

Spis treści

Od autora	IX
Sztuczna inteligencja	1
O książce	7
Dla kogo jest ta książka?	7
Narzędzia	8
Przykłady	9
Bibliografia	10
Konwencje i oznaczenia	10
1. Narzędzia	11
1.1. Język Python	12
1.2. Język R	12
1.2.1. Microsoft R Open (MRO)	14
1.2.2. Microsoft R Client (MRC)	14
1.2.3. Microsoft Machine Learning Server (MLS)	15
1.3. SQL Server 2019	15
1.3.1. Instalacja	16
1.3.2. Microsoft SQL Server Machine Learning Services	20
1.3.3. Bezpieczeństwo	29
1.3.4. Wydajność	32
1.4. PyCharm Community Edition	39
1.5. RStudio Desktop	40
1.6. Instalacja dodatkowych pakietów	42
1.7. Power BI Desktop	46
2. Praca z SQL Server Machine Learning Services	47
2.1. Wykrywanie oszustw	47
2.2. Klasyfikacja przejazdów	56
2.3. Dodatkowe funkcjonalności serwera SQL Server i usługi SQL Machine Learning Services	67
2.3.1. Zapytania predykcyjne czasu rzeczywistego	67
2.3.2. Natywne zapytania predykcyjne	69
2.3.3. Tworzenie modeli dla partycji danych	70
3. Wstęp do uczenia maszynowego	75
3.1. Rodzaje uczenia maszynowego	76
3.2. Proces uczenia	78
3.3. Modele regresji	84
3.4. Modele partycjonujące	98
3.5. Metodyka CRISP-DM	115
3.6. Metodyka TDSP	117
4. Zrozumienie danych	119

4.1. Poznanie danych wymaga zrozumienia postawionego problemu	120
4.2. Statystyki opisowe	121
4.2.1. Zmienne numeryczne	123
4.2.2. Zmienne kateryczne	133
4.3. Brakujące dane	134
4.4. Entropia	135
4.5. Ocena zmiennych za pomocą programu Power BI	136
4.6. Ocena zmiennych przy użyciu języka SQL	138
4.6.1. Automatyczny opis zmiennych	143
4.7. Wizualizacja zmiennych	147
4.8. Reprezentatywność danych	151
4.9. Korelacje między zmiennymi	153
4.9.1. Klątwa wymiarowości	154
4.9.2. Ocena przydatności zmiennych	157
4.9.3. Dwie zmienne numeryczne	158
4.9.4. Dwie zmienne kateryczne	160
4.9.5. Dwie zmienne porządkowe	161
4.9.6. Zmienna kateryczna i numeryczna	162
4.9.7. Korelacja oznacza współwystępowanie, a nie związek przyczynowo-skutkowy	163
4.10. Ocena korelacji za pomocą programu Power BI	164
4.11. Ocena korelacji przy użyciu języka SQL	166
5. Przygotowanie danych	169
5.1. Uporządkowanie danych	171
5.2. Wzbogacanie danych	176
5.2.1. Data i czas	182
5.3. Wyczyszczenie danych	183
5.3.1. Usuwanie brakujących wartości	183
5.3.2. Usuwanie duplikatów	186
5.3.3. Usuwanie błędnych danych	187
5.3.4. Usuwanie wartości odstających	188
5.4. Przekształcenie danych	190
5.4.1. Kodowanie	190
5.4.2. Generalizacja	193
5.4.3. Zaokrąglanie	195
5.4.4. Dyskretyzacja	195
5.4.5. Skalowanie	198
5.4.6. Wygładzanie	201
5.5. Redukcja danych	208
5.5.1. Selekcja zmiennych	208
5.5.2. Analiza składowych głównych	210
5.5.3. Wybór zmiennych przydatnych dla modelu	214
5.6. Podział danych	215
5.6.1. Podział warstwowy	216
5.6.2. Równoważenie danych	217
5.6.3. k-krotna walidacja krzyżowa	219
5.7. Danych walidacyjnych używa się do optymalizacji, a nie do oceny modeli	222
5.8. Kto miał szansę przeżyć katastrofę Titanica?	223

6. Analiza skupień	235
6.1. Grupowanie w celu zmniejszenia liczby przykładów	236
6.2. Algorytmy iteracyjno-optymalizacyjne	238
6.3. Segmentacja	245
7. Regresja	255
7.1. Zrozumienie problemu	255
7.2. Zrozumienie danych	257
7.3. Opisanie danych	263
7.4. Ocena przydatności danych	266
7.5. Wzbogacenie danych	267
7.6. Ocena zmiennych	268
7.7. Przekształcenie i wybór danych	269
7.8. Modelowanie	272
7.8.1. Uczenie na błędach	273
7.8.2. Regresja liniowa	280
7.8.3. Ogólny model liniowy CLM	285
7.8.4. Sztuczne sieci neuronowe	288
7.8.5. Drzewa regresyjne	305
7.8.6. Kombinacje modeli	309
7.9. Porównanie modeli	310
7.10. Wdrożenie modeli po stronie serwera SQL	314
8. Klasyfikacja	319
8.1. Klasyfikacja binarna	320
8.1.1. Przygotowanie danych	321
8.1.2. Regresja logistyczna	323
8.1.3. Sztuczne sieci neuronowe	329
8.1.4. Klasyfikacja przez indukcję drzew decyzyjnych	334
8.1.5. Kombinacje modeli	346
8.1.6. Porównanie modeli	350
8.1.7. Wdrożenie modeli po stronie serwera SQL	354
8.2. Klasyfikacja wieloklasowa	358
8.2.1. Przygotowanie danych	361
8.2.2. Regresja logistyczna	363
8.2.3. Sztuczne sieci neuronowe	366
8.2.4. Klasyfikacja przez indukcję drzew decyzyjnych	369
8.2.5. Porównanie modeli	375
8.2.6. Wdrożenie modeli po stronie serwera SQL	377
9. Ocena modeli	383
9.1. Nie ma darmowego lunchu	384
9.2. Błędy modeli predykcyjnych	385
9.2.1. Błąd systematyczny i błąd aproksymacji	387
9.3. Kryteria oceny modeli regresji	390
9.3.1. Średni błąd bezwzględny	391
9.3.2. Pierwiastek błędu średniokwadratowego	391
9.3.3. Znormalizowany błąd bezwzględny	392
9.3.4. Znormalizowany błąd kwadratowy	392
9.3.5. Współczynnik determinacji R^2	393
9.3.6. Ocena modelu prognozującego pozostały czas bezawaryjnej	

pracy urzędów	393
9.4. Kryteria oceny modeli klasyfikacji binarnej	398
9.4.1. Macierz błędów	399
9.4.2. Trafność	400
9.4.3. Precyzja	401
9.4.4. Czułość	401
9.4.5. F-miara	402
9.4.6. Współczynnik Kappa Cohena	402
9.4.7. Krzywa ROC i obszar pod krzywą	403
9.4.8. Ocena modelu klasyfikującego urządzenie jako wymagające lub niewymagające przeglądu	404
9.5. Kryteria oceny modeli klasyfikacji wieloklasowej	413
9.5.1. Macierz błędów	413
9.5.2. Metryki klasy większościowej	414
9.5.3. Metryki poszczególnych klas	414
9.5.4. Średnie makro	415
9.5.5. Średnie mikro	416
9.6. Ocena modelu klasyfikującego urządzenie do przeglądu	417
9.7. Interpretacja predykcji	423
10. Optymalizacja i wdrożenie modeli	427
10.1. Zrozumienie problemu	427
10.2. Zrozumienie i przygotowanie danych	428
10.2.1. Import danych	429
10.2.2. Ocena danych	434
10.3. Modelowanie	442
10.4. Optymalizacja	446
10.5. Wdrożenie	452
10.5.1. Analiza typu Co by było, gdyby?	456
Posłowie	461
Bibliografia	463