



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

© 2020 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Galit Shmueli, Peter C. Bruce, Peter Gedeck, and Nitin R. Patel to be identified as the authors of this work has been asserted in accordance with law.

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.

Limit of Liability/Disclaimer of Warranty

The publisher and the authors make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties; including without limitation any implied warranties of fitness for a particular purpose. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for every situation. In view of on-going research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or website is referred to in this work as a citation and/or potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising here from.

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.*

ראשית חכמה, קנה חכמה; ובכל-קניינך, קנה בינה.

– 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	xix
Foreword by Ravi Bapna	xxi
Preface to the Python Edition	xxiii
Acknowledgments	xxvii

PART I PRELIMINARIES

CHAPTER 1 Introduction	3
1.1 What Is Business Analytics?	3
1.2 What Is Data Mining?	5
1.3 Data Mining and Related Terms	5
1.4 Big Data	6
1.5 Data Science	7
1.6 Why Are There So Many Different Methods?	8
1.7 Terminology and Notation	9
1.8 Road Maps to This Book	11
Order of Topics	11
CHAPTER 2 Overview of the Data Mining Process	15
2.1 Introduction	15
2.2 Core Ideas in Data Mining	16
Classification	16
Prediction	16
Association Rules and Recommendation Systems	16
Predictive Analytics	17
Data Reduction and Dimension Reduction	17
Data Exploration and Visualization	17
Supervised and Unsupervised Learning	18
2.3 The Steps in Data Mining	19
2.4 Preliminary Steps	21
Organization of Datasets	21
Predicting Home Values in the West Roxbury Neighborhood	21

VIII CONTENTS

	Loading and Looking at the Data in Python	22
	Python Imports	25
	Sampling from a Database	25
	Oversampling Rare Events in Classification Tasks	26
	Preprocessing and Cleaning the Data	27
2.5	Predictive Power and Overfitting	34
	Overfitting	34
	Creation and Use of Data Partitions	36
2.6	Building a Predictive Model	40
	Modeling Process	40
2.7	Using Python for Data Mining on a Local Machine	44
2.8	Automating Data Mining Solutions	45
2.9	Ethical Practice in Data Mining	47
	Data Mining Software: The State of the Market (by Herb Edelstein)	52
	Problems	56

PART II DATA EXPLORATION AND DIMENSION REDUCTION

CHAPTER 3 Data Visualization 61

3.1	Introduction	61
3.2	Data Examples	64
	Example 1: Boston Housing Data	64
	Example 2: Ridership on Amtrak Trains	65
3.3	Basic Charts: Bar Charts, Line Graphs, and Scatter Plots	65
	Distribution Plots: Boxplots and Histograms	68
	Heatmaps: Visualizing Correlations and Missing Values	71
3.4	Multidimensional Visualization	74
	Adding Variables: Color, Size, Shape, Multiple Panels, and Animation	74
	Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering	77
	Reference: Trend Lines and Labels	81
	Scaling Up to Large Datasets	82
	Multivariate Plot: Parallel Coordinates Plot	83
	Interactive Visualization	83
3.5	Specialized Visualizations	88
	Visualizing Networked Data	88
	Visualizing Hierarchical Data: Treemaps	90
	Visualizing Geographical Data: Map Charts	91
3.6	Summary: Major Visualizations and Operations, by Data Mining Goal	93
	Prediction	93
	Classification	94
	Time Series Forecasting	96
	Unsupervised Learning	96
	Problems	97

CHAPTER 4 Dimension Reduction	99
4.1 Introduction	100
4.2 Curse of Dimensionality	100
4.3 Practical Considerations	100
Example 1: House Prices in Boston	101
4.4 Data Summaries	102
Summary Statistics	102
Aggregation and Pivot Tables	104
4.5 Correlation Analysis	105
4.6 Reducing the Number of Categories in Categorical Variables	106
4.7 Converting a Categorical Variable to a Numerical Variable	108
4.8 Principal Components Analysis	108
Example 2: Breakfast Cereals	109
Principal Components	114
Normalizing the Data	114
Using Principal Components for Classification and Prediction	117
4.9 Dimension Reduction Using Regression Models	119
4.10 Dimension Reduction Using Classification and Regression Trees	119
Problems	120

PART III PERFORMANCE EVALUATION

CHAPTER 5 Evaluating Predictive Performance	125
5.1 Introduction	126
5.2 Evaluating Predictive Performance	126
Naive Benchmark: The Average	127
Prediction Accuracy Measures	127
Comparing Training and Validation Performance	128
Cumulative Gains and Lift Charts	128
5.3 Judging Classifier Performance	131
Benchmark: The Naive Rule	132
Class Separation	133
The Confusion (Classification) Matrix	133
Using the Validation Data	134
Accuracy Measures	135
Propensities and Cutoff for Classification	136
Performance in Case of Unequal Importance of Classes	138
Asymmetric Misclassification Costs	140
Generalization to More Than Two Classes	144
5.4 Judging Ranking Performance	144
Gains and Lift Charts for Binary Data	144
Decile Lift Charts	147
Beyond Two Classes	148
Gains and Lift Charts Incorporating Costs and Benefits	148
Cumulative Gains as a Function of Cutoff	148

5.5	Oversampling	149
	Oversampling the Training Set	152
	Evaluating Model Performance Using a Non-oversampled Validation Set	152
	Evaluating Model Performance if Only Oversampled Validation Set Exists	152
	Problems	155

PART IV PREDICTION AND CLASSIFICATION METHODS

CHAPTER 6 Multiple Linear Regression 161

6.1	Introduction	162
6.2	Explanatory vs. Predictive Modeling	162
6.3	Estimating the Regression Equation and Prediction	164
	Example: Predicting the Price of Used Toyota Corolla Cars	165
6.4	Variable Selection in Linear Regression	169
	Reducing the Number of Predictors	169
	How to Reduce the Number of Predictors	170
	Regularization (Shrinkage Models)	176
	Appendix: Using Statmodels	179
	Problems	180

CHAPTER 7 *k*-Nearest Neighbors (*k*NN) 185

7.1	The <i>k</i> -NN Classifier (Categorical Outcome)	185
	Determining Neighbors	186
	Classification Rule	186
	Example: Riding Mowers	187
	Choosing <i>k</i>	188
	Setting the Cutoff Value	191
	<i>k</i> -NN with More Than Two Classes	192
	Converting Categorical Variables to Binary Dummies	193
7.2	<i>k</i> -NN for a Numerical Outcome	193
7.3	Advantages and Shortcomings of <i>k</i> -NN Algorithms	195
	Problems	197

CHAPTER 8 The Naive Bayes Classifier 199

8.1	Introduction	199
	Cutoff Probability Method	200
	Conditional Probability	200
	Example 1: Predicting Fraudulent Financial Reporting	201
8.2	Applying the Full (Exact) Bayesian Classifier	201
	Using the “Assign to the Most Probable Class” Method	202
	Using the Cutoff Probability Method	202
	Practical Difficulty with the Complete (Exact) Bayes Procedure	202
	Solution: Naive Bayes	203
	The Naive Bayes Assumption of Conditional Independence	204

Using the Cutoff Probability Method	204
Example 2: Predicting Fraudulent Financial Reports, Two Predictors	205
Example 3: Predicting Delayed Flights	206
8.3 Advantages and Shortcomings of the Naive Bayes Classifier	210
Problems	214
CHAPTER 9 Classification and Regression Trees	217
9.1 Introduction	218
Tree Structure	219
Decision Rules	219
Classifying a New Record	220
9.2 Classification Trees	220
Recursive Partitioning	220
Example 1: Riding Mowers	221
Measures of Impurity	223
9.3 Evaluating the Performance of a Classification Tree	228
Example 2: Acceptance of Personal Loan	228
Sensitivity Analysis Using Cross Validation	230
9.4 Avoiding Overfitting	232
Stopping Tree Growth	233
Fine-tuning Tree Parameters	234
Other Methods for Limiting Tree Size	236
9.5 Classification Rules from Trees	238
9.6 Classification Trees for More Than Two Classes	239
9.7 Regression Trees	239
Prediction	240
Measuring Impurity	240
Evaluating Performance	241
9.8 Improving Prediction: Random Forests and Boosted Trees	243
Random Forests	243
Boosted Trees	244
9.9 Advantages and Weaknesses of a Tree	246
Problems	248
CHAPTER 10 Logistic Regression	251
10.1 Introduction	252
10.2 The Logistic Regression Model	253
10.3 Example: Acceptance of Personal Loan	255
Model with a Single Predictor	255
Estimating the Logistic Model from Data: Computing Parameter Estimates	257
Interpreting Results in Terms of Odds (for a Profiling Goal)	259
10.4 Evaluating Classification Performance	261
Variable Selection	262
10.5 Logistic Regression for Multi-class Classification	264
Ordinal Classes	264

Nominal Classes	266
Comparing Ordinal and Nominal Models	267
10.6 Example of Complete Analysis: Predicting Delayed Flights	269
Data Preprocessing	270
Model Training	272
Model Interpretation	273
Model Performance	273
Variable Selection	276
Appendix: Using Statmodels	278
Problems	280

CHAPTER 11 Neural Nets 283

11.1 Introduction	284
11.2 Concept and Structure of a Neural Network	284
11.3 Fitting a Network to Data	285
Example 1: Tiny Dataset	285
Computing Output of Nodes	286
Preprocessing the Data	289
Training the Model	290
Example 2: Classifying Accident Severity	292
Avoiding Overfitting	295
Using the Output for Prediction and Classification	297
11.4 Required User Input	297
11.5 Exploring the Relationship Between Predictors and Outcome	299
11.6 Deep Learning	299
Convolutional Neural Networks (CNNs)	300
Local Feature Map	301
A Hierarchy of Features	302
The Learning Process	302
Unsupervised Learning	303
Conclusion	304
11.7 Advantages and Weaknesses of Neural Networks	305
Problems	306

CHAPTER 12 Discriminant Analysis 309

12.1 Introduction	310
Example 1: Riding Mowers	310
Example 2: Personal Loan Acceptance	310
12.2 Distance of a Record from a Class	311
12.3 Fisher’s Linear Classification Functions	314
12.4 Classification Performance of Discriminant Analysis	317
12.5 Prior Probabilities	318
12.6 Unequal Misclassification Costs	319
12.7 Classifying More Than Two Classes	319
Example 3: Medical Dispatch to Accident Scenes	319

12.8 Advantages and Weaknesses	322
Problems	324

CHAPTER 13 Combining Methods: Ensembles and Uplift Modeling 327

13.1 Ensembles	328
Why Ensembles Can Improve Predictive Power	329
Simple Averaging	330
Bagging	331
Boosting	331
Bagging and Boosting in Python	332
Advantages and Weaknesses of Ensembles	332
13.2 Uplift (Persuasion) Modeling	334
A–B Testing	334
Uplift	334
Gathering the Data	335
A Simple Model	336
Modeling Individual Uplift	337
Computing Uplift with Python	338
Using the Results of an Uplift Model	339
13.3 Summary	340
Problems	341

PART V MINING RELATIONSHIPS AMONG RECORDS

CHAPTER 14 Association Rules and Collaborative Filtering 345

14.1 Association Rules	346
Discovering Association Rules in Transaction Databases	346
Example 1: Synthetic Data on Purchases of Phone Faceplates	348
Generating Candidate Rules	348
The Apriori Algorithm	349
Selecting Strong Rules	349
Data Format	352
The Process of Rule Selection	353
Interpreting the Results	354
Rules and Chance	355
Example 2: Rules for Similar Book Purchases	357
14.2 Collaborative Filtering	357
Data Type and Format	359
Example 3: Netflix Prize Contest	360
User-Based Collaborative Filtering: “People Like You”	361
Item-Based Collaborative Filtering	363
Advantages and Weaknesses of Collaborative Filtering	364
Collaborative Filtering vs. Association Rules	366
14.3 Summary	368
Problems	370

CHAPTER 15 Cluster Analysis	375
15.1 Introduction	376
Example: Public Utilities	377
15.2 Measuring Distance Between Two Records	379
Euclidean Distance	380
Normalizing Numerical Measurements	380
Other Distance Measures for Numerical Data	381
Distance Measures for Categorical Data	383
Distance Measures for Mixed Data	384
15.3 Measuring Distance Between Two Clusters	385
Minimum Distance	385
Maximum Distance	385
Average Distance	385
Centroid Distance	385
15.4 Hierarchical (Agglomerative) Clustering	387
Single Linkage	388
Complete Linkage	388
Average Linkage	388
Centroid Linkage	389
Ward’s Method	389
Dendrograms: Displaying Clustering Process and Results	390
Validating Clusters	390
Limitations of Hierarchical Clustering	393
15.5 Non-Hierarchical Clustering: The <i>k</i> -Means Algorithm	395
Choosing the Number of Clusters (<i>k</i>)	396
Problems	401

PART VI FORECASTING TIME SERIES

CHAPTER 16 Handling Time Series	407
16.1 Introduction	408
16.2 Descriptive vs. Predictive Modeling	409
16.3 Popular Forecasting Methods in Business	409
Combining Methods	410
16.4 Time Series Components	410
Example: Ridership on Amtrak Trains	411
16.5 Data-Partitioning and Performance Evaluation	415
Benchmark Performance: Naive Forecasts	415
Generating Future Forecasts	416
Problems	419

CHAPTER 17 Regression-Based Forecasting	423
17.1 A Model with Trend	424
Linear Trend	424

Exponential Trend	426
Polynomial Trend	427
17.2 A Model with Seasonality	429
17.3 A Model with Trend and Seasonality	432
17.4 Autocorrelation and ARIMA Models	433
Computing Autocorrelation	434
Improving Forecasts by Integrating Autocorrelation Information	436
Evaluating Predictability	440
Problems	442
CHAPTER 18 Smoothing Methods	451
18.1 Introduction	452
18.2 Moving Average	452
Centered Moving Average for Visualization	452
Trailing Moving Average for Forecasting	453
Choosing Window Width (w)	455
18.3 Simple Exponential Smoothing	457
Choosing Smoothing Parameter α	458
Relation Between Moving Average and Simple Exponential Smoothing	460
18.4 Advanced Exponential Smoothing	460
Series with a Trend	460
Series with a Trend and Seasonality	461
Series with Seasonality (No Trend)	462
Problems	464
PART VII DATA ANALYTICS	
CHAPTER 19 Social Network Analytics	473
19.1 Introduction	473
19.2 Directed vs. Undirected Networks	475
19.3 Visualizing and Analyzing Networks	476
Plot Layout	476
Edge List	478
Adjacency Matrix	479
Using Network Data in Classification and Prediction	479
19.4 Social Data Metrics and Taxonomy	480
Node-Level Centrality Metrics	480
Egocentric Network	481
Network Metrics	483
19.5 Using Network Metrics in Prediction and Classification	485
Link Prediction	485
Entity Resolution	485
Collaborative Filtering	488
19.6 Collecting Social Network Data with Python	491
19.7 Advantages and Disadvantages	491
Problems	494

CHAPTER 20 Text Mining	495
20.1 Introduction	496
20.2 The Tabular Representation of Text: Term–Document Matrix and “Bag-of-Words”	496
20.3 Bag-of-Words vs. Meaning Extraction at Document Level	497
20.4 Preprocessing the Text	498
Tokenization	499
Text Reduction	501
Presence/Absence vs. Frequency	501
Term Frequency–Inverse Document Frequency (TF-IDF)	502
From Terms to Concepts: Latent Semantic Indexing	505
Extracting Meaning	505
20.5 Implementing Data Mining Methods	506
20.6 Example: Online Discussions on Autos and Electronics	506
Importing and Labeling the Records	507
Text Preprocessing in Python	508
Producing a Concept Matrix	508
Fitting a Predictive Model	508
Prediction	509
20.7 Summary	510
Problems	511

PART VIII CASES

CHAPTER 21 Cases	515
21.1 Charles Book Club	515
The Book Industry	515
Database Marketing at Charles	516
Data Mining Techniques	518
Assignment	520
21.2 German Credit	522
Background	522
Data	522
Assignment	526
21.3 Tayko Software Cataloger	527
Background	527
The Mailing Experiment	527
Data	527
Assignment	529
21.4 Political Persuasion	531
Background	531
Predictive Analytics Arrives in US Politics	531
Political Targeting	531
Uplift	532
Data	533
Assignment	533

21.5	Taxi Cancellations	535
	Business Situation	535
	Assignment	535
21.6	Segmenting Consumers of Bath Soap	537
	Business Situation	537
	Key Problems	537
	Data	538
	Measuring Brand Loyalty	538
	Assignment	538
21.7	Direct-Mail Fundraising	541
	Background	541
	Data	541
	Assignment	541
21.8	Catalog Cross-Selling	544
	Background	544
	Assignment	544
21.9	Time Series Case: Forecasting Public Transportation Demand	546
	Background	546
	Problem Description	546
	Available Data	546
	Assignment Goal	546
	Assignment	547
	Tips and Suggested Steps	547
	References	549
	Data Files Used in the Book	551
	Python Utilities Functions	555
	Index	565



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 Birkhead 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.