

# Contents

Foreword	xii
Figures	xiii
<b>1 Introduction</b>	<b>1</b>
1.1 The Age of Data: Is It Just a Hype? . . . . .	1
1.2 Why Is Data Science Relevant Now? . . . . .	2
1.3 Why Data Science with R? . . . . .	4
1.4 Who Is This Book For? . . . . .	4
1.5 Is It Possible to Learn Data Science without Math? . . . . .	5
1.6 How to Use This Book . . . . .	6
<b>2 Machine Learning, Data Science, and Artificial Intelligence</b>	<b>7</b>
2.1 Learning from History – All Just Hype? . . . . .	7
2.1.1 Data and Machines before the Dawn of AI . . . . .	7
2.1.2 The First Spring of Artificial Intelligence . . . . .	10
2.1.3 The First AI Winter . . . . .	10
2.1.4 The Second AI Spring: Expert Systems . . . . .	11
2.1.5 The Second AI Winter . . . . .	12
2.1.6 Is This a New AI Spring? . . . . .	12
2.1.7 Setbacks and New Hopes . . . . .	12
2.1.8 Technological Singularity: Do Machines Have a Mind? . . . . .	13
2.1.9 Alan Turing and the Turing Test . . . . .	14
2.2 Definitions . . . . .	15
2.2.1 Machine Learning . . . . .	15
2.2.2 Artificial Intelligence . . . . .	16
2.2.3 Data Science . . . . .	16
2.2.4 Data Analysis and Statistics . . . . .	18
2.2.5 Big Data . . . . .	18
<b>3 The Anatomy of a Data Science Project</b>	<b>21</b>
3.1 The General Flow of a Data Science Project . . . . .	21
3.1.1 The CRISP-DM Stages . . . . .	21
3.1.2 ASUM-DM . . . . .	23
3.1.3 The Data Science Workflow According to Hadley Wickham . . . . .	24
3.1.4 Which Approach Is Right for Me? . . . . .	25

3.2	Business Understanding: What Is the Problem to Be Solved?	25
3.2.1	Senior Management Support and Involvement of the Specialist Department . . . . .	25
3.2.2	Understanding Requirements . . . . .	26
3.2.3	Overcoming Resistance: Who Is Afraid of the Evil AI?	27
3.3	Basic Approaches in Machine Learning . . . . .	28
3.3.1	Supervised, Unsupervised, and Reinforcement Learning	28
3.3.2	Feature Engineering . . . . .	29
3.4	Performance Measurement . . . . .	29
3.4.1	Test and Training Data . . . . .	29
3.4.2	Not all Errors Are Created Equal: False Positives and False Negatives . . . . .	30
3.4.3	Confusion Matrix . . . . .	32
3.4.4	ROC AUC . . . . .	32
3.4.5	Precision Recall Curve . . . . .	34
3.4.6	Impact Outside the Lab . . . . .	34
3.4.7	Data Science ROI . . . . .	36
3.5	Communication with Stakeholders . . . . .	36
3.5.1	Reporting . . . . .	36
3.5.2	Storytelling . . . . .	37
3.6	From the Lab to the World: Data Science Applications in Production . . . . .	38
3.6.1	Data Pipelines and Data Lakes . . . . .	38
3.6.2	Integration with other Systems . . . . .	38
3.7	Typical Roles in a Data Science Project . . . . .	38
3.7.1	Data Scientist . . . . .	39
3.7.2	Data Engineer . . . . .	40
3.7.3	Data Science Architect . . . . .	40
3.7.4	Business Intelligence Analyst . . . . .	40
3.7.5	The Subject Matter Expert . . . . .	40
3.7.6	Project Management . . . . .	41
3.7.7	Citizen Data Scientist . . . . .	43
3.7.8	Other Roles . . . . .	43
<b>4</b>	<b>Introduction to R</b>	<b>45</b>
4.1	R: Free, Portable, and Interactive . . . . .	45
4.1.1	History . . . . .	47
4.1.2	Extension with Packages . . . . .	47
4.1.3	The IDE RStudio . . . . .	48
4.1.4	R versus Python . . . . .	49
4.1.5	Other Languages . . . . .	50
4.2	Installation and Configuration of R and RStudio . . . . .	51
4.2.1	Installation of R and Short Functional Test . . . . .	51
4.2.2	RStudio Installation . . . . .	53
4.2.3	Configuration of R and RStudio . . . . .	54

4.2.4	A Tour of RStudio . . . . .	59
4.2.5	Projects in RStudio . . . . .	63
4.2.6	The Cloud Alternative: Posit Cloud . . . . .	64
4.3	First Steps with R . . . . .	65
4.3.1	Everything in R Is an Object . . . . .	65
4.3.2	Basic Commands . . . . .	65
4.3.3	Data Types . . . . .	67
4.3.4	Reading Data . . . . .	73
4.3.5	Writing Data . . . . .	83
4.3.6	Shortcuts . . . . .	84
<b>5</b>	<b>Exploratory Data Analysis</b>	<b>87</b>
5.1	Data: Collection, Cleaning and Transformation . . . . .	88
5.1.1	Data Acquisition . . . . .	89
5.1.2	How Much Data Is Enough? . . . . .	89
5.1.3	Data Cleaning: The Different Dimensions of Data Quality . . . . .	90
5.1.4	Data Transformation: The Underestimated Effort . . . . .	91
5.2	Notebooks . . . . .	92
5.2.1	EDAs with Notebooks and Markdown . . . . .	92
5.2.2	Knitting . . . . .	97
5.3	The Tidyverse . . . . .	97
5.3.1	Why Use the Tidyverse? . . . . .	98
5.3.2	The Basic Tidyverse Verbs . . . . .	100
5.3.3	From Data Frames to Tibbles . . . . .	103
5.3.4	Data Transformation . . . . .	103
5.3.5	Regular Expressions and Mutate() . . . . .	109
5.4	Data Visualization . . . . .	110
5.4.1	Data Visualization as Part of the Analysis Process . . . . .	110
5.4.2	Data Visualization as Part of the Reporting . . . . .	111
5.4.3	Plots in Base R . . . . .	113
5.4.4	ggplot2: A Grammar of Graphics . . . . .	118
5.5	Data Analysis . . . . .	120
<b>6</b>	<b>Forecasting</b>	<b>129</b>
6.1	Linear Regression . . . . .	129
6.1.1	How the Algorithm Works . . . . .	130
6.1.2	Linear Regression in R . . . . .	133
6.1.3	Interpretation of the Results . . . . .	135
6.1.4	Advantages and Disadvantages . . . . .	137
6.1.5	Non-Linear Regression . . . . .	138
6.1.6	Small Hack: Linear Regression with Non-Linear Data . . . . .	141
6.1.7	Logistic Regression . . . . .	144
6.2	Anomaly Detection . . . . .	145
6.2.1	Time Series Analyses . . . . .	145

6.2.2	Fitting with the Forecast Package . . . . .	147
<b>7</b>	<b>Clustering</b>	<b>153</b>
7.1	Hierarchical Clustering . . . . .	153
7.1.1	Introduction to the Algorithm . . . . .	153
7.1.2	The Euclidean Distance and its Competitors . . . . .	157
7.1.3	The Distance Matrix, but Scaled . . . . .	159
7.1.4	The Dendrogram . . . . .	160
7.1.5	Dummy Variables: What If We Have No Numerical Data? . . . . .	162
7.1.6	What Do You Do with the Results? . . . . .	163
7.2	k-Means . . . . .	163
7.2.1	How the Algorithm Works . . . . .	164
7.2.2	How Do We Know k? . . . . .	166
7.2.3	Interpretation of the Results . . . . .	169
7.2.4	Is k-Means Always the Answer? . . . . .	171
<b>8</b>	<b>Classification</b>	<b>173</b>
8.1	Use cases for classification . . . . .	173
8.2	Create Training and Test Data . . . . .	175
8.2.1	The Titanic Data Set: A Brief EDA . . . . .	175
8.2.2	The Caret Package: Dummy Variables and Splitting the Data . . . . .	179
8.2.3	The pROC Package . . . . .	181
8.3	Decision Trees . . . . .	181
8.3.1	How the Algorithm Works . . . . .	182
8.3.2	Training and Test . . . . .	182
8.3.3	Interpretation of the Results . . . . .	184
8.4	Support Vector Machines . . . . .	185
8.4.1	How the Algorithm Works . . . . .	185
8.4.2	Data Preparation . . . . .	188
8.4.3	Training and Test . . . . .	188
8.4.4	Interpretations of the Results . . . . .	189
8.5	Naive Bayes . . . . .	190
8.5.1	How the Algorithm Works . . . . .	191
8.5.2	Data Preparation . . . . .	193
8.5.3	Training and Test . . . . .	193
8.5.4	Interpretation of the Results . . . . .	194
8.6	XG Boost: The Newcomer . . . . .	194
8.6.1	How the algorithm works . . . . .	195
8.6.2	Data Preparation . . . . .	196
8.6.3	Training and Test . . . . .	196
8.6.4	Interpretation of the Results . . . . .	199
8.7	Text Classification . . . . .	201
8.7.1	Prepare the Data . . . . .	201

8.7.2	Training and Test	204
8.7.3	Interpretation of the results	205
<b>9</b>	<b>Other Use Cases</b>	<b>207</b>
9.1	Shopping Cart Analysis – Association Rules	207
9.1.1	How the Algorithm Works	207
9.1.2	Data Preparation	208
9.1.3	Application of the Algorithm	210
9.1.4	Interpretations of the Results	211
9.1.5	Visualization of Association Rules	212
9.1.6	Association Rules with the Groceries Data Set	214
9.2	k-nearest Neighbors	215
9.2.1	How the Algorithm Identifies Outliers	215
9.2.2	Who Is the Furthest out of Everyone Now?	219
9.2.3	kNN as Classifier	220
9.2.4	LOF for Misclassification Analysis	223
<b>10</b>	<b>Workflows and Tools</b>	<b>225</b>
10.1	Versioning with Git	225
10.1.1	Why Versioning?	225
10.1.2	Git, GitHub, and GitLab	226
10.1.3	Basic commands	226
10.1.4	Integration with RStudio	228
10.1.5	Commit and Push Code	229
10.2	Dealing with Large Amounts of Data	234
10.2.1	Need a Bigger Computer? Cloud Computing with R	234
10.2.2	Working with Clusters: Spark and Sparklyr	235
10.2.3	data.table	243
10.3	Deploy Applications via API	243
10.3.1	What Is a REST API?	244
10.3.2	Provide an API with the “plumber” Package	244
10.3.3	The Next Step: Docker	247
10.4	Create Applications with Shiny	248
10.4.1	What Is Shiny?	248
10.4.2	UI and Server	251
10.4.3	Publish a Shiny App from RStudio	253
10.4.4	Example Applications	253
10.4.5	shinyapps.io	254
<b>11</b>	<b>Ethical Handling of Data and Algorithms</b>	<b>261</b>
11.1	Privacy	261
11.1.1	Legislations around the World	261
11.1.2	Do Users Really Care?	264
11.2	Ethics: Against Profiling and Discrimination	265
11.2.1	What Is Discrimination?	265

11.2.2	How to Prevent Discrimination . . . . .	266
11.2.3	What Is Profiling? . . . . .	268
11.2.4	How Can Profiling Be Prevented? . . . . .	271
<b>12</b>	<b>Next Steps after This Book</b>	<b>273</b>
12.1	Projects, Projects, Projects . . . . .	273
12.1.1	Putting Together a Project Portfolio . . . . .	273
12.1.2	Kaggle . . . . .	275
12.2	Where to Find Help . . . . .	277
12.2.1	RTFM . . . . .	277
12.2.2	Stack Overflow . . . . .	278
12.2.3	The R-Help Mailing List . . . . .	280
12.3	RSeek . . . . .	282
<b>13</b>	<b>Appendix: Troubleshooting</b>	<b>283</b>
13.1	Typical Error Messages and Solutions . . . . .	283
13.2	Typical Mistakes and How to Avoid Them . . . . .	283
13.3	R or RStudio Does Not Respond . . . . .	284
13.4	Typical Error Messages . . . . .	284
<b>14</b>	<b>Glossary</b>	<b>287</b>
	<b>Bibliography</b>	<b>293</b>
	<b>Index</b>	<b>297</b>