

# Inhaltsverzeichnis

<b>Einleitung</b> . . . . .	13
<b>Über die Autoren</b> . . . . .	21
<b>1 Einführung: Datenanalytisches Denken</b> . . . . .	23
1.1 Allgegenwärtige Datenerfassungsmöglichkeiten . . . . .	23
1.2 Beispiel: Hurrikan Frances . . . . .	25
1.3 Beispiel: Vorhersage der Kundenfluktuation . . . . .	26
1.4 Data Science, Engineering und datengestützte Entscheidungsfindung . . . . .	27
1.5 Datenverarbeitung und »Big Data« . . . . .	31
1.6 Von Big Data 1.0 zu Big Data 2.0 . . . . .	32
1.7 Daten und Data-Science-Fähigkeiten als strategisches Gut . . . . .	33
1.8 Datenanalytische Denkweise . . . . .	36
1.9 Dieses Buch . . . . .	38
1.10 Data Mining und Data Science . . . . .	39
1.11 In der Chemie geht es nicht um Reagenzgläser: Data Science vs. die Aufgabe des Data Scientists . . . . .	40
1.12 Zusammenfassung . . . . .	41
<b>2 Geschäftliche Aufgaben und Data-Science-Lösungen</b> . . . . .	43
2.1 Von geschäftlichen Aufgaben zum Data-Mining-Verfahren . . . . .	44
2.2 Überwachte vs. unüberwachte Verfahren . . . . .	49
2.3 Ergebnisse des Data Minings . . . . .	51
2.4 Der Data-Mining-Prozess . . . . .	52
2.4.1 Aufgabenverständnis . . . . .	53
2.4.2 Datenverständnis . . . . .	54
2.4.3 Datenaufbereitung . . . . .	56
2.4.4 Modellbildung . . . . .	57
2.4.5 Beurteilung . . . . .	57
2.4.6 Einsatz . . . . .	59
2.5 Auswirkungen auf das Management des Data-Science-Teams . . . . .	61
2.6 Weitere Analyseverfahren und -Technologien . . . . .	62
2.6.1 Statistik . . . . .	63
2.6.2 Datenbankabfragen . . . . .	65

2.6.3	Data Warehouses .....	66
2.6.4	Regressionsanalyse .....	67
2.6.5	Machine Learning und Data Mining .....	67
2.6.6	Geschäftliche Aufgaben durch diese Verfahren lösen .....	68
2.7	Zusammenfassung .....	70
<b>3</b>	<b>Einführung in die Vorhersagemodellbildung: Von der Korrelation zur überwachten Segmentierung .....</b>	<b>71</b>
3.1	Modelle, Induktion und Vorhersage .....	73
3.2	Überwachte Segmentierung .....	77
3.2.1	Auswahl informativer Merkmale .....	78
3.2.2	Beispiel: Merkmalsauswahl anhand des Informationsgewinns .....	86
3.2.3	Überwachte Segmentierung mit Baumstrukturmodellen .....	92
3.3	Segmentierungen visualisieren .....	98
3.4	Bäume als Regelsätze .....	100
3.5	Wahrscheinlichkeitsabschätzung .....	101
3.6	Beispiel: Abwanderungsrate per Entscheidungsbaum ermitteln .....	104
3.7	Zusammenfassung .....	108
<b>4</b>	<b>Ein Modell an Daten anpassen .....</b>	<b>111</b>
4.1	Klassifizierung via mathematischer Funktionen .....	113
4.1.1	Lineare Diskriminanzfunktion .....	115
4.1.2	Optimieren der Zielfunktion .....	118
4.1.3	Beispiel: Extraktion einer linearen Diskriminanzfunktion aus Daten .....	119
4.1.4	Lineare Diskriminanzfunktionen zur Beurteilung und zum Erstellen einer Rangfolge von Instanzen .....	121
4.1.5	Support Vector Machines kompakt erklärt .....	122
4.2	Regression via mathematischer Funktionen .....	125
4.3	Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit und logistische »Regression« .....	127
4.3.1	* Logistische Regression: Technische Details .....	131
4.4	Beispiel: Logistische Regression vs. Entscheidungsverfahren .....	134
4.5	Nichtlineare Funktionen, Support Vector Machines und neuronale Netze .....	138
4.6	Zusammenfassung .....	141

<b>5</b>	<b>Überanpassung erkennen und vermeiden</b> .....	<b>143</b>
5.1	Verallgemeinerungsfähigkeit .....	143
5.2	Überanpassung .....	145
5.3	Überanpassung im Detail .....	146
	5.3.1 Zurückgehaltene Daten und Fitfunktionen .....	146
	5.3.2 Überanpassung bei Entscheidungsbaumverfahren .....	149
	5.3.3 Überanpassung bei mathematischen Funktionen .....	151
5.4	Beispiel: Überanpassung linearer Funktionen .....	152
5.5	* Beispiel: Nachteile der Überanpassung .....	156
5.6	Von der Beurteilung durch Testdatenmengen zur Kreuzvalidierung .....	159
5.7	Abwanderungsdaten .....	163
5.8	Lernkurven .....	165
5.9	Überanpassung vermeiden und Steuerung der Komplexität .....	167
	5.9.1 Überanpassung von Entscheidungsbäumen vermeiden .....	167
	5.9.2 Eine allgemeine Methode zur Vermeidung von Überanpassung .....	168
	5.9.3 * Überanpassung bei der Parameteroptimierung vermeiden .....	171
5.10	Zusammenfassung .....	175
<b>6</b>	<b>Ähnlichkeit, Nachbarn und Cluster</b> .....	<b>177</b>
6.1	Ähnlichkeit und Distanz .....	178
6.2	Nächste-Nachbarn-Methoden .....	181
	6.2.1 Beispiel: Whisky-Analyse .....	181
	6.2.2 Nächste Nachbarn und Vorhersagemodelle .....	184
	6.2.3 Anzahl der Nachbarn und ihre Gewichtung .....	187
	6.2.4 Geometrische Interpretation, Überanpassung und Steuerung der Komplexität .....	189
	6.2.5 Probleme mit Nächste-Nachbarn-Methoden .....	193
6.3	Ähnlichkeit und Nachbarn: Wichtige technische Details .....	196
	6.3.1 Heterogene Merkmale .....	196
	6.3.2 * Weitere Distanzmaße .....	197
	6.3.3 * Zusammenfassende Funktionen: Scores der Nachbarn berechnen .....	200
6.4	Clustering .....	202
	6.4.1 Beispiel: Weitere Whisky-Analysen .....	203
	6.4.2 Hierarchisches Clustering .....	204
	6.4.3 Nächste Nachbarn: Clustering um Zentroiden .....	209

6.4.4	Beispiel: Clustering von Wirtschaftsnachrichten . . . . .	214
6.4.5	Das Ergebnis des Clusterings verstehen . . . . .	218
6.4.6	* Cluster-Beschreibungen durch überwachtes Lernen erzeugen . . . . .	220
6.5	Lösen von geschäftlichen Aufgaben vs. Datenerkundung . . . . .	223
6.6	Zusammenfassung . . . . .	226
<b>7</b>	<b>Entscheidungsanalyse I: Was ist ein gutes Modell?</b> . . . . .	<b>227</b>
7.1	Beurteilung von Klassifizierern . . . . .	228
7.1.1	Korrektklassifizierungsrate und damit verbundene Probleme . . . . .	229
7.1.2	Die Wahrheitsmatrix . . . . .	230
7.1.3	Klassifizierungsaufgaben mit unausgewogener Klassenverteilung . . . . .	230
7.1.4	Klassifizierungsaufgaben mit unausgewogenem Kosten-Nutzen-Verhältnis . . . . .	233
7.2	Verallgemeinerung über Klassifizierungen hinaus . . . . .	234
7.3	Ein wichtiges analytisches Tool: Der Erwartungswert . . . . .	235
7.3.1	Erwartungswerte für Klassifizierer verwenden . . . . .	236
7.3.2	Erwartungswerte zur Beurteilung von Klassifizierern verwenden . . . . .	238
7.4	Beurteilung, Leistung und die Folgen für Investitionen in Daten . . . . .	246
7.5	Zusammenfassung . . . . .	249
<b>8</b>	<b>Visualisierung der Leistung von Modellen</b> . . . . .	<b>251</b>
8.1	Rangfolge statt Klassifizierung . . . . .	252
8.2	Profitkurven . . . . .	254
8.3	ROC-Diagramme und -Kurven . . . . .	257
8.4	Die Fläche unter der ROC-Kurve . . . . .	263
8.5	Kumulative Reaktionskurven und Lift-Kurven . . . . .	263
8.6	Beispiel: Leistungsanalyse . . . . .	266
8.7	Zusammenfassung . . . . .	275
<b>9</b>	<b>Evidenz und Wahrscheinlichkeiten</b> . . . . .	<b>277</b>
9.1	Beispiel: Gezielte Kundenansprache durch Onlinewerbung . . . . .	277
9.2	Evidenzen probabilistisch kombinieren . . . . .	280
9.2.1	Verbundwahrscheinlichkeit und Unabhängigkeit . . . . .	281
9.2.2	Der Satz von Bayes . . . . .	282
9.3	Anwendung des Satzes von Bayes in der Data Science . . . . .	284
9.3.1	Bedingte Unabhängigkeit und naive Bayes-Klassifizierung . . . . .	286

	9.3.2	Vor- und Nachteile des naiven Bayes-Klassifizierers . . . . .	288
9.4		Ein Modell für den Lift der Evidenz . . . . .	290
9.5		Beispiel: Lifts der Evidenz von Facebooks-Likes . . . . .	291
	9.5.1	Evidenz in Aktion: Gezielte Kundenansprache durch Werbung . . . . .	293
9.6		Zusammenfassung . . . . .	294
<b>10</b>		<b>Texte repräsentieren und auswerten . . . . .</b>	<b>295</b>
10.1		Die Bedeutung von Text . . . . .	296
10.2		Probleme bei der Auswertung von Text. . . . .	297
10.3		Repräsentierung . . . . .	298
	10.3.1	Das Bag-of-words-Modell. . . . .	298
	10.3.2	Vorkommenshäufigkeiten . . . . .	299
	10.3.3	Inverse Dokumenthäufigkeit. . . . .	302
	10.3.4	Die Kombination aus Vorkommenshäufigkeit und inverser Dokumenthäufigkeit: TFIDF . . . . .	303
10.4		Beispiel: Jazzmusiker . . . . .	304
10.5		* Der Zusammenhang zwischen IDF und Entropie. . . . .	308
10.6		Jenseits des Bag-of-words-Modells . . . . .	310
	10.6.1	N-Gramme . . . . .	310
	10.6.2	Eigennamenerkennung . . . . .	311
	10.6.3	Topic Models. . . . .	312
10.7		Beispiel: Auswertung von Wirtschaftsnachrichten zwecks Vorhersage von Börsenkursen . . . . .	313
	10.7.1	Die Aufgabe . . . . .	314
	10.7.2	Die Daten . . . . .	316
	10.7.3	Datenvorverarbeitung. . . . .	319
	10.7.4	Ergebnisse. . . . .	320
10.8		Zusammenfassung . . . . .	324
<b>11</b>		<b>Entscheidungsanalyse II: Analytisches Engineering . . . . .</b>	<b>325</b>
11.1		Auswahl geeigneter Empfänger eines Spendenaufrufs . . . . .	326
	11.1.1	Erwartungswerte: Zerlegung in Teilaufgaben und Kombination der Teilergebnisse . . . . .	326
	11.1.2	Ein kurzer Exkurs zum Thema Auswahleffekte . . . . .	328
11.2		Eine noch ausgeklügeltere Vorhersage der Kundenabwanderung . . . . .	329
	11.2.1	Erwartungswerte: Strukturierung einer komplizierteren geschäftlichen Aufgabe . . . . .	330
	11.2.2	Den Einfluss des Anreizes beurteilen. . . . .	331

11.2.3	Von der Zerlegung eines Erwartungswerts zur Data-Science-Lösung .....	333
11.3	Zusammenfassung .....	336
<b>12</b>	<b>Weitere Verfahren und Methoden der Data Science.</b> .....	<b>339</b>
12.1	Gleichzeitiges Auftreten und Assoziationen: Zueinander passende Objekte finden .....	340
12.1.1	Unerwartetheit messen: Lift und Leverage .....	341
12.1.2	Beispiel: Bier und Lotterielose .....	342
12.1.3	Assoziationen von Facebook-Likes .....	343
12.2	Profiling: Typisches Verhalten erkennen .....	347
12.3	Verknüpfungsvorhersagen und Kontaktempfehlungen .....	352
12.4	Datenreduzierung, latente Informationen und Filmempfehlungen .....	354
12.5	Bias, Varianz und Ensemblemethoden .....	358
12.6	Datengestützte Kausalmodelle und ein Beispiel für virales Marketing .....	362
12.7	Zusammenfassung .....	363
<b>13</b>	<b>Data Science und Geschäftsstrategie</b> .....	<b>365</b>
13.1	Datenanalytische Denkweise .....	365
13.2	Durch Data Science Wettbewerbsvorteile erzielen .....	368
13.3	Durch Data Science erzielte Wettbewerbsvorteile bewahren .....	369
13.3.1	Vorteile durch historische Gegebenheiten .....	370
13.3.2	Einzigartiges geistiges Eigentum .....	370
13.3.3	Einzigartige immaterielle Werte .....	371
13.3.4	Überlegene Data Scientists .....	371
13.3.5	Überlegenes Data-Science-Management .....	373
13.4	Gewinnung und Förderung von Data Scientists und ihren Teams .....	375
13.5	Data-Science-Fallstudien .....	377
13.6	Kreative Ideen von beliebigen Quellen übernehmen .....	378
13.7	Beurteilung von Vorschlägen für Data-Science-Projekte .....	379
13.7.1	Beispiel für einen Data-Mining-Projektvorschlag .....	379
13.7.2	Mängel des Projektvorschlags von Big Red .....	380
13.8	Ausgereifte Data Science .....	382
<b>14</b>	<b>Schlussfolgerungen</b> .....	<b>385</b>
14.1	Die fundamentalen Konzepte der Data Science .....	385
14.1.1	Anwendung der fundamentalen Konzepte auf eine neue Aufgabe: Auswertung der Daten von Mobilgeräten .....	388

14.1.2	Eine neue Sichtweise auf die Lösung von geschäftlichen Aufgaben.....	391
14.2	Was Daten nicht leisten können: Der menschliche Faktor .....	392
14.3	Privatsphäre, Ethik und Auswertung der Daten von Einzelpersonen .....	396
14.4	Data Science: Steckt noch mehr dahinter? .....	397
14.5	Ein letztes Beispiel: Vom Crowd-Sourcing zum Cloud-Sourcing ...	398
14.6	Schlussworte .....	400
<b>A</b>	<b>Leitfaden zur Beurteilung von Projektvorschlägen.....</b>	<b>401</b>
A.1	Aufgaben- und Datenverständnis.....	401
A.2	Datenaufbereitung.....	402
A.3	Modellbildung .....	403
A.4	Beurteilung und Deployment.....	403
<b>B</b>	<b>Ein weiteres Beispiel für einen Projektvorschlag .....</b>	<b>405</b>
B.1	Szenario und Projektvorschlag.....	405
B.2	Mängel des Projektvorschlags von GGC .....	406
	<b>Glossar .....</b>	<b>409</b>
	<b>Quellenverzeichnis .....</b>	<b>415</b>
	<b>Stichwortverzeichnis .....</b>	<b>423</b>