
Inhaltsverzeichnis

Teil I Datenanalyseprozess

1	Fach- und Geschäftsverständnis	3
1.1	Die Neuen	3
1.2	Einführung	4
1.2.1	Data-Science-Begriffe	5
1.2.2	Datenschutz	8
1.2.3	Datensicherheit	10
1.2.4	Datenethik	11
1.3	Auf Leben und Tod	12
1.4	Praktischer Teil	17
1.5	Kaltes Wasser	32
2	Datenverständnis	35
2.1	König der Welt	35
2.2	Daten	38
2.2.1	Datentypen	39
2.2.2	Daten- und Dateiformate	42
2.2.3	Skalenniveaus	43
2.2.4	Datenqualität	45
2.2.5	Explorative Datenanalyse	46
2.2.6	Korrelationsanalyse	48
2.3	Die passenden Fragen stellen	49
2.4	Die ersten Analysen	50
2.5	Es leuchtet blau	53
3	Standard-Prozessmodell	83
3.1	Schön knusprig	83
3.2	CRISP-DM	84
3.2.1	Business Understanding	86
3.2.2	Data Understanding	87
3.2.3	Data Preparation	88
		XIII

3.2.4	Modeling	89
3.2.5	Evaluation	90
3.2.6	Deployment	91
3.3	Frauen und Kinder zuerst	92
3.4	Der erste Prozessdurchlauf	97
3.5	Frauen und Mädchen zuerst	99

Teil II Sprints

4	Modellierung mit k-Nearest Neighbor	105
4.1	Glück und Unglück sind zwei Nachbarn	105
4.2	Modeling	107
4.2.1	k-Nearest Neighbor	108
4.2.2	Partitionierung und Sampling	113
4.2.3	Kreuzvalidierung	114
4.3	Ein guter Nachbar	116
4.4	Lazy Learner	122
4.5	Einer lernt's von seinem Nachbar	124
5	Datenvorbereitung	129
5.1	Die heiße Eins	129
5.2	Feature Engineering	132
5.2.1	Feature Encoding	133
5.2.2	Feature Binning	136
5.2.3	Skalierungen und Transformationen	138
5.2.4	Textoperationen	141
5.3	Das ist kein Bug	142
5.4	Datenvarianten	151
5.5	Blut ist dicker als Wasser	156
6	Auswertung und Einsatz der Ergebnisse	159
6.1	Spiced Ham	159
6.2	Gütemaße der binären Klassifikation	160
6.2.1	Genauigkeit und Fehlerrate	162
6.2.2	Konfusionsmatrix	163
6.2.3	Sensitivität und Spezifität	163
6.2.4	Präzision und Recall	164
6.2.5	Komplexe Gütemaße	165
6.2.6	Bezeichnungen der Gütemaße	167
6.3	Entwerrung	168
6.4	Sophias Welt	171
6.5	Mindestziel	175

Teil III Klassifikation: Methoden

7	Naive Bayes	181
7.1	Das Ziegenproblem	181
7.2	Rechnen mit Wahrscheinlichkeiten	183
7.2.1	Wahrscheinlichkeiten und der Satz von Bayes	183
7.2.2	Maximum-a-posteriori-Hypothese	187
7.2.3	Naive-Bayes-Klassifizierer	189
7.2.4	Vor- und Nachteile	194
7.3	Schnelles Denken, langsames Denken	196
7.4	Basis-Setup	197
7.5	Naive Ergebnisse	205
8	Entscheidungsbäume	211
8.1	Chaos	211
8.2	Decision Tree	213
8.2.1	Beispiel	214
8.2.2	Konstruktion mit Entropie	215
8.2.3	Vor- und Nachteile	224
8.2.4	Overfitting und Pruning	224
8.3	Gratis-Mittagessen	225
8.4	Tuning	227
8.5	White Box	229
9	Neuronale Netze	237
9.1	Bergnebel	237
9.2	Lernen mit Neuronen	239
9.2.1	Biologisches Neuron	239
9.2.2	Künstliches Neuron	241
9.2.3	Multilayer-Perzeptron	245
9.2.4	Backpropagation	247
9.2.5	Gradientenverfahren	249
9.2.6	Vor- und Nachteile	253
9.3	Faustregeln	254
9.4	Topologiestudien	255
9.5	Shallow Learning	257
10	Lineare Klassifizierer	265
10.1	Nobelpreis	265
10.2	Unterstützung und Trennung	268
10.2.1	Support Vector Machine	268
10.2.2	Logistische Regression	275
10.3	Sabotage	278

10.4	Kernel, Penalty und Solver	280
10.5	Lineare Separierbarkeit	282
11	Ensemble Learning	287
11.1	Wisdom of the Crowd	287
11.2	Ensemble-Methoden	289
11.2.1	Verzerrung-Varianz-Dilemma	290
11.2.2	Ensembles, Bagging und Random Forest	293
11.2.3	Boosting und Gradient Boosted Trees	297
11.3	Ransomware	302
11.4	Feature Importance	303
11.5	Sex-Fare	305
 Teil IV Regression		
12	Vorbereitung der Regressionsanalyse	315
12.1	Junior	315
12.2	Einführung in die Regressionsanalyse	317
12.2.1	Interpolation vs. Extrapolation	317
12.2.2	Korrelationsanalyse	318
12.2.3	Korrelation vs. Regression	320
12.3	Zirkelschluss	322
12.4	FirstnameMeanAge	324
12.5	Doppelgänger	326
13	Lineare und polynomiale Regression	333
13.1	Code Red	333
13.2	Regression und Regularisierung	335
13.2.1	Methode der kleinsten Quadrate	336
13.2.2	Lineare Regression und der T-Test	337
13.2.3	Polynomiale Regression	339
13.2.4	Güte- bzw. Fehlermaße	339
13.2.5	Vor- und Nachteile	340
13.2.6	Regularisierung	341
13.3	Einfache Modelle	344
13.4	Regressionsmodelle	347
13.5	Dr. Strangelove	349
14	Bekannte Methoden	355
14.1	Gekko	355
14.2	Von Bäumen und Netzen	358
14.2.1	Regression Tree	358
14.2.2	Random Forest und Gradient Boosted Trees	361

14.2.3	Multilayer-Perzeptron	361
14.3	Suspect	362
14.4	Regressoren	362
14.5	Lebenslanges Lernen	364

Teil V Klassifikation vs. Regression

15	Mehrklassen-Klassifikation	373
15.1	Kickern	373
15.2	Viele Klassen	374
15.2.1	Klassifizierungsstrategien	374
15.2.2	Mikro- und Makro-Mittelwertbildungen	377
15.2.3	LogLoss	380
15.2.4	Softmax	381
15.3	AgeBinned	382
15.4	MKK	387
15.5	Champions	389
16	Ungleiche Verteilungen	395
16.1	Marx oder May	395
16.2	Sampling-Methoden	397
16.2.1	Undersampling vs. Oversampling	398
16.2.2	Bootstrapping und SMOTE	399
16.2.3	Gegenüberstellung	402
16.3	Research	405
16.4	Child	408
16.5	Best results ever	410
17	Konsolidierung	417
17.1	Repetitorium	417
17.2	Rekapitulation	421
17.3	Outcome	424
17.4	Ergebnisse zusammenführen	425
17.5	Neue Erkenntnisse	427

Teil VI Fortgeschrittene Techniken

18	Feature Selection	433
18.1	Dreimal schwarzer Kater	433
18.2	Auswahlstrategien	436
18.2.1	Filter	436
18.2.2	Eingebettete Methoden	437
18.2.3	Wrapper	437

18.3	Maulwurf	439
18.4	Selection Loop	440
18.5	Top Features	442
19	Hyperparameteroptimierung	451
19.1	Ambidextrie	451
19.2	Parameter-Tuning	453
19.2.1	Rastersuche	453
19.2.2	Zufallssuche	454
19.2.3	Sukzessive Halbierung	455
19.2.4	Bayes'sche Optimierung	456
19.2.5	Weitere Optimierungsverfahren	456
19.3	Max-Test	457
19.4	Parametersuche	458
19.5	Die Nadel im Heuhaufen	462
20	Präzision	469
20.1	Intervalltraining	469
20.2	Vertrauenswahrscheinlichkeit	470
20.2.1	Schwellenwert	471
20.2.2	Konstruktion der ROC	473
20.3	Kryptonite	476
20.4	Threshold Tuning	478
20.5	Konfidenz	479
 Teil VII Abschluss		
21	Freestyle	487
21.1	Plan A	487
21.2	Karls Weg	488
21.3	Verschwörung	491
21.4	Annas Weg	492
21.5	Plan B	495
22	Modell 6	497
22.1	Zurück auf Los	497
22.2	Aufbruch	499
22.3	Deepfake	500
22.4	Heureka!	501
22.5	Serendipity	503

23	Ausblick	507
23.1	Gartenparty	507
23.2	Fallstudien	508
23.2.1	Binäre Klassifikation	510
23.2.2	Mehrklassen-Klassifikation	510
23.2.3	Regression	511
23.3	Zugriff	512
23.4	Datensammlungen	513
23.5	Anaconda	514
A Softwarewerkzeuge		517
Fallstudie		519
Datenversionen		523
Mathematische Grundlagen		527
Statistische Grundlagen		535
Gütemaße		539
Sophias Welt		541
Max' Materialien		543
Karls Weg		547
Annas Weg		563
Glossar		581
Stichwortverzeichnis		589