

JOHN W. FOREMAN

BIG DATA smart mit **EXCEL** analysieren



So holen Sie das Beste aus Ihren Kundendaten heraus

 **SYBEX**
A Wiley Brand

John W. Foreman

Big Data smart mit Excel analysieren

**So holen Sie das Beste aus
Ihren Kundendaten heraus**

Deutsche Ausgabe von »Data Smart«

Übersetzung aus dem Amerikanischen von
Meinhard Schmidt



Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

1. Auflage 2015

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved including the right of reproduction in whole or in part in any form. This translation published by arrangement with John Wiley and Sons, Inc.

Copyright der englischsprachigen Originalausgabe »Data Smart« © 2014 by John Wiley and Sons, Inc.

Alle Rechte vorbehalten inklusive des Rechtes auf Reproduktion im Ganzen oder in Teilen und in jeglicher Form. Diese Übersetzung wird mit Genehmigung von John Wiley and Sons, Inc. publiziert.

Wiley, das Wiley-Logo und das Sybex-Logo sind Marken oder eingetragene Marken von John Wiley & Sons, Inc., USA, Deutschland und in anderen Ländern und dürfen nicht ohne schriftliche Genehmigung genutzt werden. Alle anderen Marken sind Eigentum ihrer jeweiligen Inhaber. John Wiley & Sons, Inc. und WILEY-VCH Verlag GmbH & Co. KGaA stehen in keiner Verbindung zu den in diesem Buch erwähnten Produkten und Herstellern.

Das vorliegende Werk wurde sorgfältig erarbeitet. Dennoch übernehmen Autoren und Verlag für die Richtigkeit von Angaben, Hinweisen und Ratschlägen sowie eventuelle Druckfehler keine Haftung.

Wir möchten Sie mit diesem Buch optimal unterstützen und freuen uns daher über Ihre Anregungen und Verbesserungsvorschläge. Notwendige Korrekturen

veröffentlichen wir im Interesse aller Leser umgehend unter www.sybex.de und berücksichtigen sie bei der nächsten Auflage. Herzlichen Dank für Ihre Unterstützung!

Ihr Sybex-Lektoratsteam

lektorat@sybex.de

Print ISBN 978-3-527-76060-2

ePub ISBN: 978-3-527-69224-8

mobi ISBN: 978-3-527-69223-1

Coverfoto © Kumbabali – Fotolia.com

Umschlaggestaltung Torge Stoffers Grafik-Design, Leipzig

Korrektur Petra Heubach-Erdmann und Jürgen Erdmann,
Düsseldorf

Satz inmedialo Digital- und Printmedien UG, Plankstadt

Über den Autor

John W. Foreman ist der Chief Data Scientist von MailChimp.com. Davor hat er als Managementberater nicht nur in großen Unternehmen (wie Coca-Cola, Royal Caribbean, Intercontinental Hotels), sondern auch für die US-Regierung (wie das Verteidigungsministerium, die Bundessteuerbehörde, das Ministerium für innere Sicherheit DHS und das FBI) als Analytiker gearbeitet. John W. Foreman ist ein gern gehörter Redner, wenn es darum geht, über die Möglichkeiten und Probleme zu informieren, die die Einrichtung von Analysesoftware in Unternehmen mit sich bringen kann.

Wenn er nicht gerade mit Daten herumspielt, verbringt er seine Zeit mit Wandern, dem Abhängen vor dem Fernseher, dem Verputzen richtig ungesunder Nahrungsmittel und dem Aufziehen von drei prima Jungen.

Inhaltsverzeichnis

Über den Autor

Einführung

Was mache ich hier?

Eine brauchbare Definition von Data Science

Was hat es denn mit Big Data auf sich?

Wer bin ich?

Wer sind Sie?

Nichts geht über eine Tabellenkalkulation

Aber Tabellenkalkulationen sind doch aus der Mode!

Verwenden Sie Excel oder LibreOffice Konventionen

Los geht's

Kapitel 1 Alles, was Sie jemals über Tabellenkalkulationen wissen wollen, sich aber nicht zu fragen getraut haben

1.1 Beispieldaten

1.2 Sich schnell mit der Steuerungstaste bewegen

1.3 Formeln und Daten schnell kopieren

1.4 Zellen formatieren

- [1.5 Inhalte einfügen](#)
- [1.6 Diagramme hinzufügen](#)
- [1.7 Die Menüs »Suchen« und »Ersetzen«](#)
- [1.8 Formeln für das Auffinden und Entnehmen von Werten](#)
- [1.9 SVERWEIS verwenden, um Daten zusammenzuführen](#)
- [1.10 Filtern und sortieren](#)
- [1.11 Pivot-Tabellen verwenden](#)
- [1.12 Array-Formeln verwenden](#)
- [1.13 Probleme mit dem Solver lösen](#)
- [1.14 OpenSolver: Ich wünschte, wir würden ihn nicht benötigen. Dem ist aber nicht so](#)
- [1.15 Zusammenfassung](#)

[Kapitel 2 Clusteranalyse Teil I: Die Kundenbasis mit k-Means aufteilen](#)

- [2.1 Mädchen tanzen mit Mädchen, und Jungens kratzen sich am Kopf](#)
- [2.2 Es wird ernst: k-Means-Clusterbildung bei Abonnenten eines E-Mail-Marketings](#)
 - [2.2.1 Joey Bag O' Donuts Weinhandel](#)
 - [2.2.2 Die Ausgangsdaten](#)
 - [2.2.3 Festlegen, was zu bewerten ist](#)
 - [2.2.4 Mit vier Clustern beginnen](#)
 - [2.2.5 Euklidischer Abstand: Abstandsmessung auf kürzestem Weg](#)
 - [2.2.6 Abstände und Clusterzuweisungen für jedermann](#)
 - [2.2.7 Clusterzentren bestimmen](#)
 - [2.2.8 Aus den Ergebnissen schlau werden](#)

2.2.9 Die Top-Verkäufe je Cluster erhalten

2.2.10 Die Silhouette: Ein guter Weg, um es unterschiedliche k-Werte unter sich ausfechten zu lassen

2.2.11 Was halten Sie von fünf Clustern?

2.2.12 Eine Lösung für fünf Cluster

2.2.13 Die Top-Verkäufe der fünf Cluster erhalten

2.2.14 Die Silhouette für die 5-Means-Clusterbildung berechnen

2.3 K-Medians-Clusterbildung und asymmetrische Abstandsmessungen

2.3.1 Die k-Medians-Clusterbildung

2.3.2 Eine geeignetere Abstandsmetrik erhalten

2.3.3 Bringen Sie das alles in Excel unter

2.3.4 Die Top-Verkäufe der 5-Medians-Cluster

2.4 Zusammenfassung

Kapitel 3 Naives Bayes und wie unglaublich leicht es ist, ein Idiot zu sein

3.1 Wenn Sie ein Produkt »Mandrill« nennen, erhalten Sie Signale und Nebengeräusche

3.2 Die kürzeste Einführung in die Wahrscheinlichkeitsrechnung der Welt

3.2.1 Bedingte Wahrscheinlichkeiten summieren

3.2.2 Die Verbundwahrscheinlichkeit, die Kettenregel und die Unabhängigkeit

3.2.3 Was geschieht in einer abhängigen Situation?

3.2.4 Die Bayes-Regel

3.3 Die Bayes-Regel verwenden, um ein KI-Modell zu erstellen

3.3.1 Klassenwahrscheinlichkeiten auf hohem Niveau werden oft miteinander gleichgesetzt

3.3.2 Und noch ein paar Kleinigkeiten

3.4 Auf geht's mit Excel

3.4.1 Für die Sache irrelevante Interpunktion entfernen

3.4.2 An Leerzeichen auftrennen

3.4.3 Token zählen und Wahrscheinlichkeiten berechnen

3.4.4 Wir haben ein Modell! Nutzen wir es!

3.5 Zusammenfassung

Kapitel 4 Optimierungsmodellierung: Weil der »frisch gepresste« Orangensaft sich nicht selbst herstellt

4.1 Warum sollten Data Scientists wissen, was Optimierung bedeutet?

4.2 Mit einem einfachen Zielkonflikt geht es los

4.2.1 Das Problem als Polytop darstellen

4.2.2 Lösen durch Verschieben der Niveaumenge

4.2.3 Das Simplex-Verfahren: in den Ecken herumstöbern

4.2.4 Mit Excel arbeiten

4.2.5 Am Ende dieses Kapitels wartet ein Monster

4.3 Frisch vom Baum in Ihr Glas ... mit einem kurzen Boxenstopp fürs Mischen

4.3.1 Sie verwenden für das Mischen ein Modell

4.3.2 Beginnen wir mit ein paar Spezifikationen

4.3.3 Zurück zum gleichbleibenden Geschmack

4.3.4 Die Daten in Excel eintragen

4.3.5 Das Problem in Solver eingeben

4.3.6 Die Standards herabsetzen

4.3.7 Ein totes Eichhörnchen loswerden: der Minimax-Ansatz

4.3.8 Wenn-Dann- und die Big-M-Bedingung

4.3.9 Variablen vervielfachen: das Volumen bis auf 11 hochtreiben

4.4 Modellierungsrisiko

4.4.1 Normal verteilte Daten

4.5 Zusammenfassung

Kapitel 5 Clusteranalyse Teil II: Netzwerkdiagramme und die Entdeckung der Community

5.1 Was ist ein Netzwerkdiagramm?

5.2 Einen einfachen Graphen darstellen

5.3 Eine kurze Einführung in Gephi

5.3.1 Die Installation von Gephi und die Vorbereitung der Dateien

5.3.2 Den Graphen gestalten

5.3.3 Rangfolge von Knoten

5.3.4 Drucken

5.3.5 Dem Graphen an die Daten gehen

5.4 Aus den Daten des Weinhandels einen Graphen bilden

- [5.4.1 Eine Kosinus-Ähnlichkeitsmatrix erstellen](#)
- [5.4.2 Einen r-Nachbarschaftsgraphen entwickeln](#)
- [5.5 Wie viel ist eine Kante wert? Normale Punkte und Penaltys bei der Modularität von Graphen](#)
- [5.5.1 Was ist ein Punkt und woraus besteht ein Penalty?](#)
- [5.5.2 Das Arbeitsblatt für die Bewertungen einrichten](#)
- [5.6 Lassen Sie uns Cluster bilden!](#)
- [5.6.1 Aufteilung Nummer 1](#)
- [5.6.2 Aufteilung 2: Electric Boogaloo](#)
- [5.6.3 Und ... Aufteilung 3: Aufteilung mit Vergeltung](#)
- [5.6.4 Die Communitys decodieren und analysieren](#)
- [5.7 Einmal hin und wieder zurück: eine Gephi-Tabelle](#)
- [5.8 Zusammenfassung](#)

[Kapitel 6 Der Großvater der betreuten künstlichen Intelligenz - die Regression](#)

- [6.1 He, was bist du? Schwanger?](#)
- [6.2 Machen Sie sich nicht selbst verrückt](#)
- [6.3 Die Schwangerschaft von Kundinnen bei RetailMart mithilfe der linearen Regression vorhersagen](#)
- [6.3.1 Welche Funktionen benötigt werden](#)
- [6.3.2 Die Trainingsdaten zusammenstellen](#)
- [6.3.3 Dummy-Variablen erzeugen](#)

6.3.4 Backen wir uns unsere eigene lineare Regression

6.3.5 Statistiken und lineare Regression: R-Quadrat, F-Test und t-Tests

6.3.6 Vorhersagen anhand neuer Daten tätigen und die Leistungsfähigkeit messen

6.4 Mit einer logistischen Regression Schwangerschaften in Kundenhaushalten vorhersagen

6.4.1 Als Erstes benötigen Sie eine Verknüpfungsfunktion

6.4.2 Die logistische Funktion einbinden und alles neu optimieren

6.4.3 Eine echte logistische Regression zusammenbauen

6.4.4 Modellauswahl - die Leistungsfähigkeit des linearen mit der des logistischen Modells vergleichen

6.5 Wenn Sie mehr wissen wollen

6.6 Zusammenfassung

Kapitel 7 Ensemble-Modelle: eine Menge mieser Pizza

7.1 Die Daten aus Kapitel 6 verwenden

7.2 Bagging: zufällig anordnen, trainieren, wiederholen

7.2.1 Decision Stump ist keine sehr sexy Bezeichnung für eine blöde Vorhersage

7.2.2 Das sieht für mich gar nicht mal so dumm aus!

7.2.3 Das Modell untersuchen

7.3 Boosting: Wenn das Ergebnis falsch ist, verstärken Sie es und versuchen es auf ein Neues

7.3.1 Das Modell trainieren - jedes Merkmal wird angesprochen

7.3.2 Das verstärkte Modell auswerten

7.4 Zusammenfassung

Kapitel 8 Prognosen: Atmen Sie tief durch, Sie können nicht gewinnen

8.1 Der Handel mit Schwertern stottert

8.2 Mit Zeitreihen vertraut werden

8.3 Langsam Fahrt aufnehmen mit einer einfachen exponentiellen Glättung

8.3.1 Prognosen mit der einfachen exponentiellen Glättung einrichten

8.4 Es könnte ein Trend vorliegen

8.5 Die lineare exponentielle Glättung nach Holt

8.5.1 Die lineare exponentielle Glättung nach Holt in einem Arbeitsblatt einrichten

8.5.2 Sind Sie nun fertig? Einen Blick auf Autokorrelationen werfen

8.6 Die multiplikative Glättung nach Holt-Winters

8.6.1 Die Anfangswerte für Niveau, Trend und Saisonabhängigkeit festlegen

8.6.2 Die Prognose ins Rollen bringen

8.6.3 Optimieren!

8.6.4 Bestätigen Sie mir jetzt bitte, dass wir fertig sind

8.6.5 Um die Prognose einen Vorhersagebereich legen

8.6.6 Für die Galerie: Ein Fan-Chart anlegen

8.7 Zusammenfassung

Kapitel 9 Die Entdeckung von Ausreißern: Nur weil sie sonderbar sind, heißt das nicht, dass sie auch unwichtig sind

9.1 Auch Ausreißer sind nur (schlechte?) Menschen

9.2 Der faszinierende Fall von Hadlum gegen Hadlum

9.2.1 Tukey-Begrenzungen

9.2.2 Tukey-Begrenzungen in einem Arbeitsblatt anwenden

9.2.3 Die Grenzen dieser einfachen Vorgehensweise

9.3 In nichts wirklich schlecht, aber auch nirgends wirklich gut

9.3.1 Daten für einen Graphen vorbereiten

9.3.2 Einen Graphen erstellen

9.3.3 Die k nächsten Nachbarn erhalten

9.3.4 Methode 1 zum Entdecken von Ausreißern in einem Graphen: Verwenden Sie einfach den Indegree

9.3.5 Methode 2 zum Entdecken von Ausreißern in einem Graphen: Differenzierte Ergebnisse mit k-Abstand erhalten

9.3.6 Methode 3 zum Entdecken von Ausreißern in einem Graphen: Local Outlier Factors sind dort, wo die Musik spielt

9.4 Zusammenfassung

Kapitel 10 Von der Tabellenkalkulation zu R wechseln

10.1 Mit R loslegen

10.1.1 Ein paar einfache Fingerübungen

10.1.2 Daten in R einlesen

10.2 Sich aktiv mit Data Science beschäftigen

10.2.1 Ein paar Zeilen sphärisches k-Means für Wein-Daten

10.3 Mit den Schwangerschaftsdaten ein KI- Modell entwickeln

10.3.1 Prognosen in R tätigen

10.3.2 Sich um das Entdecken von Ausreißern kümmern

10.4 Zusammenfassung

Stichwortverzeichnis

Einführung

Was mache ich hier?

Möglicherweise sind Sie in den Medien, in Büchern, die sich mit unternehmensbezogenen Themen beschäftigen, in Zeitschriften oder auf Konferenzen schon einmal über den Begriff *Data Science* gestolpert. Data Science (oder – grob übersetzt – die Wissenschaft von den Daten) ist in der Lage, Präsidentschaftswahlkämpfe in Hektik zu versetzen, mehr über Ihre Kaufgewohnheiten aufzudecken, als Sie von sich selbst wissen, und präzise Auskunft darüber zu geben, seit wie vielen Jahren diese ausgesprochen leckeren Käse-Cracker für Ihren Cholesterinspiegel verantwortlich sind. *Data Scientists*, die »Datenwissenschaftler«, die gleichzeitig die Elite derer bilden, die die Kunst der Data Science praktizieren, sind in einem Artikel im Harvard Business Review sogar schon als »sexy« bezeichnet worden. Dies sollten Sie nicht zu ernst nehmen, denn der Stellenwert dieser Behauptung lässt sich mit dem Stellenwert von Aussagen wie der vergleichen, dass ein Einhorn sexy sei. Dieser Teil des Artikels kann im Moment nicht bestätigt werden, aber wenn Sie mich dabei beobachten könnten, wie ich dieses Buch schreibe, mit zerwühlten Haaren und den müden Augen eines Vaters von drei Jungen, können Sie sich sicherlich vorstellen, dass sexy ein wenig übertrieben ist.

Aber ich schweife ab. In Wirklichkeit geht es darum, dass heutzutage ziemlich viel Wirbel um Data Science gemacht wird, was wiederum ziemlich viel Druck auf bestimmte Geschäftszweige ausübt. Wenn Sie sich nicht um Data Science kümmern, hängt Sie der Wettbewerb ab. Irgendjemand bringt ein neues Produkt mit dem Namen

»BlahBlahBlahBigDataGraphDing« auf den Markt und macht damit Ihr Unternehmen kaputt.

Atmen Sie ganz tief durch.

Die Wahrheit sieht so aus, dass die meisten Menschen falsche Vorstellungen von Data Science haben. Das beginnt damit, dass sie sich die entsprechenden Werkzeuge kaufen und Berater anheuern. Sie geben ihr ganzes Geld aus, bevor sie überhaupt wissen, was sie wollen, weil heute in vielen Unternehmen schon ein Kaufauftrag mit Erfolg gleichgesetzt wird.

Wenn Sie dieses Buch lesen, bekommen Sie diesen Spaßvögeln gegenüber einen großen Vorteil, weil Sie hier genau erfahren, was es mit den Techniken der Data Science auf sich hat und wie Sie sie anwenden können. Wenn dann die Zeit der Planung, des Anheuerns von Beratern und des Einkaufens gekommen ist, wissen Sie bereits, wie Sie herausfinden können, was in Ihrer Organisation an Data Science möglich ist.

Dieses Buch hat den Sinn, Ihnen die Data-Science-Praxis auf angenehme Weise und unterhaltsam vorzustellen. Wenn Sie das Buch durchgelesen haben, hoffe ich, dass viele Ängste, die mit Data Science zu tun haben, durch Neugier und Ideen darüber ersetzt worden sind, was Sie mit Daten machen können, um Ihr Unternehmen weiter nach vorn zu bringen.

Eine brauchbare Definition von Data Science

Der Ausdruck *Data Science* dient in gewisser Weise auch als Synonym für Begriffe wie *Business Analytics* (betriebswirtschaftliche Auswertungen), *Operations Research*(Unternehmensforschung), *Business Intelligence* (mit diesem Begriff werden Verfahren und Prozesse zur systematischen Analyse von Daten bezeichnet; er wird auch als BI abgekürzt), *Competitive Intelligence* (was mit

Wettbewerbsforschung oder -analyse übersetzt werden könnte), *Data Analysis And Modeling* (Datenanalyse und Datenmodellierung) und *Knowledge Extraction* (das Extrahieren von Erkenntnissen, was auch *Knowledge Discovery In Databases* oder *KDD* genannt wird). Letztendlich handelt es sich bei *Data Science* nur um eine neue Bezeichnung für etwas, das in Unternehmen schon seit Langem getan wird - und das auch im Deutschen gerne mit englischen Ausdrücken belegt wird. Diese Ausdrücke haben sich inzwischen oft zu Fachbegriffen gemausert, die wir, wie hier, zumindest einmal mit einer deutschsprachigen Entsprechung versehen und in den Index aufgenommen haben, damit Sie eine bessere Vorstellung davon bekommen, worum es geht. Nun ist aber auch im Umfeld der Datenanalyse nicht alles englisch, was glänzt. Wenn es im fachspezifischen Umfeld (womit nicht populärwissenschaftliche Artikel in Computer- und Managementzeitschriften, sondern primär Wissenschaft und Unternehmen gemeint sind, die sich hauptberuflich mit unserer Thematik beschäftigen) »normal« ist, deutschsprachig zu agieren, wird in der Übersetzung auf Englisch insoweit verzichtet, als dass die deutschsprachigen Begriffe verwendet werden und ihre englische Entsprechung zumindest einmal als Information aufgeführt wird. Auch in diesem Fall hilft der Index dabei, sich zurechtzufinden.

Seit der Blütezeit dieser »synonymen« Begriffe hat es eine nicht unbeträchtliche technologische Weiterentwicklung gegeben. Diese Weiterentwicklungen bei der Hardware und der Software haben dafür gesorgt, dass das Sammeln, Speichern und Auswerten großer Datenmengen aus dem Vertrieb und dem Marketing, aus HTTP-Anfragen an Ihre Website, aus Daten des Kundendienstes und so weiter einfacher und kostengünstiger geworden ist. Endlich sind auch kleinere Unternehmen und nicht kommerzielle

Organisationen in der Lage, sich mit Analysen zu beschäftigen, die bis dahin ausschließlich großen Unternehmen vorbehalten waren. Da der Begriff *Data Science* heutzutage für so gut wie alles verwendet wird, was mit einer Analyse unternehmensbezogener Daten zu tun hat, wird er häufig mit den Techniken des Data-Minings gleichgesetzt, zu denen beispielsweise die künstliche Intelligenz (KI), die Clusterbildung und das Erkennen von Ausreißern gehören. Dank der fulminanten, auf Transaktionen beruhenden Vermehrung von Unternehmensdaten haben diese rechenintensiven Techniken in den letzten Jahren einen Fuß in die Tür von Unternehmen bekommen, für die es sich bis dahin nicht gelohnt hat, so etwas produktiv zu verwenden.

Ich vertrete in diesem Buch eine sehr weit gefasste Definition des Begriffs Data Science. Sie sieht so aus:

Data Science ist die Umwandlung von Daten mithilfe der Mathematik und statistischer Methoden in wertvolle Erkenntnisse, Entscheidungen und Produkte.

Dies ist eine *unternehmensbezogene* Definition. Dort geht es um ein nützliches und wertvolles Endergebnis, das aus Daten abgeleitet wird. Warum? Mir geht es hier weder um Marktforschung noch glaube ich, dass Daten ästhetische Werte aufweisen. Ich kümmere mich um Data Science, damit mein Unternehmen besser funktioniert und Werte hervorbringt. Und ich kann mir vorstellen, dass es Ihnen ähnlich ergeht.

Dieses Buch nimmt obige Definition als Grundlage und behandelt zentrale Analysetechniken, zu denen nicht nur Optimierung, Prognosen und Simulationen, sondern auch »heißere« Themen wie künstliche Intelligenz, Netzwerkdiagramme, Clusterbildung und das Entdecken von Ausreißern gehören.

Einige dieser Techniken sind Jahrzehnte alt. Andere wurden erst in den letzten fünf Jahren entwickelt. Und Sie werden

sehen, dass Alter nichts mit Problemen oder Nutzen zu tun hat. Alle vorgestellten Techniken sind unabhängig davon, wie aktuell sie gerade sind, im richtigen Unternehmensumfeld gleich nützlich.

Damit kennen Sie auch schon den Grund dafür, warum Sie verstehen müssen, wie diese Techniken funktionieren, wie Sie die für ein Problem geeignete Technik auswählen und damit erste Schritte unternehmen können. Dort draußen gibt es viele Typen, die sich zwar mit einer oder zwei dieser Techniken auskennen, die aber den Rest nicht auf ihrem Radar haben. Wenn es in meiner Werkzeugkiste nur einen Hammer gibt, neige ich – wie mein zweijähriger Sohn – dazu, alle Probleme dadurch zu lösen, dass ich hart zuschlage.

Da ist es doch wohl besser, ein paar zusätzliche Werkzeuge zur Auswahl zu haben.

Was hat es denn mit Big Data auf sich?

Höchstwahrscheinlich sind Sie öfter über *Big Data* als über *Data Science* gestolpert. Handelt dieses Buch von Big Data? Das hängt davon ab, wie Sie Big Data definieren. Wenn Sie unter Big Data das Berechnen einfacher, zusammenfassender Statistiken anhand unstrukturierter Daten verstehen, die in riesigen, horizontal skalierbaren Datenbanken liegen, die nichts mit SQL zu tun haben, dann hat dieses Buch nichts mit Big Data zu tun.

Wenn Sie Big Data aber als Umwandlung geschäftlicher Daten in Entscheidungen und Erkenntnisse definieren, wobei für diese Umwandlung (ohne Rücksicht darauf, wo die Daten gespeichert sind) innovative Analysemethoden verwendet werden, dann handelt dieses Buch auch von Big Data.

Dieses Buch beschäftigt sich nicht mit Datenbanktechnologien wie MongoDB oder HBase. Dieses Buch behandelt auch keine Projekte zur Data-Science-Kodierung wie Mahout, NumPy, die verschiedenen R-Bibliotheken und so weiter. Um diese Themen kümmern sich andere Bücher.

Und das ist auch gut so. Dieses Buch ignoriert die Werkzeuge, die Speicherung und den Code. Stattdessen konzentriert es sich so weit wie möglich auf die Techniken. Dort draußen gibt es viele Menschen, die glauben, dass Big Data nichts als Datenspeicherung und Datenabfrage ist, wobei die Daten ein wenig bereinigt und zusammengefasst werden.

Sie irren. Dieses Buch bringt Sie auf eine Ebene, die über dem liegt, was Sie von den Verkäufern von Big-Data-Software und von Bloggern zu hören bekommen, und es zeigt Ihnen, was Sie wirklich aus Ihren Daten herausholen können. Und das Beste daran ist, dass der Umfang Ihrer Daten für die meisten dieser Techniken keine Rolle spielt. Sie müssen nicht erst über ein Petabyte an Daten verfügen und die entsprechenden Kosten bewältigen, bevor Sie sich mit den Interessen Ihrer Kunden auseinandersetzen dürfen. Wenn Sie einen großen Datenbestand haben, ist das prima, aber genauso gibt es Unternehmen, die so etwas nicht aufweisen, nicht benötigen und niemals haben werden. Wie das zum Beispiel bei meinem Metzger der Fall ist. Das bedeutet aber noch lange nicht, dass sein E-Mail-Marketing nicht von einem Würstchen-im-Vergleich-mit-Schinken-Cluster profitieren könnte.

Wenn Bücher über Data Science Trainingsunterlagen wären, hätten Sie es nur mit Lockerungsübungen zu tun – keine Gewichte, nichts Ergometrisches. Wenn ein Buch aber weiter geht und Sie verstanden haben, wie Sie die Techniken nur mit einem Grundstock an Werkzeugen implementieren können, sind Sie auch in der Lage, diese Implementierungen

in einer Vielzahl von Technologien vorzunehmen, auf ihnen problemlos etwas aufzubauen, bei Beratern die richtigen Data-Science-Produkte zu erwerben, Ihren Entwicklern die richtige Vorgehensweise an die Hand zu geben und so weiter.

Wer bin ich?

Gönnen Sie mir eine kurze Unterbrechung, um Ihnen etwas über mich zu erzählen. Es würde zu weit gehen, Ihnen zu erklären, warum ich Data Science so lehre, wie ich das tue. Bis vor einigen Jahren war ich Unternehmensberater. Ich beschäftigte mich bei Organisationen wie dem FBI, dem US-amerikanischen Verteidigungsministerium, der Coca-Cola Company, der Intercontinental Hotels Group und der Royal Caribbean International mit Analyseproblemen. Und jedes Mal, wenn ich irgendwo wegging, verstärkte sich das Gefühl, dass viel mehr Menschen als nur die, die hauptberuflich als Data Scientists arbeiten, Data Science verstehen sollten.

Ich habe mit Managern zusammengearbeitet, die Simulationen gekauft haben, obwohl sie Optimierungsmodelle benötigten. Ich habe mit Analysten zusammengearbeitet, die ausschließlich mit Gantt-Diagrammen umgehen konnten, weshalb alles über Gantt-Diagramme gelöst werden musste. Ein Berater konnte leicht einen Kunden mit einer alten Publikation und einer gekonnt gemachten PowerPoint-Präsentation beeindrucken, der KI nicht von BI unterscheiden kann.

In diesem Buch geht es darum, ein größeres Publikum in die Lage zu versetzen, Data-Science-Techniken zu verstehen und zu implementieren. Ich habe nicht die Absicht, aus Ihnen gegen Ihren Willen einen »Datenwissenschaftler« zu machen. Ich möchte nur die Rolle, die Sie bisher im Unternehmen spielen, um die Fähigkeiten erweitern, mit Data Science umzugehen.

Wer sind Sie?

Keine Angst, aber ich habe nicht vor, Sie mithilfe von Data Science auszuspionieren. Ich habe keine Ahnung, wer Sie sind, aber vielen Dank dafür, dass Sie für dieses Buch Geld ausgegeben haben. Vielleicht unterstützen Sie aber auch Ihre örtliche Bibliothek. Das wäre auch gut.

Hier ein paar Archetypen (oder *Personas*, wie sie in Marketingkreisen genannt werden), die in meinem Kopf herumspukten, als ich dieses Buch schrieb. Vielleicht sind Sie:

- Die stellvertretende Leiterin der Marketingabteilung, die die Daten der geschäftlichen Transaktionen strategischer als bisher für die Preisgestaltung und die Einteilung der Kunden nutzen möchte. Aber Sie verstehen die Vorgehensweise Ihrer Entwickler und der überbezahlten Berater nicht.
- Die Person, die Bedarfsprognosen untersucht und die weiß, dass sich in den Verkaufsdaten des letzten Quartals mehr über die Kunden des Unternehmens verbirgt als nur eine Vorschau für das nächste Quartal. Aber Sie wissen nicht, wie Sie an diese verborgenen Schätze gelangen können.
- Die Geschäftsführerin eines Online-Start-ups, die auf der Basis der letzten Einkäufe eines Kunden vorhersagen möchte, ob dieser Kunde auch am Kauf eines anderen Artikels interessiert sein könnte.
- Der für die Business Intelligence zuständige Analyst, der zusieht, wie viel Geld für Infrastrukturmaßnahmen und die Lieferkette des Unternehmens sinnlos ausgegeben wird, der aber nicht weiß, wie kostensparende Entscheidungen systematisch gefällt werden.
- Der Fachmann für Onlinemarketing, der mehr mit den E-Mail- oder Facebook- und Twitter-Reaktionen von Kunden

anfangen möchte, als sie nur zu lesen und abzuspeichern.

Ich stelle mir vor, dass Sie ein Leser sind, der direkten Nutzen daraus zieht, mehr über Data Science zu wissen, der es aber bisher noch nicht geschafft hat, einen Fuß in die Tür zu diesen Techniken zu bekommen. Sinn dieses Buches ist es, alle Irritationen zu beseitigen, die sich um Data Science ranken (den Code, die Werkzeuge und den ganzen Rummel), und Ihnen die entsprechenden Techniken beizubringen. Dabei verwende ich Fallstudien, die jeder verstehen kann, der sich in der Schule zumindest grundsätzlich mit linearer Algebra oder Infinitesimalrechnung beschäftigt hat. Sollte das bei Ihnen nicht der Fall sein, lesen Sie einfach langsamer und greifen Sie auf Wikipedia zu.

Nichts geht über eine Tabellenkalkulation

Dies ist kein Buch über das Codieren. Ich bin sogar bereit, dies (mit der kleinen Ausnahme von [Kapitel 10](#)) zu garantieren. Warum?

Ganz einfach: Ich habe kein Interesse daran, zu Beginn dieses Buches hundert Seiten damit zu vergeuden, mich mit Git abzugeben, Umgebungsvariablen einzurichten und den Spagat zwischen Emacs und Vi zu wagen.

Vielleicht laufen bei Ihnen nur Windows und Microsoft Office. Oder Sie sind bei einer Organisation beschäftigt, die es nicht zulässt, dass Sie auf Ihrem Computer irgendwelches Open-Source-Zeugs herunterladen und installieren. Und selbst wenn Ihnen in der Schule schon Ihr Taschenrechner eine Heidenangst einjagen konnte, müssen Sie sich keine Sorgen machen.

Sollten Sie wissen, wie Code geschrieben wird, um die meisten der hier vorgestellten Techniken in eine automatisierte, produktive Form zu bringen? Unbedingt! Auf

jeden Fall müssen Sie mindestens jemanden kennen, der mit Code umgehen kann und Speichertechnologien beherrscht. Müssen Sie wissen, wie Code geschrieben wird, um diese Techniken zu verstehen, sie zu unterscheiden und auf sie aufbauen zu können? Natürlich nicht!

Aus diesem Grund behandle ich jede Technik mithilfe einer Tabellenkalkulation.

Okay, in meinen Aussagen ist eine kleine Lüge versteckt. Das letzte Kapitel dieses Buches handelt von der auf Data Science ausgerichteten Programmiersprache R. Es soll denen unter Ihnen als Sprungbrett dienen, die sich intensiver mit Dingen dieser Art beschäftigen wollen.

Aber Tabellenkalkulationen sind doch aus der Mode!

Tabellenkalkulationen sind nicht gerade das aufregendste Werkzeug, das man sich vorstellen kann. Letztendlich gehören sie sogar zu den langweiligsten Analysewerkzeugen auf dieser Erde. Aber sie erlauben Ihnen, die Daten zu sehen und zu berühren (oder wenigstens anzuklicken). Wenn es darum geht, die entsprechenden Techniken kennenzulernen, benötigen Sie etwas Unspektakuläres, etwas, das jeder versteht und mit dem Sie gleichzeitig parallel zu Ihrem Lernfortschritt schnell und ohne großen Aufwand weiterkommen. Und genau das geht prima mit einer Tabellenkalkulation.

Tabellenkalkulationen sind ein erstklassiges Werkzeug, wenn es um das Entwickeln von ersten Ansätzen geht. Sie werden wohl kaum für Ihren Online-Vertrieb ein produktives KI-Modell aus Excel heraus ablaufen lassen, was aber nicht heißt, dass Sie sich in diesem Programm keine Verkaufsdaten anschauen können, nicht mit Funktionen herumspielen sollten, die das Interesse an Produkten vorhersagen können, und nicht in der Lage sind,

Zielvorgaben festzulegen. Um so etwas zu tun, bietet eine Tabellenkalkulation den perfekten Rahmen.

Verwenden Sie Excel oder LibreOffice

Alle Beispiele, die Sie durcharbeiten, setzen in diesem Buch Excel voraus. Auf der Webseite zu diesem Buch (www.wiley-vch.de/publish/dt/books/ISBN3-527-76060-1) können Sie zu den einzelnen Kapiteln Arbeitsmappen herunterladen, die Bestandteil einer großen Demodatei sind und die Ihnen leichter machen, die Aufgaben zu verfolgen. Wenn Sie dann vielleicht die Abenteuerlust packt, können Sie in den Arbeitsblättern alles bis auf die Anfangsdaten löschen und die gesamte Übung selbstständig nachvollziehen. Das Buch ist kompatibel zu Excel 2007, 2010, 2011 für den Mac und 2013. [Kapitel 1](#) geht genau auf die Unterschiede der einzelnen Versionen ein.

Die meisten von Ihnen haben Zugriff auf Excel, und vielleicht nutzen Sie es schon, um auf der Arbeit Berichte zu erstellen oder Daten festzuhalten. Wenn Sie aber aus irgendeinem Grund kein Excel besitzen, sollten Sie diese Software erwerben oder auf LibreOffice(www.libreoffice.org) zugreifen.

Hinweis

Was ist mit Google Drive?

Vielleicht denken einige von Ihnen darüber nach, Google Drive zu verwenden. Dies ist eine verlockende Möglichkeit, da sich Google Drive in der Cloud befindet und auch von Ihren mobilen Geräten aus erreichbar ist. Aber das, was wir hier vorhaben, funktioniert dort nicht.

Google Drive eignet sich gut für einfache Arbeitsblätter einer Tabellenkalkulation, aber dort, wo Sie sich hinbegeben, kann Google nicht folgen. Das Hinzufügen von Zeilen und Spalten ist in Drive eine mehr als nervige Sache, die Einbindung von Solver ist haarsträubend, und die Diagramme besitzen noch nicht einmal Trendlinien. Ich wünsche mir, es wäre anders.

Bei LibreOffice handelt es sich um kostenlose Open-Source-Software, die über fast dieselben Funktionen wie Excel verfügt. Ich bin sogar der Meinung, dass der Solver von LibreOffice dem von Excel vorzuziehen ist. Wenn Sie diesen Weg einschlagen wollen, hindert Sie nichts daran.

Konventionen

Damit Sie das meiste aus dem Text herausholen und den Geschehnissen auf der Spur bleiben können, verwende ich in diesem Buch einige Konventionen.

Hinweis

Informationen wie die gerade zu Google Drive beziehen sich in der Regel auf Themen auf der aktuellen Seite und ergänzen diese Themen.

Warnung

Warnungen enthalten wichtige Informationen, die Sie nicht vergessen dürfen, und die für den unmittelbar benachbarten Text von Bedeutung sind.

Tipp

Anmerkungen dieser Art enthalten Tipps, Hinweise, Tricks und Randbemerkungen, die zum aktuellen Thema gehören.

Ich verweise im Text so auf Codestückchen:

```
=VERKETTEN("Dies ist Text";"in Excel!")
```

Neue und/oder wichtige Begriffe werden bei ihrer ersten Verwendung optisch *hervorgehoben*. Dateinamen, Bezeichnungen von Verzeichnissen weisen ebenfalls diese kursive Formatierung auf, während auf URLs so hingewiesen wird: www.wiley-vch.de

Wenn es im Text um eine Formel wie =SUMME(A4:T32) oder um Funktionen oder Bezeichnungen geht, wird ebenfalls die »Formelschriftart« verwendet.

Los geht's

Im ersten Kapitel möchte ich weiße Flecken bei Ihren Excel-Kenntnissen mit Leben füllen. Danach geht es sofort mit Fallstudien los. Am Ende dieses Buches kennen Sie nicht nur die folgenden Techniken, sondern Sie wissen auch, wie sie von Grund auf eingerichtet werden:

- Lineare und ganzzahlige Optimierung
- Arbeiten mit Zeitreihen, Erkennen von Trends und saisonbedingten Mustern und Erstellen von Prognosen mithilfe von exponentiellen Glättungen