

Spis treści

Przedmowa	13
Wstęp	15
Podziękowania	17
Informacje o książce	19
Informacje o autorach	27
Informacje o autorach przedmowy	29
CZĘŚĆ I WPROWADZENIE DO ANALIZY DANYCH	31
1. Proces analizy danych	33
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
2. Wprowadzenie do języka R i danych	51
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

3. Eksploracja danych	85
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
4. Zarządzanie danymi	121
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 vtreat 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
5. Inżynieria i kształtowanie danych	147
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

CZĘŚĆ II METODY MODELOWANIA	195
6. Wybór i ocena modeli	197
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. Grupowanie — praca bez znajomości zmiennych docelowych	200
6.1.4. Odwzorowanie problemu na metodę	202
6.2. Ocenianie modeli	202
6.2.1. Przetrenowanie	204
6.2.2. Wskaźniki wydajności modelu	208
6.2.3. Ocenianie modeli klasyfikacyjnych	209
6.2.4. Ocenianie modelu obliczania wyników	218
6.2.5. Ocenianie modeli prawdopodobieństwa	222
6.3. Metoda lokalnie wytłumaczalnych wyjaśnień niezależnych od modelu służąca do wyjaśniania przewidywań modelu	229
6.3.1. LIME — zautomatyzowane sprawdzanie poprawności działania systemu	231
6.3.2. Stosowanie metody LIME — mały przykład	231
6.3.3. Metoda LIME w klasyfikacji tekstu	238
6.3.4. Uczenie klasyfikatora tekstu	241
6.3.5. Wyjaśnianie przewidywań klasyfikatora	242
Podsumowanie	247
7. Regresja liniowa i logistyczna	249
7.1. Stosowanie regresji liniowej	250
7.1.1. Mechanizm działania regresji liniowej	251
7.1.2. Tworzenie modelu regresji liniowej	256
7.1.3. Uzyskiwanie predykcji	257
7.1.4. Wyszukiwanie relacji i wydobywanie przydatnych informacji	262
7.1.5. Odczytywanie podsumowania modelu i określanie jakości współczynników	264
7.1.6. Kluczowe wnioski na temat regresji liniowej	271
7.2. Stosowanie regresji logistycznej	271
7.2.1. Mechanizm działania regresji logistycznej	272
7.2.2. Tworzenie modelu regresji logistycznej	276
7.2.3. Uzyskiwanie przewidywań	277
7.2.4. Wyszukiwanie relacji i wydobywanie użytecznych informacji z modeli logistycznych	282
7.2.5. Odczytywanie podsumowania modelu i charakteryzowanie współczynników	284
7.2.6. Kluczowe wnioski na temat regresji logistycznej	291
7.3. Regularyzacja	291
7.3.1. Przykład quasi-separacji	292
7.3.2. Rodzaje regresji regularyzowanej	296

7.3.3. Regresja regularyzowana przy użyciu pakietu glmnet	298
Podsumowanie	307
8. Zaawansowane przygotowywanie danych	309
8.1. Cel pakietu vtreat	310
8.2. Konkurs KDD i zestaw danych KDD Cup 2009	312
8.2.1. Pierwsze kroki z danymi KDD Cup 2009	313
8.2.2. Metoda „słonia w składzie porcelany”	315
8.3. Podstawowe przygotowywanie danych do zadań klasyfikacji	318
8.3.1. Ramka oceny zmiennej	319
8.3.2. Odpowiednie stosowanie planu naprawy	324
8.4. Zaawansowane przygotowywanie danych do zadań klasyfikacji	325
8.4.1. Korzystanie z metody mkCrossFrameCExperiment()	325
8.4.2. Budowanie modelu	328
8.5. Przygotowywanie danych do zadań regresji	332
8.6. Opanowanie pakietu vtreat	334
8.6.1. Fazy mechanizmu vtreat	335
8.6.2. Brakujące wartości	337
8.6.3. Zmienne wskaźnikowe	338
8.6.4. Kodowanie wpływu	339
8.6.5. Plan naprawy	341
8.6.6. Ramka krzyżowa	341
Podsumowanie	345
9. Metody nienadzorowane	347
9.1. Analiza skupień	348
9.1.1. Odległości	349
9.1.2. Przygotowanie danych	352
9.1.3. Hierarchiczna analiza skupień za pomocą funkcji hclust()	354
9.1.4. Algorytm centroidów	367
9.1.5. Przypisywanie nowych punktów do skupień	374
9.1.6. Kluczowe wnioski na temat analizy skupień	376
9.2. Reguły asocjacyjne	377
9.2.1. Przegląd reguł asocjacyjnych	377
9.2.2. Przykładowy problem	379
9.2.3. Wydobywanie reguł asocjacyjnych za pomocą pakietu arules	380
9.2.4. Kluczowe wnioski na temat reguł asocjacyjnych	388
Podsumowanie	388
10. Zaawansowane metody uczenia maszynowego	391
10.1. Metody drzewa	393
10.1.1. Podstawowe drzewo decyzyjne	394
10.1.2. Usprawnianie przewidywań za pomocą agregacji	397
10.1.3. Dalsze usprawnianie przewidywań za pomocą lasów losowych	399
10.1.4. Drzewa wzmacniane gradientowe	405
10.1.5. Kluczowe wnioski na temat modeli bazujących na drzewach	414

10.2. Wykrywanie relacji niemonotonicznych za pomocą uogólnionych modeli addytywnych	414
10.2.1. Mechanizm działania modelu GAM	415
10.2.2. Przykład regresji jednowymiarowej	415
10.2.3. Wydobywanie relacji nieliniowych	420
10.2.4. Stosowanie modelu GAM na rzeczywistych danych	422
10.2.5. Stosowanie modelu GAM w regresji logistycznej	425
10.2.6. Kluczowe wnioski na temat modelu GAM	427
10.3. Rozwiązywanie problemów „nierozdzielnych” 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. Opis Twojego rozwiązania	492
12.3.4. Omówienie wyników i przyszłych planów	492
12.3.5. Kluczowe wnioski na temat prezentacji przeznaczonej dla partnerów	493
Podsumowanie	494
Dodatek A Korzystanie z R i innych narzędzi	497
Dodatek B Ważne pojęcia z dziedziny statystyki	523
Dodatek C Bibliografia	559

oprac. BPK