MARCIN SZELIGA

DATA SCIENCE
I UCZENIE MASZYNOWE

PWN

Spis treści

Wstęp XVII

O czym jest ta książka? XVII
Data science XVIII
Uczenie maszynowe XX
Dla kogo jest ta książka? XXI
Narzędzia XXII
Usługa Azure ML XXIII
Język R XXIV
Microsoft R Open XXV
Przykładowe dane XXVI
Konwencje i oznaczenia XXVI

1. Uczenie maszynowe jako element eksperymentów *data science* 1

1.1. Eksploracja danych jako technika wspomagania decyzji 2
1.2. Modelowanie 4
1.3. Wiedza i proces uczenia 6
1.4. Hipotezy 9
1.5. Założenia eksperymentu data science 10
1.6. Dwa typy analiz 12
1.7. *Data science* jako metoda naukowa 12
1.8. Przykładowy eksperyment – optymalizacja kampanii marketingowej 14
1.8.1. Zrozumienie problemu i określenie celów eksperymentu 15
1.8.2. Zrozumienie danych 16
1.8.3. Wstępne przetwarzanie danych 17
1.8.4. Modelowanie 18
1.8.5. Ocena 18
1.8.6. Wdrożenie 20
Podsumowanie 23

2. Ocena przydatności danych ... 25

2.1. Dane źródłowe ... 26
2.2. Zmienne ... 27
2.2.1. Rozkład częstości zmiennych ... 30
2.2.2. Graficzna prezentacja danych ... 42
2.2.3. Korelacje (związki między zmiennymi) ... 44
2.3. Reprezentatywność danych ... 50
2.4. Duplikaty ... 54
2.5. Szeregi czasowe ... 56
Podsumowanie ... 63

3. Wstępne przetwarzanie danych ... 65

3.1. Uzupełnianie brakujących danych ... 66
3.2. Poprawianie błędnych danych ... 71
3.3. Zmienne numeryczne ... 71
3.3.1. Instalowanie dodatkowych bibliotek *R* w Azure ML ... 72
3.3.2. Wartości nietypowe (odstające) ... 73
3.3.3. Normalizacja ... 75
3.3.4. Dyskretyzacja ... 77
3.4. Zmienne kategoryczne ... 78
3.4.1. Problem jakości danych tekstowych ... 79
3.4.2. Uogólnienie (generalizacja) ... 80
3.4.3. Numerowanie stanów ... 81
3.4.4. Zmienne porządkowe ... 83
3.5. Szeregi czasowe ... 83
3.6. Wyrażenia języka naturalnego ... 89
3.7. Redukcja wymiarów ... 94
3.7.1. Usuwanie zmiennych na podstawie ich zdolności predykcyjnych ... 95
3.7.2. Analiza głównych składowych (PCA) ... 97
Podsumowanie ... 99

4. Wzbogacanie danych ... 101

4.1. Równoważenie danych ... 102
4.1.1. Usunięcie części przykładów większościowych ... 103
4.1.2. Nadpróbkowanie ... 104
4.2. Zmienne wyliczeniowe ... 106
4.3. Zastąpienie zmiennych wspólnym rozkładem prawdopodobieństwa ... 108
4.4. Wydzielenie danych testowych ... 111
4.4.1. Szeregi czasowe ... 115
4.4.2. Modele rekomendujące ... 116
4.4.3. Modele wykrywania oszustw ... 116
4.5. Wzorzec eksperymentu *data science* ... 116
Podsumowanie ... 117

5. Klasyfikacja ... 119

5.1. Klasyfikacja poprzez indukcję drzew decyzyjnych ... 121
5.1.1. Drzewa decyzyjne – definicja ... 121
5.1.2. Pojedyncze drzewa decyzyjne ... 124
5.1.3. Kombinacje drzew decyzyjnych ... 126

5.2. Klasyfikacja z użyciem maszyny wektorów nośnych ... 141
5.2.1. Przetwarzanie języka naturalnego przy użyciu maszyny wektorów nośnych ... 143
5.2.2. Modele maszyny wektorów nośnych i lokalnie głębokiej maszyny wektorów nośnych ... 152
5.3. Klasyfikacja probabilistyczna ... 153
5.3.1. Sieć Bayesa ... 157
5.3.2. Maszyna punktów Bayesa ... 159
5.4. Inne klasyfikatory dostępne w Studiu Azure ML ... 161
5.4.1. Inne klasyfikatory – omówienie ... 161
5.4.2. Modele eksploracji danych w języku R ... 163
5.5. Klasyfikatory binarne a klasyfikacja wieloklasowa ... 164
5.6. Wykrywanie oszustw jako przykład klasyfikacji binarnej ... 167
5.6.1. Oznaczenie obserwacji ... 167
5.6.2. Zrównoważenie danych i wydzielenie danych testowych ... 169
5.6.3. Wzbogacenie danych ... 169
Podsumowanie ... 172

6. Regresja ... 173

6.1. Model regresji wielorakiej ... 179
6.1.1 Wieloraka regresja liniowa ... 181
6.1.2. Estymacja bayesowska modelu regresji liniowej ... 183
6.2. Zmienne kategoryczne w modelach regresji ... 185
6.2.1. Regresja Poissona ... 186
6.2.2. Regresja porządkowa ... 188
6.3. Regresja kwantylowa ... 188
6.4. Regresja poprzez indukcję drzew decyzyjnych ... 191
6.5. Sztuczne sieci neuronowe ... 193
6.5.1. Perceptron ... 198
6.5.2. Sieci neuronowe a regresja ... 200
6.5.3. Metody minimalizacji błędu ... 202
6.5.4. Wsteczna propagacja błędów ... 203
6.5.5. Regresja z użyciem sieci neuronowej ... 205
6.5.6. Głębokie sieci neuronowe ... 209
Podsumowanie ... 218

7. Grupowanie (analiza skupień) ... 221

7.1. Na czym polega grupowanie ... 221
7.2. Algorytmy grupowania ... 225
7.2.1. Grupowanie hierarchiczne ... 226
7.2.2. Grupowanie iteracyjno-optymalizacyjne ... 231
7.3. Grupowanie w celu znajdowania podobnych obiektów ... 236
7.4. Grupowanie w celu kompresji ... 239
7.5. Wykrywanie anomalii ... 240
Podsumowanie ... 244

8. Rekomendowanie ... 245

8.1. Systemy rekomendujące ... 245
8.2. Odkrywanie asocjacji ... 250
8.3. Model Matchbox Recommender ... 258
8.3.1. Rekomendowanie przez filtrowanie kolektywne ... 258
8.3.2. Rekomendowanie przez filtrowanie cech przedmiotów i użytkowników (hybrydowe) ... 267
Podsumowanie ... 269

9. Prognozowanie ... 271

9.1. Szeregi czasowe ... 272
9.2. Naiwne metody prognozowania ... 274
9.3. Modele średniej ważonej ... 274
9.4. Modele *ARIMA* ... 283
9.5. Modele nieliniowe ... 288
9.6. Prognozowanie w Studiu Azure ML ... 290
Podsumowanie ... 292

10. Ocena i poprawa jakości modeli ... 293

10.1. Reguła powrotu do średniej ... 293
10.2. Kryteria oceny modeli eksploracji danych ... 295
10.2.1. Łatwość interpretacji ... 296
10.2.2. Trafność ... 296
10.2.3. Wiarygodność ... 297
10.2.4. Wydajność i skalowalność ... 297
10.2.5. Przydatność ... 297
10.3. Ocena jakości modeli klasyfikacyjnych ... 298
10.3.1. Moduł *Evaluate Model* ... 298
10.3.2. Macierz pomyłek ... 299
10.3.3. Krzywa ROC ... 302
10.3.4. Wykres precyzja w funkcji czułości i wykres zysku ... 304
10.3.5. Trafność klasyfikacji ... 305
10.3.6. Klasyfikatory wieloklasowe ... 307
10.4. Ocena jakości modeli regresyjnych ... 308
10.4.1. Miary oceny modeli ... 308
10.4.2. Walidacja krzyżowa ... 310
10.5. Ocena jakości modeli grupujących ... 313
10.6. Ocena jakości modeli rekomendujących ... 315
10.7. Ocena jakości modeli prognozujących ... 317
10.8. Porównanie jakości modeli ... 322
10.9. Poprawa jakości modeli ... 326
10.9.1. Automatyczna poprawa jakości modeli uczenia nadzorowanego ... 326
10.9.2. Znalezienie optymalnej liczby klastrów ... 330
10.10. Cykl życia eksperymentu *data science* ... 333
Podsumowanie ... 334

11. Publikacja modeli eksploracji danych jako usług WWW ... 339

11.1. Wzorcowy eksperyment *data science* ... 340
11.2. Predykcyjne usługi WWW ... 345
11.2.1. Zapytania predykcyjne ad-hoc ... 348
11.2.2. Wsadowe zapytania predykcyjne ... 349
Podsumowanie ... 352

Bibliografia ... 353

Dodatek A ... 361

Dodatek B ... 367