

# Spis treści

---

<i>Przedmowa</i>	13
<i>Wstęp</i>	15
<i>Podziękowania</i>	17
<i>Informacje o książce</i>	19
<i>Informacje o autorach</i>	27
<i>Informacje o autorach przedmowy</i>	29
<b>CZĘŚĆ I WPROWADZENIE DO ANALIZY DANYCH</b>	<b>31</b>
<b>1. Proces analizy danych</b>	<b>33</b>
1.1. Role w projekcie analizy danych	34
1.1.1. Role w projekcie	34
1.2. Etapy projektu analizy danych	36
1.2.1. Definiowanie celu	37
1.2.2. Gromadzenie danych i zarządzanie nimi	39
1.2.3. Modelowanie	41
1.2.4. Ocena i krytyka modelu	43
1.2.5. Prezentacja i dokumentowanie	45
1.2.6. Wdrażanie i utrzymywanie modelu	47
1.3. Wyznaczanie oczekiwań	47
1.3.1. Określenie dolnego pułapu wydajności modelu	48
Podsumowanie	49
<b>2. Wprowadzenie do języka R i danych</b>	<b>51</b>
2.1. Początki z R	52
2.1.1. Instalowanie R, narzędzi i przykładów	53
2.1.2. Programowanie w R	53
2.2. Praca z danymi przechowywanymi w plikach	63
2.2.1. Praca z danymi ustrukturyzowanymi z poziomu plików lub adresów URL	63
2.2.2. Praca z mniej ustrukturyzowanymi danymi	68
2.3. Praca z relacyjnymi bazami danych	71
2.3.1. Przykładowe dane o rozmiarze produkcyjnym	72
Podsumowanie	83
<b>3. Eksploracja danych</b>	<b>85</b>
3.1. Wykrywanie problemów za pomocą statystyk podsumowujących	87
3.1.1. Typowe problemy wykrywane za pomocą podsumowania danych	88

3.2.	Wykrywanie problemów za pomocą grafiki i wizualizacji	92
3.2.1.	Wizualne sprawdzanie rozkładów dla jednej zmiennej	94
3.2.2.	Wizualne sprawdzanie relacji pomiędzy dwiema zmiennymi	104
	Podsumowanie	119
<b>4.</b>	<b>Zarządzanie danymi</b>	<b>121</b>
4.1.	Oczyszczanie danych	121
4.1.1.	Oczyszczanie danych specyficznych dla danej dziedziny	122
4.1.2.	Naprawianie brakujących wartości	124
4.1.3.	Pakiet <i>vtreat</i> służący do automatycznego naprawiania brakujących danych	128
4.2.	Przekształcenia danych	131
4.2.1.	Normalizacja	132
4.2.2.	Środkowanie i skalowanie	133
4.2.3.	Przekształcenia logarytmiczne rozkładów nierównomiernych i szerokich	137
4.3.	Losowanie danych do modelowania i walidacji	140
4.3.1.	Zbiory uczący i testowy	141
4.3.2.	Tworzenie kolumny grupowania próby	142
4.3.3.	Grupowanie rekordów	143
4.3.4.	Pochodzenie danych	144
	Podsumowanie	144
<b>5.</b>	<b>Inżynieria i kształtowanie danych</b>	<b>147</b>
5.1.	Dobieranie danych	150
5.1.1.	Wyznaczanie podzbiorów rzędów i kolumn	150
5.1.2.	Usuwanie rekordów z brakującymi danymi	155
5.1.3.	Wyznaczanie kolejności rzędów	158
5.2.	Podstawowe przekształcenia danych	162
5.2.1.	Dodawanie nowych kolumn	162
5.2.2.	Inne proste operacje	168
5.3.	Przekształcenia agregacyjne	168
5.3.1.	Łączenie wielu rzędów w rzędy podsumowujące	168
5.4.	Wielotablicowe przekształcenia danych	172
5.4.1.	Szybkie łączenie co najmniej dwóch uporządkowanych ramek danych	172
5.4.2.	Główne metody łączenia danych pochodzących z wielu tabel	177
5.5.	Transformacje przestawiające	184
5.5.1.	Przenoszenie danych z formy szerokiej do wysokiej	184
5.5.2.	Przenoszenie danych z formy wysokiej do szerokiej	188
5.5.3.	Współrzędne danych	193
	Podsumowanie	194
<b>CZĘŚĆ II</b>	<b>METODY MODELOWANIA</b>	<b>195</b>
<b>6.</b>	<b>Wybór i ocena modeli</b>	<b>197</b>
6.1.	Odwzorowywanie problemów na zadania uczenia maszynowego	197
6.1.1.	Zadania klasyfikacji	199
6.1.2.	Zadania obliczania wyniku	199

6.1.3.	<i>Grupowanie — praca bez znajomości zmiennych docelowych</i>	200
6.1.4.	<i>Odwzorowanie problemu na metodę</i>	202
6.2.	<i>Ocenianie modeli</i>	202
6.2.1.	<i>Przetrenowanie</i>	204
6.2.2.	<i>Wskaźniki wydajności modelu</i>	208
6.2.3.	<i>Ocenianie modeli klasyfikacyjnych</i>	209
6.2.4.	<i>Ocenianie modelu obliczania wyników</i>	218
6.2.5.	<i>Ocenianie modeli prawdopodobieństwa</i>	222
6.3.	<i>Metoda lokalnie wytłumaczalnych wyjaśnień niezależnych od modelu służąca do wyjaśniania przewidywań modelu</i>	229
6.3.1.	<i>LIME — zautomatyzowane sprawdzanie poprawności działania systemu</i>	231
6.3.2.	<i>Stosowanie metody LIME — mały przykład</i>	231
6.3.3.	<i>Metoda LIME w klasyfikacji tekstu</i>	238
6.3.4.	<i>Uczenie klasyfikatora tekstu</i>	241
6.3.5.	<i>Wyjaśnianie przewidywań klasyfikatora</i>	242
	<i>Podsumowanie</i>	247
7.	<i>Regresja liniowa i logistyczna</i>	249
7.1.	<i>Stosowanie regresji liniowej</i>	250
7.1.1.	<i>Mechanizm działania regresji liniowej</i>	251
7.1.2.	<i>Tworzenie modelu regresji liniowej</i>	256
7.1.3.	<i>Uzyskiwanie predykcji</i>	257
7.1.4.	<i>Wyszukiwanie relacji i wydobywanie przydatnych informacji</i>	262
7.1.5.	<i>Odczytywanie podsumowania modelu i określanie jakości współczynników</i>	264
7.1.6.	<i>Kluczowe wnioski na temat regresji liniowej</i>	271
7.2.	<i>Stosowanie regresji logistycznej</i>	271
7.2.1.	<i>Mechanizm działania regresji logistycznej</i>	272
7.2.2.	<i>Tworzenie modelu regresji logistycznej</i>	276
7.2.3.	<i>Uzyskiwanie przewidywań</i>	277
7.2.4.	<i>Wyszukiwanie relacji i wydobywanie użytecznych informacji z modeli logistycznych</i>	282
7.2.5.	<i>Odczytywanie podsumowania modelu i charakteryzowanie współczynników</i>	284
7.2.6.	<i>Kluczowe wnioski na temat regresji logistycznej</i>	291
7.3.	<i>Regularyzacja</i>	291
7.3.1.	<i>Przykład quasi-separacji</i>	292
7.3.2.	<i>Rodzaje regresji regularyzowanej</i>	296
7.3.3.	<i>Regresja regularyzowana przy użyciu pakietu glmnet</i>	298
	<i>Podsumowanie</i>	307
8.	<i>Zaawansowane przygotowywanie danych</i>	309
8.1.	<i>Cel pakietu vtreat</i>	310
8.2.	<i>Konkurs KDD i zestaw danych KDD Cup 2009</i>	312
8.2.1.	<i>Pierwsze kroki z danymi KDD Cup 2009</i>	313
8.2.2.	<i>Metoda „słonia w składzie porcelany”</i>	315

8.3.	Podstawowe przygotowywanie danych do zadań klasyfikacji	318
8.3.1.	<i>Ramka oceny zmiennej</i>	319
8.3.2.	<i>Odpowiednie stosowanie planu naprawy</i>	324
8.4.	Zaawansowane przygotowywanie danych do zadań klasyfikacji	325
8.4.1.	<i>Korzystanie z metody <code>mkCrossFrameCExperiment()</code></i>	325
8.4.2.	<i>Budowanie modelu</i>	328
8.5.	Przygotowywanie danych do zadań regresji	332
8.6.	Opanowanie pakietu <code>vtreat</code>	334
8.6.1.	<i>Fazy mechanizmu <code>vtreat</code></i>	335
8.6.2.	<i>Brakujące wartości</i>	337
8.6.3.	<i>Zmienne wskaźnikowe</i>	338
8.6.4.	<i>Kodowanie wpływu</i>	339
8.6.5.	<i>Plan naprawy</i>	341
8.6.6.	<i>Ramka krzyżowa</i>	341
	Podsumowanie	345
9.	<b>Metody nienadzorowane</b> .....	<b>347</b>
9.1.	Analiza skupień	348
9.1.1.	<i>Odległości</i>	349
9.1.2.	<i>Przygotowanie danych</i>	352
9.1.3.	<i>Hierarchiczna analiza skupień za pomocą funkcji <code>hclust()</code></i>	354
9.1.4.	<i>Algorytm centroidów</i>	367
9.1.5.	<i>Przypisywanie nowych punktów do skupień</i>	374
9.1.6.	<i>Kluczowe wnioski na temat analizy skupień</i>	376
9.2.	Reguły asocjacyjne	377
9.2.1.	<i>Przegląd reguł asocjacyjnych</i>	377
9.2.2.	<i>Przykładowy problem</i>	379
9.2.3.	<i>Wydobywanie reguł asocjacyjnych za pomocą pakietu <code>arules</code></i>	380
9.2.4.	<i>Kluczowe wnioski na temat reguł asocjacyjnych</i>	388
	Podsumowanie	388
10.	<b>Zaawansowane metody uczenia maszynowego</b> .....	<b>391</b>
10.1.	Metody drzewa	393
10.1.1.	<i>Podstawowe drzewo decyzyjne</i>	394
10.1.2.	<i>Usprawnianie przewidywań za pomocą agregacji</i>	397
10.1.3.	<i>Dalsze usprawnianie przewidywań za pomocą lasów losowych</i>	399
10.1.4.	<i>Drzewa wzmacniane gradientowo</i>	405
10.1.5.	<i>Kluczowe wnioski na temat modeli bazujących na drzewach</i>	414
10.2.	Wykrywanie relacji niemonotonicznych za pomocą uogólnionych modeli addytywnych	414
10.2.1.	<i>Mechanizm działania modelu GAM</i>	415
10.2.2.	<i>Przykład regresji jednowymiarowej</i>	415
10.2.3.	<i>Wydobywanie relacji nieliniowych</i>	420
10.2.4.	<i>Stosowanie modelu GAM na rzeczywistych danych</i>	422
10.2.5.	<i>Stosowanie modelu GAM w regresji logistycznej</i>	425
10.2.6.	<i>Kluczowe wnioski na temat modelu GAM</i>	427

- 10.3. Rozwiązywanie problemów „nicrozdzielnych”  
za pomocą maszyn wektorów nośnych 427
  - 10.3.1. *Używanie maszyn SVM do rozwiązywania problemów* 428
  - 10.3.2. *Mechanizm działania maszyn wektorów nośnych* 433
  - 10.3.3. *Mechanizm działania funkcji jądra* 435
  - 10.3.4. *Kluczowe wnioski na temat maszyn wektorów nośnych i metod  
z użyciem jądra* 438
- Podsumowanie 438

## **CZĘŚĆ III PRACA W PRAWDZIWYM ŚWIECIE ..... 441**

### **11. Dokumentowanie i wdrażanie ..... 443**

- 11.1. Przewidywanie szumu medialnego 445
- 11.2. Tworzenie dokumentacji poszczególnych etapów  
za pomocą formatu R Markdown 446
  - 11.2.1. *Czym jest R Markdown?* 447
  - 11.2.2. *Szczegóły techniczne silnika knitr* 449
  - 11.2.3. *Dokumentowanie danych Buzz i tworzenie modelu  
za pomocą pakietu knitr* 450
- 11.3. Sporządzanie dokumentacji bieżącej za pomocą komentarzy i kontroli wersji 454
  - 11.3.1. *Pisanie przydatnych komentarzy* 454
  - 11.3.2. *Rejestrowanie historii za pomocą kontroli wersji* 456
  - 11.3.3. *Eksplorowanie modelu za pomocą kontroli wersji* 461
  - 11.3.4. *Udostępnianie pracy za pomocą kontroli wersji* 463
- 11.4. Wdrażanie modeli 468
  - 11.4.1. *Wdrażanie wersji demonstracyjnych za pomocą narzędzia Shiny* 468
  - 11.4.2. *Wdrażanie modeli jako usług HTTP* 471
  - 11.4.3. *Wdrażanie modeli poprzez eksportowanie* 472
  - 11.4.4. *Kluczowe wnioski* 475
- Podsumowanie 476

### **12. Tworzenie użytecznych prezentacji ..... 477**

- 12.1. Prezentowanie rezultatów sponsorowi projektu 479
  - 12.1.1. *Podsumowanie celów projektu* 479
  - 12.1.2. *Określanie wyników projektu* 481
  - 12.1.3. *Uzupełnianie szczegółów* 482
  - 12.1.4. *Sporządzanie zaleceń i omawianie przyszłych planów* 484
  - 12.1.5. *Kluczowe wnioski na temat prezentacji  
przeznaczonej dla sponsora projektu* 485
- 12.2. Prezentowanie modelu użytkownikom końcowym 485
  - 12.2.1. *Podsumowanie celów projektu* 486
  - 12.2.2. *Omówienie dopasowania modelu do cyklu pracy* 486
  - 12.2.3. *Prezentowanie sposobu korzystania z modelu* 487
  - 12.2.4. *Kluczowe wnioski na temat prezentacji przeznaczonej  
dla użytkowników końcowych* 489
- 12.3. Prezentowanie pracy innym analitykom danych 490
  - 12.3.1. *Wprowadzenie do problemu* 491
  - 12.3.2. *Omówienie powiązanej pracy* 491

12.3.3.	<i>Opis Twojego rozwiązania</i>	492
12.3.4.	<i>Omówienie wyników i przyszłych planów</i>	492
12.3.5.	<i>Kluczowe wnioski na temat prezentacji przeznaczonych dla partnerów</i>	493
	<b>Podsumowanie</b>	494
	<b><i>Dodatek A Korzystanie z R i innych narzędzi</i></b> .....	<b>497</b>
	<b><i>Dodatek B Ważne pojęcia z dziedziny statystyki</i></b> .....	<b>523</b>
	<b><i>Dodatek C Bibliografia</i></b> .....	<b>559</b>