zmiennej, którą można byłoby uwzględnić. Wpłynie ona na regresję. Została więc utworzona zmienna *IcedCoffeeSales*, która jest skorelowana z wartością *Temperature*, a nie *SummerMonth*. Przyjrzyjmy się, co się stanie z regresją, jeśli weźmiemy pod uwagę tę zmienną razem z pozostałymi, czyli *Temperature* i *SummerMonth*. *SummerMonth* to zmienna binarna równa 1 lub 0, która wskazuje, czy miesiącem był lipiec albo sierpień (1), czy też jakikolwiek inny (0)):

```
## R (dane wyjściowe nie zostały zaprezentowane)
> 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
...
```

Widać, że wartość współczynnika dla zmiennej *Temperature* zmieniła się dramatycznie w porównaniu z poprzednim przykładem i teraz jest ujemna. Wysokie wartości p dla zmiennych *Temperature* i *IcedCoffeeSales* byłyby zwykle traktowane jako oznaka problemu, ale ponieważ wartość p dla zmiennej *Temperature* jest „gorsza", analityk może wywnioskować, że powinien ją usunąć z regresji. Jak to jest możliwe?

Prawda kryjąca się za danymi (która jest znana, ponieważ dla tego przykładu specjalnie utworzono odpowiednie relacje i wygenerowano losowe dane) jest taka, że gdy robi się gorąco, ludzie chętniej kupują mrożoną kawę. W upalne dni chętniej kupuje się też więcej lodów. Jednak sam zakup kawy mrożonej nie sprawia, że klienci są mniej lub bardziej skłonni do zakupu lodów. Miesiące letnie również nie są skorelowane z zakupami kawy mrożonej, ponieważ dzieci w wieku szkolnym nie są istotnym czynnikiem popytu na nią (matematyczne wyjaśnienie jest dostępne w ramce na stronie sąsiedniej).

Na rysunku 1.4 przedstawiono pozytywną korelację między sprzedażą kawy mrożonej a sprzedażą lodów, ponieważ obie wartości rosną, gdy jest cieplej. Jednak każdy wzrost sprzedaży kawy mrożonej w miesiącach letnich można wyjaśnić wspólną korelacją ze zmienną odpowiadającą temperaturze. Gdy algorytm regresji próbuje wyjaśnić sprzedaż lodów za pomocą trzech dostępnych zmiennych, moc wyjaśniająca teorii wykorzystującej temperaturę, opisującej sprzedaż kawy mrożonej, zostaje uwzględniona w zmiennej temperaturowej. Z drugiej strony sprzedaż kawy mrożonej musi zostać skompensowana ze

Dokładniejsza analiza techniczna: co się zdarzyło?

Oto równanie dotyczące sprzedaży lodów, które zostało użyte do wygenerowania symulowanych danych:

$$IceCreamSales := 1\,000 \cdot Temperature + 20\,000 \cdot SummerMonth + \varepsilon_1$$

W tym równaniu ε_1 reprezentuje losowy szum o średniej zerowej, a znak „:=" wskazuje, że definiuje lub konstruuje się w nim zmienną występującą po lewej stronie, czyli *IceCreamSales*.

Jednak równanie, które służy do prognozowania w regresji liniowej, jest następujące:

$$IceCreamSales = \beta_T \cdot Temperature + \beta_S \cdot SummerMonth + \beta_C \cdot IcedCoffeeSales$$

Oto rzeczywiste równanie, które zostało użyte do wygenerowania wartości sprzedaży mrożonej kawy:

$$IcedCoffeeSales := 1\,000 \cdot Temperature + \varepsilon_2$$

Oznacza to, że można odpowiednio zmodyfikować poprzednie równanie:

$$IceCreamSales = \beta_T \cdot Temperature + \beta_S \cdot SummerMonth + \beta_C \cdot (1\,000 \cdot Temperature + \varepsilon_2) = (\beta_T + 1\,000\,\beta_C) \cdot Temperature + \beta_S \cdot SummerMonth$$

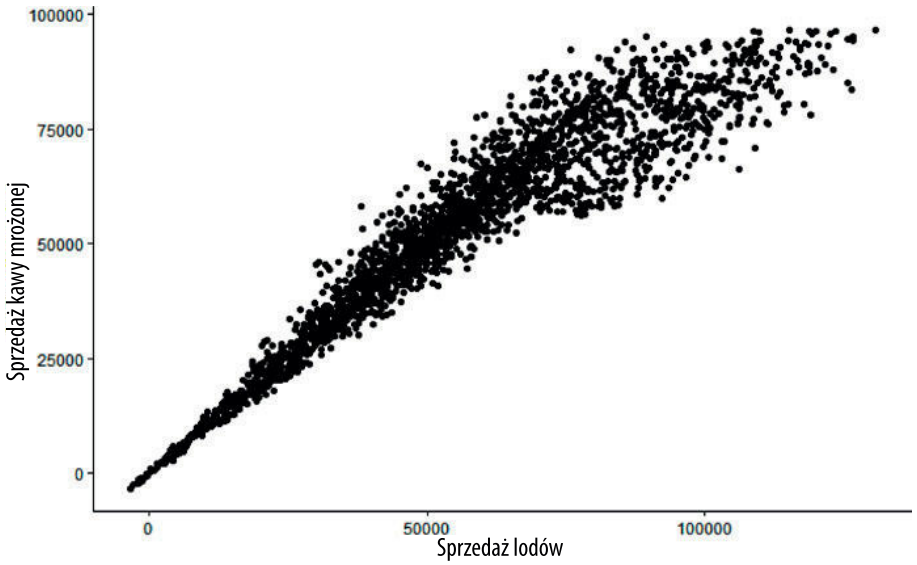
Pomijając losowe przypadki, współczynnik β_S powinien być bliski prawdziwej wartości 20 000. Jednak w przypadku temperatury oprogramowanie będzie próbowało rozwiązać poniższe równanie:

$$\beta_T + 1\,000 \cdot \beta_C = 1\,000$$

Jest to jedno równanie z dwiema niewiadomymi, więc ma nieskończoną liczbę rozwiązań. Na przykład sprawdzą się wartości $\beta_T = 0$ i $\beta_C = 1$, jak również $\beta_T = 500$ i $\beta_C = 0.5$, czy też $\beta_T = 5\,000$ i $\beta_C = -4$. Metoda najmniejszych kwadratów wybierze kombinację, która zapewni najwyższą wartość współczynnika determinacji *R*2, ale nie będzie wiarygodna (chociaż generalnie rzecz ujmując, zawodność i tak będzie znacznie mniejsza w praktyce niż w przypadku tego symulowanego przykładu). Mówiąc językiem formalnym, pojawiła się wielowspółliniowość.

względu na nadmierny wpływ tejże temperatury. Mimo że sprzedaż kawy mrożonej nie jest statystycznie istotna, a współczynnik jest stosunkowo niski, wartość sprzedaży jest znacznie wyższa niż wzrost temperatury. Ostatecznie sprzedaż kawy mrożonej równoważy więc inflację współczynnika temperatury.

Dodanie zmiennej *IcedCoffeeSales* do regresji przysłoniło istnienie związku między temperaturą a sprzedażą lodów. Niestety, może się również pojawić odwrotna sytuacja – uwzględnienie niewłaściwej zmiennej w regresji może stworzyć iluzję związku, którego nie ma.

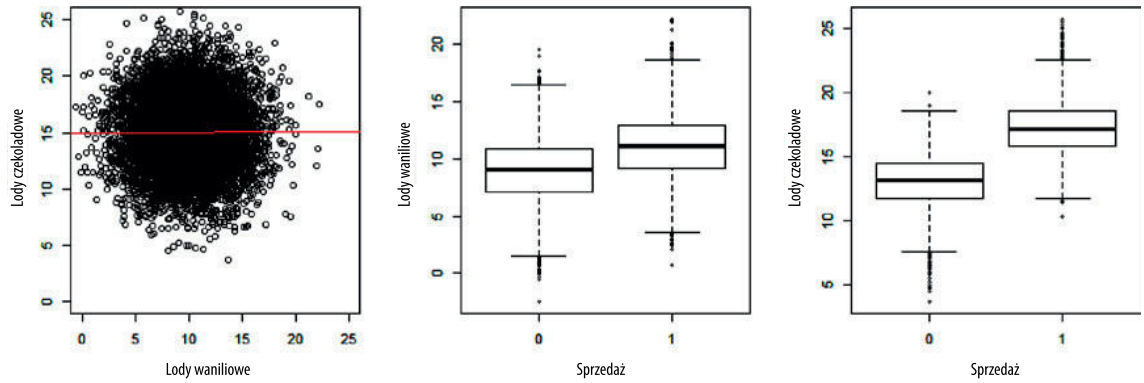


Rysunek 1.4 *Wykres zależności sprzedaży kawy mrożonej od sprzedaży lodów*

Pozostańmy jeszcze przez chwilę przy przykładzie sprzedaży lodów w stoiskach firmy C-Mart. Załóżmy, że menedżer zarządzający kategoriami produktów chce zrozumieć gusta klientów, więc prosi swojego pracownika, aby stanął przed sklepem i pytał przechodzące osoby o to, jak bardzo lubią lody waniliowe i czekoladowe (odpowiedzi powinny być liczbami w skali od 0 do 25), a także, czy kiedykolwiek kupowali lody na stoisku. Aby uprościć sprawę, załóżmy, że stoisko sprzedaje wyłącznie lody czekoladowe i waniliowe. Również przyjmijmy, że smaki lodów waniliowych i czekoladowych są ze sobą całkowicie nieskorelowane. Niektórzy ludzie lubią jedne, ale nie drugie, inni w równym stopniu obydwa, jeszcze inni jedne *bardziej* niż drugie itd. Wszystkie te preferencje wpływają na zmienną binarną (tak/nie) oznaczającą, czy ktoś kupuje lody na stoisku.

Ponieważ zmienna *Shopped* jest binarna, można byłoby skorzystać z regresji logistycznej, gdyby należało zmierzyć wpływ jednej ze zmiennych *Taste* na zachowanie związane z zakupami. Ponieważ te dwie zmienne *Taste* nie są skorelowane, można byłoby dostrzec regularną chmurę bez widocznej korelacji, gdyby je nanieść na siebie. Jednak każda z nich wpływa na prawdopodobieństwo zakupów w lodziarni (rysunek 1.5).

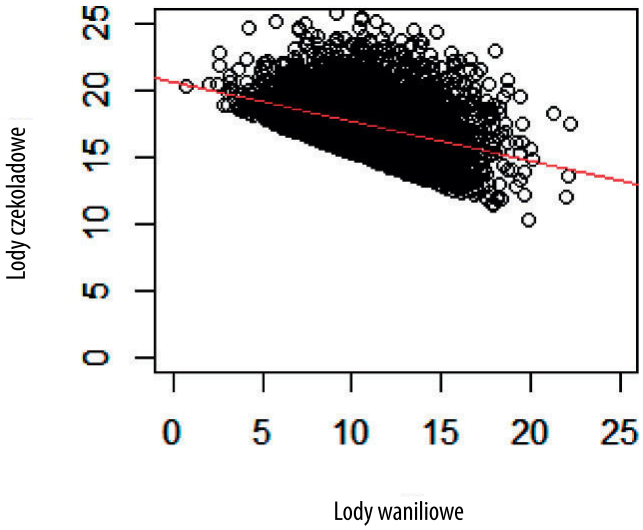
Na pierwszym wykresie została dodatkowo umieszczona linia najlepszego dopasowania, która jest prawie idealnie pozioma i odzwierciedla brak korelacji między zmiennymi (współczynnik korelacji wynosi 0,004 i odpowiada błędowi próbkowania). Na drugim i trzecim wykresie widać, że smaki waniliowy i czekoladowy są ogólnie rzecz ujmując częściej wybierane przez klientów sklepu (*Shopped* = 1), niż przez osoby, które nimi nie są. Ma to sens.



Rysunek 1.5 *Po lewej: smaki waniliowy i czekoladowy nie są skorelowane w całej populacji; na środku: smak waniliowy jest częściej wybierany przez osoby robiące zakupy w lodziarniach niż te, które tego nie robią; po prawej: ten sam wynik uzyskuje się w przypadku smaku czekoladowego*

Na razie rozumowanie przebiega poprawnie. Załóżmy, że po przeanalizowaniu danych z ankiety zarząd firmy stwierdził, iż rozważa możliwość wprowadzenia kuponu zachęcającego do korzystania stoiska z lodami. Kupując lody klient otrzyma bon umożliwiający uzyskanie zniżki podczas przyszłych odwiedzin. Ta zachęta lojalnościowa nie wpłynie na osoby, które nigdy nie robiły zakupów na stoisku, więc odpowiednią grupą będą klienci sklepu. Zarząd rozważa także, by w przypadku kuponów zastosować ograniczenia smakowe w celu zrównoważenia zapasów. Nie wie jednak, jak bardzo może to wpłynąć na wybór smaku. Czy ktoś, kto kupił lody waniliowe i otrzymał bon na lody czekoladowe z 50% zniżką, zrobiłby coś więcej poza wyrzuceniem kolejnego papierka do kosza? W jakim stopniu pozytywnie oceniają lody czekoladowe te osoby, które lubią waniliowe?

Ponownie wygenerujmy ten sam wykres, ale tym razem ograniczając dane do osób, które odpowiedziały „Tak” na pytanie dotyczące zakupów w stoisku (rysunek 1.6).



Rysunek 1.6 *Opinie kupujących dotyczące smaków waniliowego i czekoladowego*

Pomiędzy tymi dwiema zmiennymi występuje teraz silna ujemna korelacja (współczynnik korelacji wynosi –0,39). Co się stało? Czy miłośnicy smaku waniliowego, którzy korzystali ze stoiska sklepu, zaczęli żywić odrazę do smaku czekoladowego (i odwrotnie)? Oczywiście nie jest to właściwa odpowiedź. Ta korelacja została sztucznie stworzona, ponieważ podczas analizy ograniczono się do samych klientów.

Wróćmy do prawdziwych związków przyczynowych: im bardziej ktoś lubi smak waniliowy, tym większe prawdopodobieństwo istnieje, że będzie robił zakupy na stoisku sklepu. Podobna sytuacja występuje w przypadku smaku czekoladowego. Oznacza to, że pojawił się łączny wpływ tych dwóch zmiennych. Jeśli ktoś nie lubi zarówno lodów waniliowych, jak i czekoladowych, jest bardzo mało prawdopodobne, że będzie robić zakupy na stoisku. Innymi słowy, większość klientów nieprzepadających za smakiem waniliowym uwielbia smak czekoladowy. Z drugiej strony, jeśli ktoś lubi smak waniliowy, również może robić zakupy na stoisku, nawet jeśli nie gustuje w smaku czekoladowym. Można to zauważyć na wcześniejszym wykresie – dla wysokich wartości ocen smaku waniliowego (przyjmijmy, że powyżej 15) istnieją punkty danych z niższymi wartościami dla ocen smaku czekoladowego (poniżej 15), podczas gdy dla niskich wartości ocen smaku waniliowego (poniżej 5) jedyne punkty na wykresie odpowiadają wysokim wartościom ocen dla smaku czekoladowego (powyżej 17). Żadne preferencje się nie zmieniły, ale osoby, które nie przepadają zarówno za smakiem waniliowym, jak i czekoladowym, zostały wykluczone ze zbioru danych.

W literaturze naukowej takie zjawisko jest znane pod nazwą paradoksu Berksona (*https://oreil.ly/KwJ1R*), jednak Judea Pearl i Dana Mackenzie określają je bardziej intuicyjną nazwą jako „efekt usprawiedliwienia". Jeśli któryś z klientów bardzo lubi smak waniliowy, to całkowicie wyjaśnia, dlaczego robi zakupy na stoisku sklepu. Nie musi przy tym wcale przepadać za smakiem czekoladowym. Z drugiej strony, jeśli klienci nie lubią smaku waniliowego, nie można na podstawie tego sądzić, że z tego powodu robią zakupy na stoisku sklepu. Chętnie jednak wybierają oni lody o smaku czekoladowym.

Paradoks Berksona jest sprzeczny z intuicją, więc z początku trudny do zrozumienia. W zależności od sposobu gromadzenia danych może wprowadzać zniekształcenia. Dzieje się tak nawet przed rozpoczęciem jakiejkolwiek analizy. Klasycznym przykładem tworzenia sztucznych korelacji jest to, że niektóre choroby wykazują wyższy stopień korelacji wśród pacjentów szpitali w porównaniu z całym społeczeństwem. Wyjaśnienie polega na tym, że nie każda osoba zostaje skierowana do szpitala z powodu pojedynczej choroby. Hospitalizacja dochodzi do skutku tylko wtedy, gdy stan zdrowia pacjenta staje się rzeczywiście zły z powodu większej liczby chorób współistniejących[3].

3 Formalnie rzecz ujmując, jest to nieco inna sytuacja, ponieważ zamiast dwóch liniowych (lub logistycznych) zależności pojawia się efekt progowy. Nadal jednak obowiązuje podstawowa zasada polegająca na tym, że uwzględnienie niewłaściwej zmiennej może wygenerować sztuczne korelacje.

Podsumowanie

Analiza predykcyjna odniosła ogromny sukces w ciągu ostatnich kilku dekad, a w przyszłości również będzie odgrywać dużą rolę. Jeśli jednak celem jest zrozumienie – i co ważniejsze – zmiana ludzkich zachowań, ciekawą alternatywą staje się analiza przyczynowa.

Taka analiza wymaga jednak innego podejścia niż to, jakie jest stosowane w przypadku analizy predykcyjnej. Miejmy nadzieję, że przykłady zaprezentowane w tym rozdziale przekonały Czytelnika, że nie można po prostu wrzucić wielu zmiennych do regresji liniowej lub logistycznej, a następnie mieć nadzieję na uzyskanie dobrych wyników (taką metodę można byłoby opisać słowami „Użyj wszystkich danych. Bóg rozpozna swoje”[4]). Nadal jednak można się zastanawiać, czy nie warto byłoby wykorzystać innych rodzajów modeli i algorytmów. Czy modele wzmacniania gradientowego lub uczenia głębokiego są w jakiś sposób odporne na czynniki zakłócające, wielowspółliniowość i fałszywe korelacje? Niestety, odpowiedź brzmi – nie. Modele te są „czarnymi skrzynkami”, co oznacza, że zakłócenia mogą być jeszcze trudniejsze do wykrycia.

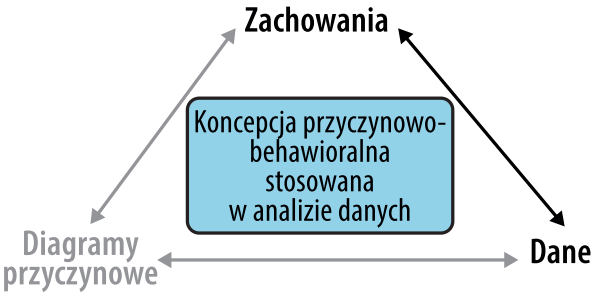
W następnym rozdziale dowiemy się, jak należy traktować same dane behawioralne.

4 Trawestacja łacińskiej frazy „Caedite eos. Novit enim Dominus qui sunt eius” („Zabijcie wszystkich! Bóg rozpozna swoich”) – *przyp. tłum.*

Rozdział 2

Zrozumienie danych behawioralnych

W rozdziale 1. stwierdzono, że głównym celem tej książki jest wyjaśnienie, jak należy wykorzystywać analizę danych, aby zrozumieć, co kieruje ludzkimi zachowaniami. Wymaga to zrozumienia związku zachodzącego między danymi a zachowaniami, który został przedstawiony za pomocą strzałki na schemacie koncepcji przyczynowo-behawioralnej (patrz rysunek 2.1).



Rysunek 2.1 *Schemat koncepcji przyczynowo-behawioralnej, na którym wyróżniono strzałkę łączącą elementy omawiane w tym rozdziale*

Niech Czytelnik wybaczy odniesienie do popkultury, ale jeśli oglądał film *Matrix*, zapewne przypomina sobie, że główny bohater mógł obserwować otaczający go świat i dostrzegać związane z nim liczby. W tym rozdziale dowiemy się, jak można się nauczyć traktować dane w taki sposób, by dostrzegać związane z nimi zachowania.

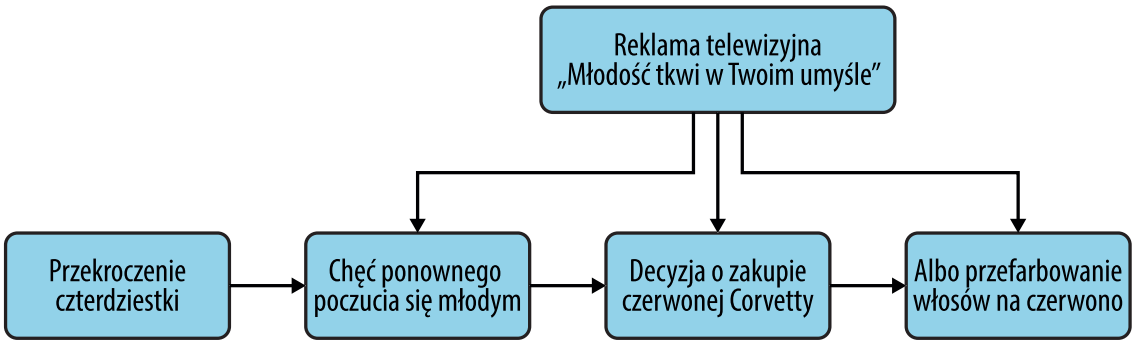
Treść pierwszego podrozdziału jest skierowana głównie do osób z pewnym doświadczeniem biznesowym lub dotyczącym analizowania danych, dlatego zawiera wprowadzenie do podstawowych pojęć wykorzystywanych w naukach behawioralnych. Jeśli Czytelnik jest z wykształcenia behawiorystą, prawdopodobnie nie znajdzie tu wielu nowości. Zawsze jednak może przejrzeć ten podrozdział tylko po to, aby poznać określone terminy, które będą używane w książce.

W drugim podrozdziale, przy wykorzystaniu ogólnej metody rozumowania zaprezentujemy, jak można się przyglądać danym przez pryzmat behawioralny i dla każdej ze zmiennych definiować koncepcję behawioralną. Niestety, w wielu przypadkach zmienne mogą być początkowo luźno powiązane z odpowiadającymi im zachowaniami. Dowiemy się więc, jak można „behawioryzować" takie krnąbrne zmienne.

Podstawowy model ludzkiego zachowania

„Zachowanie" to jedno z tych słów, które są bardzo znane ze względu na częste występowanie. To słowo jest jednak rzadko (jeśli w ogóle) właściwie definiowane. Zapytałem kiedyś partnerkę biznesową, do jakiego zachowania stara się zachęcać ludzi. Jej odpowiedź zaczęła się od słów: „Chcemy, żeby wiedzieli, że…" W tym momencie zdałem sobie sprawę z dwóch rzeczy: (1) pomagając w sprecyzowaniu celów mógłbym zapewnić projektowi większą wartość dodaną niż się początkowo spodziewałem; (2) wprowadzenie do nauk behawioralnych, które wcześniej jej przekazałem, było naprawdę nic nie warte, jeśli nadal myślała, że wiedza o czymś jest zachowaniem. Mam nadzieję, że tym razem spiszę się lepiej, a Czytelnik po przeczytaniu tego podrozdziału będzie mógł dostarczyć swojej organizacji większą wartość dodaną.

Rzeczywiście można uwierzyć w to, że jedną z kluczowych korzyści z przyjęcia behawioralnego sposobu myślenia jest skłonienie ludzi do dokładniejszego zastanowienia się nad tym, co próbują zrobić. Zmiana sposobu myślenia jakiejś osoby to nie to samo, co wpływ na jej działania (i odwrotnie). W książce zaprezentujemy uproszczony, ale miejmy nadzieję, działający model ludzkiego zachowania. Najpierw zostanie on zilustrowany przykładem prezentującym, jak firma specjalizująca się w pielęgnacji ciała może wykorzystać kryzys wieku średniego u klientów (rysunek 2.2).



Rysunek 2.2 *Model ludzkiego zachowania w przypadku kryzysu wieku średniego*

W tym przykładzie cechy osobowe (przekroczenie czterdziestki) prowadzą do powstania określonych emocji i myśli (chęci ponownego poczucia się młodo), co z kolei powoduje pojawienie się intencji (decyzji o zakupie czerwonej corvetty). W zależności od zachowań biznesowych (reklamy telewizyjnej) ta intencja może spowodować powstanie odpowiedniego zachowania lub skutkować czymś innym (na przykład przefarbowaniem włosów na czerwono).

W niektórych okolicznościach można wpływać nie na zachowania klientów, ale pracowników, dostawców itd. Należałoby wtedy odpowiednio dostosować model, ale intuicyjnie rzecz ujmując, jego zasada działania pozostałaby taka sama – z jednej strony mamy do czynienia z człowiekiem, na którego zachowanie próbujemy wpływać, a z drugiej jako przedsiębiorstwo zarządzamy wszystkimi procesami i regułami, a także podejmujemy odpowiednie decyzje (rysunek 2.3).