DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES AND APPLICATIONS IN PYTHON

GALIT SHMUELI | PETER C. BRUCE
PETER GEDECK | NITIN R. PATEL

WILEY

DATA MINING
FOR BUSINESS ANALYTICS

DATA MINING FOR BUSINESS ANALYTICS

Concepts, Techniques, and Applications in Python

GALIT SHMUELI
PETER C. BRUCE
PETER GEDECK
NITIN R. PATEL

WILEY

This edition first published 2020

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data applied for

Hardback: 9781119549840

Cover Design: Wiley
Cover Image: © Achim Mittler, Frankfurt am Main/Gettyimages

Set in 11.5/14.5pt BemboStd by Aptara Inc., New Delhi, India
Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

The beginning of wisdom is this:
Get wisdom, and whatever else you get, get insight.

רֵאשִׁית חָכְמָה, קְנֵה חָכְמָה; וּבְכָל-קִנְיָנְךָ, קְנֵה בִינָה.

– Proverbs 4:7

In memory of Professor Ayala Cohen (1940–2019)
who combined wisdom, insight, enthusiasm, and care

Peter Gedeck dedicates this book to his son, Victor

Contents

Foreword by Gareth James

The field of statistics has existed in one form or another for 200 years, and by the second half of the 20th century had evolved into a well-respected and essential academic discipline. However, its prominence expanded rapidly in the 1990s with the explosion of new, and enormous, data sources. For the first part of this century, much of this attention was focused on biological applications, in particular, genetics data generated as a result of the sequencing of the human genome. However, the last decade has seen a dramatic increase in the availability of data in the business disciplines, and a corresponding interest in business-related statistical applications.

The impact has been profound. Ten years ago, when I was able to attract a full class of MBA students to my new statistical learning elective, my colleagues were astonished because our department struggled to fill most electives. Today, we offer a Masters in Business Analytics, which is the largest specialized masters program in the school and has application volume rivaling those of our MBA programs. Our department's faculty size and course offerings have increased dramatically, yet the MBA students are still complaining that the classes are all full. Google's chief economist, Hal Varian, was indeed correct in 2009 when he stated that "the sexy job in the next 10 years will be statisticians."

This demand is driven by a simple, but undeniable, fact. Business analytics solutions have produced significant and measurable improvements in business performance, on multiple dimensions and in numerous settings, and as a result, there is a tremendous demand for individuals with the requisite skill set. However, training students in these skills is challenging given that, in addition to the obvious required knowledge of statistical methods, they need to understand business-related issues, possess strong communication skills, and be comfortable dealing with multiple computational packages. Most statistics texts concentrate on abstract training in classical methods, without much emphasis on practical, let alone business, applications.

This book has by far the most comprehensive review of business analytics methods that I have ever seen, covering everything from classical approaches such as linear and logistic regression, through to modern methods like neural

networks, bagging and boosting, and even much more business specific procedures such as social network analysis and text mining. If not the bible, it is at the least a definitive manual on the subject. However, just as important as the list of topics, is the way that they are all presented in an applied fashion using business applications. Indeed the last chapter is entirely dedicated to 10 separate cases where business analytics approaches can be applied.

In this latest edition, the authors have added support for Python, a programming language that is rapidly gaining popularity among data scientists. The book provides detailed descriptions and code involving applications of Python in numerous business settings, ensuring that the reader will actually be able to apply their knowledge to real-life problems. I'm confident that this book will be an indispensable tool for any business analytics course using Python.

We recently introduced a business analytics course into our required MBA core curriculum and I intend to make heavy use of this book in developing the syllabus. I'm confident that it will be an indispensable tool for any such course.

GARETH JAMES
Marshall School of Business, University of Southern California, 2019

Foreword by Ravi Bapna

Data is the new gold—and mining this gold to create business value in today's context of a highly networked and digital society requires a skillset that we haven't traditionally delivered in business or statistics or engineering programs on their own. For those businesses and organizations that feel overwhelmed by today's Big Data, the phrase *you ain't seen nothing yet* comes to mind. Yesterday's three major sources of Big Data—the 20+ years of investment in enterprise systems (ERP, CRM, SCM, etc.), the 3 billion plus people on the online social grid, and the close to 5 billion people carrying increasingly sophisticated mobile devices—are going to be dwarfed by tomorrow's smarter physical ecosystems fueled by the Internet of Things (IoT) movement.

The idea that we can use sensors to connect physical objects such as homes, automobiles, roads, even garbage bins and streetlights, to digitally optimized systems of governance goes hand in glove with bigger data and the need for deeper analytical capabilities. We are not far away from a smart refrigerator sensing that you are short on, say, eggs, populating your grocery store's mobile app's shopping list, and arranging a Task Rabbit to do a grocery run for you. Or the refrigerator negotiating a deal with an Uber driver to deliver an evening meal to you. Nor are we far away from sensors embedded in roads and vehicles that can compute traffic congestion, track roadway wear and tear, record vehicle use and factor these into dynamic usage-based pricing, insurance rates, and even taxation. This brave new world is going to be fueled by analytics and the ability to harness data for competitive advantage.

Business Analytics is an emerging discipline that is going to help us ride this new wave. This new Business Analytics discipline requires individuals who are grounded in the fundamentals of business such that they know the right questions to ask, who have the ability to harness, store, and optimally process vast datasets from a variety of structured and unstructured sources, and who can then use an array of techniques from machine learning and statistics to uncover new insights for decision-making. Such individuals are a rare commodity today, but their creation has been the focus of this book for a decade now. This book's forte is that it relies on explaining the core set of concepts required for today's business analytics professionals using real-world data-rich cases in a hands-on

manner, without sacrificing academic rigor. It provides a modern day foundation for Business Analytics, the notion of linking the x's to the y's of interest in a predictive sense. I say this with the confidence of someone who was probably the first adopter of the zeroth edition of this book (Spring 2006 at the Indian School of Business).

After the publication of the R edition in 2018, the new Python edition is an important addition. Python is gaining in popularity among analytics professionals, and the two open source languages constitute the primary statistical modeling and machine learning programming environments in data science.

I look forward to using the book in multiple fora, in executive education, in MBA classrooms, in MS-Business Analytics programs, and in Data Science bootcamps. I trust you will too!

RAVI BAPNA
Carlson School of Management, University of Minnesota, 2019

Preface to the Python Edition

This textbook first appeared in early 2007 and has been used by numerous students and practitioners and in many courses, including our own experience teaching this material both online and in person for more than 15 years. The first edition, based on the Excel add-in Analytic Solver Data Mining (previously XLMiner), was followed by two more Analytic Solver editions, a JMP edition, an R edition, and now this Python edition, with its companion website, www.dataminingbook.com.

This new Python edition, which relies on the free and open-source Python programming language, presents output from Python, as well as the code used to produce that output, including specification of the appropriate packages and functions, the dominant one being scikit-learn. Unlike computer-science or statistics-oriented textbooks, the focus in this book is on data mining concepts, and how to implement the associated algorithms in Python. We assume a basic familiarity with Python.

For this Python edition, a new co-author, Peter Gedeck comes on board bringing extensive data science experience in business. In addition to providing Python code and output, this edition also incorporates updates and new material based on feedback from instructors teaching MBA, MS, undergraduate, diploma, and executive courses, and from their students as well. Importantly, this edition includes for the first time an extended section on Data Ethics (Section 2.9).

A note about the book's title: The first two editions of the book used the title *Data Mining for Business Intelligence*. Business Intelligence today refers mainly to reporting and data visualization ("what is happening now"), while Business Analytics has taken over the "advanced analytics," which include predictive analytics and data mining. In this new edition, we therefore use the updated terms.

This Python edition includes the material that was recently added in the third edition of the original (Analytic Solver based) book:

- Social network analysis
- Text mining
- Ensembles

- Uplift modeling
- Collaborative filtering

Since the appearance of the (Analytic Solver based) second edition, the landscape of the courses using the textbook has greatly expanded: whereas initially, the book was used mainly in semester-long elective MBA-level courses, it is now used in a variety of courses in Business Analytics degrees and certificate programs, ranging from undergraduate programs, to post-graduate and executive education programs. Courses in such programs also vary in their duration and coverage. In many cases, this textbook is used across multiple courses. The book is designed to continue supporting the general "Predictive Analytics" or "Data Mining" course as well as supporting a set of courses in dedicated business analytics programs.

A general "Business Analytics," "Predictive Analytics," or "Data Mining" course, common in MBA and undergraduate programs as a one-semester elective, would cover Parts I–III, and choose a subset of methods from Parts IV and V. Instructors can choose to use cases as team assignments, class discussions, or projects. For a two-semester course, Part VI might be considered, and we recommend introducing the new Part VII (Data Analytics).

For a set of courses in a dedicated business analytics program, here are a few courses that have been using our book:

Predictive Analytics—Supervised Learning: In a dedicated Business Analytics program, the topic of Predictive Analytics is typically instructed across a set of courses. The first course would cover Parts I–IV and instructors typically choose a subset of methods from Part IV according to the course length. We recommend including the Chapter 13 on ensembles in such a course, as well as "Part VII: Data Analytics."

Predictive Analytics—Unsupervised Learning: This course introduces data exploration and visualization, dimension reduction, mining relationships, and clustering (Parts III and V). If this course follows the Predictive Analytics: Supervised Learning course, then it is useful to examine examples and approaches that integrate unsupervised and supervised learning, such as the new part on "Data Analytics."

Forecasting Analytics: A dedicated course on time series forecasting would rely on Part VI.

Advanced Analytics: A course that integrates the learnings from Predictive Analytics (supervised and unsupervised learning). Such a course can focus on Part VII: Data Analytics, where social network analytics and text mining are introduced. Some instructors choose to use the Cases (Chapter 21) in such a course.

In all courses, we strongly recommend including a project component, where data are either collected by students according to their interest or provided by the instructor (e.g., from the many data mining competition datasets available). From our experience and other instructors' experience, such projects enhance the learning and provide students with an excellent opportunity to understand the strengths of data mining and the challenges that arise in the process.

GALIT SHMUELI, PETER C. BRUCE, PETER GEDECK, AND NITIN R. PATEL
2019

Acknowledgments

We thank the many people who assisted us in improving the book from its inception as *Data Mining for Business Intelligence* in 2006 (using XLMiner, now Analytic Solver), through the recent editions now called *Data Mining for Business Analytics*, including two later XLMiner editions, a JMP edition, an R edition, and now for the first time, a Python edition.

Anthony Babinec, who has been using earlier editions of this book for years in his data mining courses at Statistics.com, provided us with detailed and expert corrections. Dan Toy and John Elder IV greeted our project with early enthusiasm and provided detailed and useful comments on initial drafts. Ravi Bapna, who used an early draft in a data mining course at the Indian School of Business and later at University of Minnesota, has provided invaluable comments and helpful suggestions since the book's start.

Many of the instructors, teaching assistants, and students using earlier editions of the book have contributed invaluable feedback both directly and indirectly, through fruitful discussions, learning journeys, and interesting data mining projects that have helped shape and improve the book. These include MBA students from the University of Maryland, MIT, the Indian School of Business, National Tsing Hua University, and Statistics.com. Instructors from many universities and teaching programs, too numerous to list, have supported and helped improve the book since its inception. Scott Nestler has been a helpful friend of this book project from the beginning.

Kuber Deokar, instructional operations supervisor at Statistics.com, has been unstinting in his assistance, support, and detailed attention. We also thank Anuja Kulkarni, assistant teacher at Statistics.com. Valerie Troiano has shepherded many instructors and students through the Statistics.com courses that have helped nurture the development of these books.

Colleagues and family members have been providing ongoing feedback and assistance with this book project. Boaz Shmueli and Raquelle Azran gave detailed editorial comments and suggestions on the first two editions; Bruce McCullough and Adam Hughes did the same for the first edition. Noa Shmueli provided careful proofs of the third edition. Ran Shenberger offered design tips. Che Lin and Boaz Shmueli provided feedback on Deep Learning. Ken Strasma,

founder of the microtargeting firm HaystaqDNA and director of targeting for the 2004 Kerry campaign and the 2008 Obama campaign, provided the scenario and data for the section on uplift modeling. We also thank Jen Golbeck, director of the Social Intelligence Lab at the University of Maryland and author of *Analyzing the Social Web*, whose book inspired our presentation in the chapter on social network analytics. Randall Pruim contributed extensively to the chapter on visualization. Inbal Yahav, co-author of the R edition, helped improve the social network analytics and text mining chapters.

Marietta Tretter at Texas A&M shared comments and thoughts on the time series chapters, and Stephen Few and Ben Shneiderman provided feedback and suggestions on the data visualization chapter and overall design tips.

Susan Palocsay and Mia Stephens have provided suggestions and feedback on numerous occasions, as have Margret Bjarnadottir, and, specifically for this Python edition, Mohammad Salehan. We also thank Catherine Plaisant at the University of Maryland's Human–Computer Interaction Lab, who helped out in a major way by contributing exercises and illustrations to the data visualization chapter. Gregory Piatetsky-Shapiro, founder of KDNuggets.com, has been generous with his time and counsel in the early years of this project.

We thank colleagues at the MIT Sloan School of Management for their support during the formative stage of this book—Dimitris Bertsimas, James Orlin, Robert Freund, Roy Welsch, Gordon Kaufmann, and Gabriel Bitran. As teaching assistants for the data mining course at Sloan, Adam Mersereau gave detailed comments on the notes and cases that were the genesis of this book, Romy Shioda helped with the preparation of several cases and exercises used here, and Mahesh Kumar helped with the material on clustering.

Colleagues at the University of Maryland's Smith School of Business: Shrivardhan Lele, Wolfgang Jank, and Paul Zantek provided practical advice and comments. We thank Robert Windle, and University of Maryland MBA students Timothy Roach, Pablo Macouzet, and Nathan Birckhead for invaluable datasets. We also thank MBA students Rob Whitener and Daniel Curtis for the heatmap and map charts.

Anand Bodapati provided both data and advice. Jake Hofman from Microsoft Research and Sharad Borle assisted with data access. Suresh Ankolekar and Mayank Shah helped develop several cases and provided valuable pedagogical comments. Vinni Bhandari helped write the Charles Book Club case.

We would like to thank Marvin Zelen, L. J. Wei, and Cyrus Mehta at Harvard, as well as Anil Gore at Pune University, for thought-provoking discussions on the relationship between statistics and data mining. Our thanks to Richard Larson of the Engineering Systems Division, MIT, for sparking many stimulating ideas on the role of data mining in modeling complex systems. Over two decades ago, they helped us develop a balanced philosophical perspective on the emerging field of data mining.