

Auf einen Blick

| | | |
|----|---|-----|
| 1 | Einleitung | 15 |
| 2 | Machine Learning, Data Science und künstliche Intelligenz | 25 |
| 3 | Ablauf eines Data-Science-Projekts | 39 |
| 4 | Einführung in R | 67 |
| 5 | Explorative Datenanalyse | 111 |
| 6 | Anwendungsfall Prognosen | 159 |
| 7 | Clustering | 185 |
| 8 | Klassifikation | 207 |
| 9 | Weitere Anwendungsfälle | 245 |
| 10 | Workflows und Werkzeuge | 267 |
| 11 | Ethischer Umgang mit Daten und Algorithmen | 307 |
| 12 | Was kommt nach diesem Buch? | 325 |

Inhalt

| | |
|---|-----------|
| Materialien zum Buch | 13 |
| 1 Einleitung | 15 |
| 1.1 Warum dieses Buch? | 15 |
| 1.2 Das Zeitalter der Daten – alles nur ein Hype? | 16 |
| 1.3 Warum nun Data Science? | 17 |
| 1.4 Warum Data Science mit R? | 19 |
| 1.5 Für wen ist dieses Buch? | 20 |
| 1.6 Kann man Data Science ohne Mathe lernen? | 20 |
| 1.7 Wie Sie dieses Buch verwenden können | 22 |
| 1.8 Materialien und Kontakt | 22 |
| 1.9 Danksagungen | 22 |
| 2 Machine Learning, Data Science und künstliche Intelligenz | 25 |
| 2.1 Aus der Geschichte lernen – alles nur ein Hype? | 25 |
| 2.1.1 Daten und Maschinen vor den Anfängen der KI | 25 |
| 2.1.2 Der erste Frühling der künstlichen Intelligenz | 28 |
| 2.1.3 Der erste KI-Winter | 29 |
| 2.1.4 Der zweite KI-Frühling: Expertensysteme | 29 |
| 2.1.5 Der zweite KI-Winter | 30 |
| 2.1.6 Kommt nun der dritte KI-Frühling? | 31 |
| 2.1.7 Rückschläge | 31 |
| 2.1.8 Technologische Singularität: Haben Maschinen ein Bewusstsein? | 32 |
| 2.1.9 Alan Turing und der Turing-Test | 33 |
| 2.2 Begriffsdefinitionen | 34 |
| 2.2.1 Machine Learning | 34 |

| | | |
|-------|----------------------------------|----|
| 2.2.2 | Künstliche Intelligenz | 35 |
| 2.2.3 | Data Science | 35 |
| 2.2.4 | Datenanalyse und Statistik | 37 |
| 2.2.5 | Big Data | 37 |

3 Ablauf eines Data-Science-Projekts 39

| | | |
|------------|--|----|
| 3.1 | Der allgemeine Ablauf eines Data-Science-Projekts | 39 |
| 3.1.1 | Die CRISP-DM Stages | 39 |
| 3.1.2 | ASUM-DM | 41 |
| 3.1.3 | Der Ablauf nach Hadley Wickham | 42 |
| 3.1.4 | Welcher Ansatz ist für mich der richtige? | 43 |
| 3.2 | Business Understanding: Welches Problem soll gelöst werden? | 43 |
| 3.2.1 | Senior-Management-Unterstützung und Einbeziehung der Fachabteilung | 43 |
| 3.2.2 | Anforderungen verstehen | 44 |
| 3.2.3 | Widerstände überwinden: Wer hat Angst vor der bösen KI? | 45 |
| 3.3 | Grundsätzliche Ansätze im Machine Learning | 47 |
| 3.3.1 | Supervised versus Unsupervised versus Reinforcement Learning | 47 |
| 3.3.2 | Feature Engineering | 48 |
| 3.4 | Performancemessung | 49 |
| 3.4.1 | Test- und Trainingsdaten | 49 |
| 3.4.2 | Fehler ist nicht gleich Fehler: False Positives und False Negatives | 50 |
| 3.4.3 | Confusion Matrix | 52 |
| 3.4.4 | ROC AUC | 53 |
| 3.4.5 | Precision Recall Curve | 54 |
| 3.4.6 | Wirkung des Modells | 55 |
| 3.4.7 | Data Science ROI | 56 |
| 3.5 | Kommunikation mit Stakeholdern | 57 |
| 3.5.1 | Reporting | 57 |
| 3.5.2 | Storytelling | 57 |
| 3.6 | Aus dem Labor in die Welt: Data-Science-Applikationen in Produktion | 58 |
| 3.6.1 | Von Data Pipelines und Data Lakes | 59 |
| 3.6.2 | Integration in andere Systeme | 59 |

| | | |
|------------|---|----|
| 3.7 | Die verschiedenen Rollen in einem Data-Science-Projekt | 59 |
| 3.7.1 | Data Scientist | 60 |
| 3.7.2 | Data Engineer | 61 |
| 3.7.3 | Data Science Architect | 61 |
| 3.7.4 | Business Intelligence Analyst | 61 |
| 3.7.5 | Der Subject Matter Expert | 62 |
| 3.7.6 | Projektmanagement | 62 |
| 3.7.7 | Citizen Data Scientist | 64 |
| 3.7.8 | Weitere Rollen | 65 |

4 Einführung in R 67

| | | |
|------------|---|-----|
| 4.1 | R: kostenlos, portierbar und interaktiv | 67 |
| 4.1.1 | Geschichte | 69 |
| 4.1.2 | Erweiterung mit Paketen | 70 |
| 4.1.3 | Die IDE RStudio | 71 |
| 4.1.4 | R versus Python | 71 |
| 4.1.5 | Andere Sprachen | 73 |
| 4.2 | Installation und Konfiguration von R und RStudio | 74 |
| 4.2.1 | Installation von R und kurzer Funktionstest | 74 |
| 4.2.2 | Installation von RStudio | 77 |
| 4.2.3 | Konfiguration von R und RStudio | 78 |
| 4.2.4 | Ein Rundgang durch RStudio | 82 |
| 4.2.5 | Projekte in RStudio | 86 |
| 4.2.6 | Die Cloud-Alternative: RStudio Cloud | 88 |
| 4.3 | Erste Schritte mit R | 89 |
| 4.3.1 | Alles in R ist ein Objekt | 89 |
| 4.3.2 | Grundlegende Befehle | 89 |
| 4.3.3 | Datentypen | 91 |
| 4.3.4 | Daten einlesen | 97 |
| 4.3.5 | Daten schreiben | 108 |
| 4.3.6 | Tipps zum professionellen und schnellen Arbeiten | 108 |

| | | |
|------------|---|-----|
| 5 | Explorative Datenanalyse | 111 |
| 5.1 | Daten: Sammlung, Reinigung und Transformation | 112 |
| 5.1.1 | Datenakquise | 113 |
| 5.1.2 | Wie viel Daten sind genug? | 114 |
| 5.1.3 | Datenreinigung: Die verschiedenen Dimensionen der Datenqualität | 115 |
| 5.1.4 | Datentransformation: Der unterschätzte Aufwand | 116 |
| 5.2 | Notebooks | 117 |
| 5.2.1 | EDAs mit Notebooks und Markdown | 117 |
| 5.2.2 | Knitting | 122 |
| 5.3 | Das Tidyverse | 123 |
| 5.3.1 | Warum das Tidyverse nutzen? | 123 |
| 5.3.2 | Die Grundverben | 126 |
| 5.3.3 | Von Data Frames zu Tibbles | 129 |
| 5.3.4 | Transformation von Daten | 129 |
| 5.3.5 | Reguläre Ausdrücke und mutate() | 136 |
| 5.4 | Datenvisualisierung | 137 |
| 5.4.1 | Datenvisualisierung als Teil der Analyse | 137 |
| 5.4.2 | Datenvisualisierung als Teil des Reportings | 138 |
| 5.4.3 | Plots in Base R | 140 |
| 5.4.4 | ggplot2: A Grammar of Graphics | 146 |
| 5.5 | Datenanalyse | 148 |
| 6 | Anwendungsfall Prognosen | 159 |
| 6.1 | Lineare Regression | 159 |
| 6.1.1 | Wie der Algorithmus funktioniert | 160 |
| 6.1.2 | Wie wird die lineare Regression in R durchgeführt? | 163 |
| 6.1.3 | Interpretation der Ergebnisse | 167 |
| 6.1.4 | Vor- und Nachteile | 168 |
| 6.1.5 | Nicht lineare Regression | 169 |
| 6.1.6 | Kleiner Hack: Lineare Regression bei nicht linearen Daten | 172 |
| 6.1.7 | Logistische Regression | 175 |

| | | |
|------------|---|-----|
| 6.2 | Anomalie-Erkennung | 176 |
| 6.2.1 | Ein kleiner Exkurs: Zeitreihenanalysen | 176 |
| 6.2.2 | Fitting mit dem Forecast-Package | 179 |
| 7 | Clustering | 185 |
| 7.1 | Hierarchisches Clustering | 185 |
| 7.1.1 | Einführung in die Vorgehensweise | 185 |
| 7.1.2 | Die euklidische Distanz und ihre Konkurrenten | 189 |
| 7.1.3 | Die Distanzmatrix, aber skaliert | 193 |
| 7.1.4 | Das Dendrogramm | 194 |
| 7.1.5 | Dummy-Variablen: Was, wenn wir keine numerischen Daten haben? | 196 |
| 7.1.6 | Was macht man nun mit den Ergebnissen? | 197 |
| 7.2 | k-Means | 197 |
| 7.2.1 | Wie der Algorithmus funktioniert | 198 |
| 7.2.2 | Woher kennen wir eigentlich k? | 201 |
| 7.2.3 | Interpretation der Ergebnisse | 204 |
| 7.2.4 | Ist k-Means immer die Antwort? | 206 |
| 8 | Klassifikation | 207 |
| 8.1 | Anwendungsfälle für eine Klassifikation | 207 |
| 8.2 | Trainings- und Testdaten erstellen | 209 |
| 8.2.1 | Der Titanic-Datensatz: Eine kurze EDA | 210 |
| 8.2.2 | Das Caret-Package: Dummy-Variablen und Aufteilen der Daten | 214 |
| 8.2.3 | Das pROC-Package | 216 |
| 8.3 | Decision Trees | 217 |
| 8.3.1 | Wie der Algorithmus funktioniert | 217 |
| 8.3.2 | Training und Test | 217 |
| 8.4 | Support Vector Machines | 221 |
| 8.4.1 | Wie der Algorithmus funktioniert | 221 |
| 8.4.2 | Vorbereitung der Daten | 224 |

| | | |
|------------|--|-----|
| 8.4.3 | Training und Test | 224 |
| 8.4.4 | Interpretationen der Ergebnisse | 225 |
| 8.5 | Naive Bayes | 226 |
| 8.5.1 | Wie der Algorithmus funktioniert | 228 |
| 8.5.2 | Vorbereitung der Daten | 229 |
| 8.5.3 | Training und Test | 229 |
| 8.5.4 | Interpretation der Ergebnisse | 231 |
| 8.6 | XG Boost: Der Newcomer | 232 |
| 8.6.1 | Wie der Algorithmus funktioniert | 232 |
| 8.6.2 | Vorbereitung der Daten | 233 |
| 8.6.3 | Training und Test | 233 |
| 8.6.4 | Interpretation der Ergebnisse | 236 |
| 8.7 | Klassifikation von Text | 238 |
| 8.7.1 | Vorbereiten der Daten | 239 |
| 8.7.2 | Training und Test | 242 |
| 8.7.3 | Interpretation der Ergebnisse | 242 |
| 9 | Weitere Anwendungsfälle | 245 |
| 9.1 | Warenkorbanalyse – Association Rules | 245 |
| 9.1.1 | Wie der Algorithmus funktioniert | 245 |
| 9.1.2 | Vorbereitung der Daten | 246 |
| 9.1.3 | Anwendung des Algorithmus | 248 |
| 9.1.4 | Interpretationen der Ergebnisse | 249 |
| 9.1.5 | Visualisierung von Assoziationsalgorithmen | 251 |
| 9.1.6 | Association Rules mit dem Datensatz »Groceries« | 252 |
| 9.2 | k-nearest Neighbours | 254 |
| 9.2.1 | Wie der Algorithmus Ausreißer identifiziert | 254 |
| 9.2.2 | Wer ist denn jetzt am weitesten draußen von allen? | 258 |
| 9.2.3 | kNN als Klassifikator | 260 |
| 9.2.4 | LOF zur Analyse der Fehlklassifikationen | 263 |

| | | |
|-------------|---|-----|
| 10 | Workflows und Werkzeuge | 267 |
| 10.1 | Versionierung mit Git | 267 |
| 10.1.1 | Warum Versionierung? | 267 |
| 10.1.2 | Git, GitHub und GitLab | 268 |
| 10.1.3 | Basisbefehle | 269 |
| 10.1.4 | Integration in RStudio | 270 |
| 10.1.5 | Code committen und pushen | 274 |
| 10.2 | Mit großen Datenmengen umgehen | 277 |
| 10.2.1 | Größerer Computer gefällig? Cloud-Computing mit R | 278 |
| 10.2.2 | Mit Clustern arbeiten: Spark und Sparklyr | 278 |
| 10.2.3 | data.table | 287 |
| 10.3 | Applikationen via API bereitstellen | 287 |
| 10.3.1 | Was ist eine REST API? | 288 |
| 10.3.2 | Eine API mit dem Package »plumber« bereitstellen | 288 |
| 10.3.3 | Der nächste Schritt: Docker | 291 |
| 10.4 | Applikationen erstellen mit Shiny | 292 |
| 10.4.1 | Was ist Shiny? | 292 |
| 10.4.2 | UI and Server | 295 |
| 10.4.3 | Veröffentlichen einer Shiny-App aus RStudio | 298 |
| 10.4.4 | Beispielapplikationen | 298 |
| 10.4.5 | ShinyApps.io | 300 |
| 11 | Ethischer Umgang mit Daten und Algorithmen | 307 |
| 11.1 | Datenschutz | 307 |
| 11.1.1 | Die Entwicklung des Datenschutzes am Beispiel von Deutschland | 307 |
| 11.1.2 | Gesetze sind das eine, das eigene Verhalten das andere | 309 |
| 11.1.3 | Was bedeutet Datenschutz für Data-Science-Projekte? | 313 |
| 11.1.4 | Die Datenschutz-Folgeabschätzung | 316 |
| 11.2 | Ethik: Gegen Profiling und Diskriminierung | 317 |
| 11.2.1 | Was ist Diskriminierung? | 318 |
| 11.2.2 | Wie kann Diskriminierung verhindert werden? | 319 |
| 11.2.3 | Was ist Profiling? | 320 |
| 11.2.4 | Wie kann Profiling verhindert werden? | 322 |


| | | |
|-------------------|--|-----|
| 12 | Was kommt nach diesem Buch? | 325 |
| 12.1 | Projekte, Projekte, Projekte | 325 |
| 12.1.1 | Ein Projektportfolio zusammenstellen | 325 |
| 12.1.2 | Kaggle | 328 |
| 12.2 | Wer hilft Ihnen jetzt weiter? | 329 |
| 12.2.1 | RTFM | 329 |
| 12.2.2 | Stack Overflow | 331 |
| 12.2.3 | Die R-Help-Mailingliste | 334 |
| 12.3 | RSeek | 335 |
| Anhang | | 337 |
| A | Typische Fehlermeldungen und geeignete Lösungen | 339 |
| B | Glossar | 343 |
| C | Literatur | 347 |
| Index | | 353 |

Materialien zum Buch

Auf der Webseite zu diesem Buch stehen folgende Materialien für Sie zum Download bereit:

► **alle Beispielprogramme aus dem Buch**

Gehen Sie auf www.rheinwerk-verlag.de/5341. Klicken Sie auf den Reiter MATERIALIEN. Sie sehen die herunterladbaren Dateien samt einer Kurzbeschreibung des Dateiinhalts. Klicken Sie auf den Button HERUNTERLADEN, um den Download zu starten. Je nach Größe der Datei (und Ihrer Internetverbindung) kann es einige Zeit dauern, bis der Download abgeschlossen ist.

Diese Leseprobe haben Sie beim
 edv-buchversand.de heruntergeladen.
 Das Buch können Sie online in unserem
 Shop bestellen.
[Hier zum Shop](#)