

Spis treści

Przedmowa	17
1. Wstęp: myślenie w kategoriach analityki danych	25
Wszelchobecność możliwości pozyskiwania danych	25
Przykład: huragan Frances	27
Przykład: prognozowanie odpływu klientów	27
Nauka o danych, inżynieria i podejmowanie decyzji na podstawie danych	28
Przetwarzanie danych i Big Data	31
Od Big Data 1.0 do Big Data 2.0	32
Dane i potencjał nauki o danych jako aktywa strategiczne	32
Myślenie w kategoriach analityki danych	35
Nasza książka	37
Eksploracja danych i nauka o danych, nowe spojrzenie	37
Chemia to nie próbówki: nauka o danych kontra praca badacza danych	38
Podsumowanie	39
2. Problemy biznesowe a rozwiązania z zakresu nauki o danych	41
Podstawowe pojęcia: <i>Zbiór kanonicznych zadań związanych z eksploracją danych; Proces eksploracji danych; Nadzorowana i nienadzorowana eksploracja danych.</i>	
Od problemów biznesowych do zadań eksploracji danych	41
Metody nadzorowane i nienadzorowane	45
Eksploracja danych i jej wyniki	47
Proces eksploracji danych	47
Zrozumienie uwarunkowań biznesowych	49
Zrozumienie danych	49
Przygotowanie danych	51
Modelowanie	52
Ewaluacja	52
Wdrożenie	53
Implikacje w sferze zarządzania zespołem nauki o danych	55
Inne techniki i technologie analityczne	56
Statystyka	56
Zapytania do baz danych	58
Magazynowanie danych	59
Analiza regresji	59
Uczenie maszynowe i eksploracja danych	60
Odpowiadanie na pytania biznesowe z wykorzystaniem tych technik	61
Podsumowanie	62

3. Wprowadzenie do modelowania predykcyjnego: od korelacji do nadzorowanej segmentacji	63
Podstawowe pojęcia: <i>Identyfikowanie atrybutów informatywnych; Segmentowanie danych za pomocą progresywnej selekcji atrybutów.</i>	
Przykładowe techniki: <i>Wyszukiwanie korelacji; Wybór atrybutów/zmiennych; Indukcja drzew decyzyjnych.</i>	
Modele, indukcja i predykcja	64
Nadzorowana segmentacja	67
Wybór atrybutów informatywnych	68
Przykład: wybór atrybutu z wykorzystaniem przyrostu informacji	74
Nadzorowana segmentacja z użyciem modeli o strukturze drzewa	79
Wizualizacja segmentacji	83
Drzewa jako zbiory reguł	86
Szacowanie prawdopodobieństwa	86
Przykład: rozwiązywanie problemu odpływu abonentów z wykorzystaniem indukcji drzewa	88
Podsumowanie	92
4. Dopasowywanie modelu do danych	95
Podstawowe pojęcia: <i>Znajdowanie „optymalnych” parametrów modelu na podstawie danych; Wybieranie celu eksploracji danych; Funkcje celu; Funkcje straty.</i>	
Przykładowe techniki: <i>Regresja liniowa; Regresja logistyczna; Maszyny wektorów wspierających.</i>	
Klasyfikacja za pomocą funkcji matematycznych	96
Liniowe funkcje dyskryminacyjne	97
Optymalizacja funkcji celu	100
Przykład wydobywania dyskryminatora liniowego z danych	101
Liniowe funkcje dyskryminacyjne do celów scoringu i szeregowania wystąpień	102
Maszyny wektorów wspierających w skrócie	103
Regresja za pomocą funkcji matematycznych	106
Szacowanie prawdopodobieństwa klas i „regresja” logistyczna	108
* Regresja logistyczna: kilka szczegółów technicznych	111
Przykład: indukcja drzew decyzyjnych a regresja logistyczna	113
Funkcje nieliniowe, maszyny wektorów wspierających i sieci neuronowe	117
Podsumowanie	119
5. Nadmierne dopasowanie i jego unikanie	121
Podstawowe pojęcia: <i>Generalizacja; Dopasowanie i nadmierne dopasowanie; Kontrola złożoności.</i>	
Przykładowe techniki: <i>Sprawdzian krzyżowy; Wybór atrybutów; Przycinanie drzew; Regularyzacja.</i>	
Generalizacja	121
Nadmierne dopasowanie („przeliczenie”)	122

Badanie nadmiernego dopasowania	123
Dane wydzielone i wykresy dopasowania	123
Nadmierne dopasowanie w indukcji drzew decyzyjnych	125
Nadmierne dopasowanie w funkcjach matematycznych	127
Przykład: nadmierne dopasowanie funkcji liniowych	128
* Przykład: dlaczego nadmierne dopasowanie jest niekorzystne?	131
Od ewaluacji danych wydzielonych do sprawdzianu krzyżowego	133
Zbiór danych dotyczących odpływu abonentów — nowe spojrzenie	136
Krzywe uczenia się	137
Unikanie nadmiernego dopasowania i kontrola złożoności	139
Unikanie nadmiernego dopasowania w indukcji drzew decyzyjnych	139
Ogólna metoda unikania nadmiernego dopasowania	141
* Unikanie nadmiernego dopasowania w celu optymalizacji parametrów	142
Podsumowanie	145

6. Podobieństwo, sąsiedzi i klastry **147**

<i>Podstawowe pojęcia: Obliczanie podobieństwa obiektów opisanych przez dane; Wykorzystywanie podobieństwa do celów predykcji; Klastrowanie jako segmentacja oparta na podobieństwie.</i>	
<i>Przykładowe techniki: Poszukiwanie podobnych jednostek; Metody najbliższych sąsiadów; Metody klastrowania; Miary odległości do obliczania podobieństwa.</i>	
Podobieństwo i odległość	148
Wnioskowanie metodą najbliższych sąsiadów	150
Przykład: analityka whisky	150
Najbliżsi sąsiedzi w modelowaniu predykcyjnym	152
Ilu sąsiadów i jak duży wpływ?	154
Interpretacja geometryczna, nadmierne dopasowanie i kontrola złożoności	156
Problemy z metodami najbliższych sąsiadów	158
Kilka istotnych szczegółów technicznych dotyczących podobieństw i sąsiadów	162
Atrybuty heterogeniczne	162
* Inne funkcje odległości	163
* Funkcje łączące: obliczanie wskaźników na podstawie sąsiadów	165
Klastrowanie	167
Przykład: analityka whisky — nowe spojrzenie	167
Klastrowanie hierarchiczne	168
Najbliżsi sąsiedzi na nowo: klastrowanie wokół centroidów	172
Przykład: klastrowanie wiadomości biznesowych	176
Zrozumienie wyników klastrowania	179
* Wykorzystywanie uczenia nadzorowanego do generowania opisów klastrów	181
Krok wstecz: rozwiązywanie problemu biznesowego kontra eksploracja danych	183
Podsumowanie	185

7. Myślenie w kategoriach analityki decyzji I: co to jest dobry model? **187**

Podstawowe pojęcia: *Staranne rozważenie, czego oczekujemy od wyników nauki o danych; Wartość oczekiwana jako kluczowa platforma ewaluacji; Uwzględnianie odpowiednich porównawczych punktów odniesienia.*

Przykładowe techniki: *Różne miary ewaluacji; Szacowanie kosztów i korzyści; Obliczanie oczekiwanego zysku; Tworzenie metod bazowych dla porównań.*

Ewaluacja klasyfikatorów	188
Zwykła dokładność i jej problemy	189
Macierz pomyłek	189
Problemy z niezerównoważonymi klasami	190
Problemy nierównych kosztów i korzyści	191
Generalizowanie poza klasyfikacją	193
Kluczowa platforma analityczna: wartość oczekiwana	193
Wykorzystywanie wartości oczekiwanej do systematyzowania zastosowania klasyfikatora	194
Wykorzystywanie wartości oczekiwanej do systematyzowania ewaluacji klasyfikatora	195
Ewaluacja, skuteczność bazowa oraz implikacje dla inwestowania w dane	201
Podsumowanie	205

8. Wizualizacja skuteczności modelu **207**

Podstawowe pojęcia: *Wizualizacja skuteczności modelu przy różnych rodzajach niepewności; Dalsze rozważania odnośnie tego, czego należy oczekiwać od wyników eksploracji danych.*

Przykładowe techniki: *Krzywe zysku; Krzywe łącznej reakcji; Krzywe przyrostu; Krzywe ROC.*

Ranking zamiast klasyfikowania	207
Krzywe zysku	209
Wykresy i krzywe ROC	212
Pole pod krzywą ROC (AUC)	216
Krzywe łącznej reakcji i krzywe przyrostu	216
Przykład: analityka skuteczności w modelowaniu odpływu abonentów	219
Podsumowanie	226

9. Dowody i prawdopodobieństwa **227**

Podstawowe pojęcia: *Jednoznaczne łączenie dowodów za pomocą twierdzenia Bayesa; Wnioskowanie probabilistyczne poprzez założenia warunkowej niezależności.*

Przykładowe techniki: *Klasyfikacja bayesowska; Przyrost wartości dowodu.*

Przykład: targetowanie klientów reklam internetowych	227
Probabilistyczne łączenie dowodów	229

Prawdopodobieństwo łączne i niezależność	230
Twierdzenie Bayesa	231
Zastosowanie twierdzenia Bayesa w nauce o danych	232
Niezależność warunkowa i naiwny klasyfikator bayesowski	234
Zalety i wady naiwnego klasyfikatora bayesowskiego	235
Model „przyrostu” wartości dowodu	237
Przykład: przyrosty wartości dowodów z „polubień” na Facebooku	238
Dowody w akcji: targetowanie klientów reklamami	240
Podsumowanie	240

10. Reprezentacja i eksploracja tekstu **243**

Podstawowe pojęcia: <i>Znaczenie konstruowania przyjaznych eksploracji reprezentacji danych; Reprezentacja tekstu do celów eksploracji danych.</i>	
Przykładowe techniki: <i>Reprezentacja worka słów (bag of words); Kalkulacja TFIDF; N-gramy; Sprowadzanie do formy podstawowej (stemming,); Ekstrakcja wyrażeń nazwowych; Modele tematyczne.</i>	
Dlaczego tekst jest istotny	244
Dlaczego tekst jest trudny	244
Reprezentacja	245
Worek słów (bag of words)	245
Częstość termów	246
Mierzenie rzadkości (sparseness): odwrotna częstość w dokumentach	248
Łączenie reprezentacji: TFIDF	249
Przykład: muzycy jazzowi	250
* Związek IDF z entropią	253
Oprócz worka słów	255
N-gramy	255
Ekstrakcja wyrażeń nazwowych	255
Modele tematyczne	256
Przykład: eksploracja wiadomości w celu prognozowania zmian cen akcji	257
Zadanie	257
Dane	259
Wstępne przetwarzanie danych	262
Wyniki	262
Podsumowanie	266

11. Myślenie w kategoriach analityki decyzji II: w kierunku inżynierii analitycznej **267**

Podstawowe pojęcie: <i>Rozwiązywanie problemów biznesowych z wykorzystaniem nauki o danych rozpoczyna się od inżynierii analitycznej: projektowania rozwiązania analitycznego z wykorzystaniem dostępnych danych, narzędzi i technik.</i>	
Przykładowa technika: <i>Wartość oczekiwana jako platforma opracowania rozwiązania z zakresu nauki o danych.</i>	
Targetowanie najlepszych potencjalnych klientów przesylek organizacji pozyskujących fundusze	268

Platforma wartości oczekiwanej: rozkład problemu biznesowego i ponowne zestawienie elementów rozwiązania	268
Krótką dygresją na temat stroniczości selekcji	270
Nowe, jeszcze bardziej zaawansowane spojrzenie na nasz przykład odpływu abonentów	271
Platforma wartości oczekiwanej: strukturyzacja bardziej skomplikowanego problemu biznesowego	271
Ocena wpływu zachęty	272
Od rozkładu wartości oczekiwanej do rozwiązania z obszaru nauki o danych	274
Podsumowanie	277

12. Inne zadania i techniki nauki o danych **279**

Podstawowe pojęcia: *Nasze podstawowe pojęcia jako baza wielu typowych technik nauki o danych; Znaczenie wiedzy o elementach składowych nauki o danych.*

Przykładowe techniki: *Zależność i współwystępowanie; Profilowanie zachowań; Predykcja połączeń; Redukcja danych; Eksploracja informacji ukrytych; Rekomendowanie filmów; Rozkład błędu pod względem stroniczości — wariancji; Zespoły modeli; Wnioskowanie przyczynowe z danych.*

Współwystąpienia i zależności: znajdowanie elementów, które idą w parze	280
Pomiar zaskoczenia: przyrost i dźwignia	281
Przykład: piwo i kupony loteryjne	282
Zależności pomiędzy polubieniami na Facebooku	282
Profilowanie: znajdowanie typowego zachowania	285
Predykcja połączeń i rekomendacje społecznościowe	290
Redukcja danych, informacje ukryte i rekomendacje filmów	291
Stroniczość, wariancja i metody zespalań	294
Oparte na danych wyjaśnianie przyczynowe i przykład marketingu wirusowego	297
Podsumowanie	298

13. Nauka o danych i strategia biznesowa **301**

Podstawowe pojęcia: *Nasze zasady jako podstawa sukcesu firmy działającej na podstawie danych; Zdobywanie i utrzymywanie przewagi konkurencyjnej za pomocą nauki o danych; Znaczenie dbałości o potencjał nauki o danych.*

Myślenie w kategoriach analityki danych, raz jeszcze	301
Osiąganie przewagi konkurencyjnej przy pomocy nauki o danych	303
Utrzymywanie przewagi konkurencyjnej przy pomocy nauki o danych	304
Nadzwyczajna przewaga historyczna	305
Wyjątkowa własność intelektualna	305
Wyjątkowe niematerialne aktywa zabezpieczające	306
Lepsi badacze danych	306

Lepsze zarządzanie zespołem nauki o danych	308
Pozyskiwanie badaczy danych i ich zespołów oraz opieka nad nimi	309
Badanie studiów przypadku z zakresu nauki o danych	311
Gotowość do przyjmowania kreatywnych pomysłów z każdego źródła	312
Gotowość do oceny propozycji projektów z zakresu nauki o danych	312
Przykładowa propozycja eksploracji danych	313
Błędy w propozycji Big Red	313
Dojrzałość firmy w sferze nauki o danych	315
14. Zakończenie	317
Podstawowe pojęcia nauki o danych	317
Zastosowanie naszych podstawowych pojęć do nowego problemu: eksploracji danych urządzeń przenośnych	320
Zmiana sposobu myślenia o rozwiązaniach problemów biznesowych	322
Czego dane nie mogą dokonać: nowe spojrzenie na decydentów	323
Prywatność, etyka i eksploracja danych dotyczących konkretnych osób	326
Czy jest coś jeszcze w nauce o danych?	327
Ostatni przykład: od crowdsourcingu do cloudsourceingu	328
Kilka słów na zakończenie	329
A. Przewodnik dotyczący oceny propozycji	331
Zrozumienie uwarunkowań biznesowych i zrozumienie danych	331
Przygotowanie danych	332
Modelowanie	332
Ewaluacja i wdrożenie	333
B. Jeszcze jedna przykładowa propozycja	335
Scenariusz i propozycja	335
Wady propozycji GGC	336
C. Słowniczek	339
D. Bibliografia	345
Skorowidz	351