

MACHINE LEARNING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES,
AND APPLICATIONS IN R

SECOND EDITION

Galit Shmueli • Peter C. Bruce
Peter Gedeck • Inbal Yahav
Nitin R. Patel



WILEY

**MACHINE LEARNING
FOR BUSINESS ANALYTICS**

MACHINE LEARNING FOR BUSINESS ANALYTICS

Concepts, Techniques, and Applications in R

Second Edition

GALIT SHMUELI

National Tsing Hua University

PETER C. BRUCE

statistics.com

PETER GEDECK

Collaborative Drug Discovery

INBAL YAHAV

Tel Aviv University

NITIN R. PATEL

Cytel, Inc.

WILEY

This edition first published 2023
© 2023 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, Inbal Yahav, 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

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.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

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: 9781119835172

Cover Design: Wiley

Cover Image: © Hiroshi Watanabe/Getty Images

Set in 11.5/14.5pt BemboStd by Straive, Chennai, India

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

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

–Proverbs 4:7



Contents

Foreword by Ravi Bapna	xix
Foreword by Gareth James	xxi
Preface to the Second R Edition	xxiii
Acknowledgments	xxvi

PART I PRELIMINARIES

CHAPTER 1 Introduction	3
1.1 What Is Business Analytics?	3
1.2 What Is Machine Learning?	5
1.3 Machine Learning, AI, and Related Terms	5
1.4 Big Data	7
1.5 Data Science	8
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	13
CHAPTER 2 Overview of the Machine Learning Process	17
2.1 Introduction	17
2.2 Core Ideas in Machine Learning	18
Classification	18
Prediction	18
Association Rules and Recommendation Systems	18
Predictive Analytics	19
Data Reduction and Dimension Reduction	19
Data Exploration and Visualization	19
Supervised and Unsupervised Learning	20
2.3 The Steps in a Machine Learning Project	21
2.4 Preliminary Steps	23
Organization of Data	23
Predicting Home Values in the West Roxbury Neighborhood	23
Loading and Looking at the Data in R	24
Sampling from a Database	26
Oversampling Rare Events in Classification Tasks	27
Preprocessing and Cleaning the Data	28
2.5 Predictive Power and Overfitting	35
Overfitting	36

	Creating and Using Data Partitions	38
2.6	Building a Predictive Model	41
	Modeling Process	41
2.7	Using R for Machine Learning on a Local Machine	46
2.8	Automating Machine Learning Solutions	47
	Predicting Power Generator Failure	48
	Uber’s Michelangelo	50
2.9	Ethical Practice in Machine Learning	52
	Machine Learning Software: The State of the Market (by Herb Edelstein)	53
	Problems	57

PART II DATA EXPLORATION AND DIMENSION REDUCTION

CHAPTER 3 Data Visualization 63

3.1	Uses of Data Visualization	63
	Base R or ggplot?	65
3.2	Data Examples	65
	Example 1: Boston Housing Data	65
	Example 2: Ridership on Amtrak Trains	67
3.3	Basic Charts: Bar Charts, Line Charts, and Scatter Plots	67
	Distribution Plots: Boxplots and Histograms	70
	Heatmaps: Visualizing Correlations and Missing Values	73
3.4	Multidimensional Visualization	75
	Adding Variables: Color, Size, Shape, Multiple Panels, and Animation	76
	Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering	79
	Reference: Trend Lines and Labels	83
	Scaling Up to Large Datasets	85
	Multivariate Plot: Parallel Coordinates Plot	85
	Interactive Visualization	88
3.5	Specialized Visualizations	91
	Visualizing Networked Data	91
	Visualizing Hierarchical Data: Treemaps	93
	Visualizing Geographical Data: Map Charts	95
3.6	Major Visualizations and Operations, by Machine Learning Goal	97
	Prediction	97
	Classification	97
	Time Series Forecasting	97
	Unsupervised Learning	98
	Problems	99

CHAPTER 4 Dimension Reduction 101

4.1	Introduction	101
4.2	Curse of Dimensionality	102
4.3	Practical Considerations	102
	Example 1: House Prices in Boston	103
4.4	Data Summaries	103
	Summary Statistics	104
	Aggregation and Pivot Tables	104
4.5	Correlation Analysis	107
4.6	Reducing the Number of Categories in Categorical Variables	109

4.7	Converting a Categorical Variable to a Numerical Variable	111
4.8	Principal Component Analysis	111
	Example 2: Breakfast Cereals	111
	Principal Components	116
	Normalizing the Data	117
	Using Principal Components for Classification and Prediction	120
4.9	Dimension Reduction Using Regression Models	121
4.10	Dimension Reduction Using Classification and Regression Trees	121
	Problems	123

PART III PERFORMANCE EVALUATION

CHAPTER 5 Evaluating Predictive Performance 129

5.1	Introduction	130
5.2	Evaluating Predictive Performance	130
	Naive Benchmark: The Average	131
	Prediction Accuracy Measures	131
	Comparing Training and Holdout Performance	133
	Cumulative Gains and Lift Charts	133
5.3	Judging Classifier Performance	136
	Benchmark: The Naive Rule	136
	Class Separation	136
	The Confusion (Classification) Matrix	137
	Using the Holdout Data	138
	Accuracy Measures	139
	Propensities and Threshold for Classification	139
	Performance in Case of Unequal Importance of Classes	143
	Asymmetric Misclassification Costs	146
	Generalization to More Than Two Classes	149
5.4	Judging Ranking Performance	150
	Cumulative Gains and Lift Charts for Binary Data	150
	Decile-wise Lift Charts	153
	Beyond Two Classes	154
	Gains and Lift Charts Incorporating Costs and Benefits	154
	Cumulative Gains as a Function of Threshold	155
5.5	Oversampling	156
	Creating an Over-sampled Training Set	158
	Evaluating Model Performance Using a Non-oversampled Holdout Set	159
	Evaluating Model Performance If Only Oversampled Holdout Set Exists	159
	Problems	162

PART IV PREDICTION AND CLASSIFICATION METHODS

CHAPTER 6 Multiple Linear Regression 167

6.1	Introduction	167
6.2	Explanatory vs. Predictive Modeling	168
6.3	Estimating the Regression Equation and Prediction	170
	Example: Predicting the Price of Used Toyota Corolla Cars	171
	Cross-validation and <code>caret</code>	175
6.4	Variable Selection in Linear Regression	176

Reducing the Number of Predictors	176
How to Reduce the Number of Predictors	178
Regularization (Shrinkage Models)	183
Problems	188

CHAPTER 7 *k*-Nearest Neighbors (*k*NN) 193

7.1 The <i>k</i> -NN Classifier (Categorical Outcome)	193
Determining Neighbors	194
Classification Rule	194
Example: Riding Mowers	195
Choosing <i>k</i>	196
Weighted <i>k</i> -NN	199
Setting the Cutoff Value	200
<i>k</i> -NN with More Than Two Classes	201
Converting Categorical Variables to Binary Dummies	201
7.2 <i>k</i> -NN for a Numerical Outcome	201
7.3 Advantages and Shortcomings of <i>k</i> -NN Algorithms	204
Problems	205

CHAPTER 8 The Naive Bayes Classifier 207

8.1 Introduction	207
Threshold Probability Method	208
Conditional Probability	208
Example 1: Predicting Fraudulent Financial Reporting	208
8.2 Applying the Full (Exact) Bayesian Classifier	209
Using the “Assign to the Most Probable Class” Method	210
Using the Threshold Probability Method	210
Practical Difficulty with the Complete (Exact) Bayes Procedure	210
8.3 Solution: Naive Bayes	211
The Naive Bayes Assumption of Conditional Independence	212
Using the Threshold Probability Method	212
Example 2: Predicting Fraudulent Financial Reports, Two Predictors	213
Example 3: Predicting Delayed Flights	214
Working with Continuous Predictors	218
8.4 Advantages and Shortcomings of the Naive Bayes Classifier	220
Problems	223

CHAPTER 9 Classification and Regression Trees 225

9.1 Introduction	226
Tree Structure	227
Decision Rules	227
Classifying a New Record	227
9.2 Classification Trees	228
Recursive Partitioning	228
Example 1: Riding Mowers	228
Measures of Impurity	231
9.3 Evaluating the Performance of a Classification Tree	235
Example 2: Acceptance of Personal Loan	236
9.4 Avoiding Overfitting	239
Stopping Tree Growth	242

Pruning the Tree	243
Best-Pruned Tree	245
9.5 Classification Rules from Trees	247
9.6 Classification Trees for More Than Two Classes	248
9.7 Regression Trees	249
Prediction	250
Measuring Impurity	250
Evaluating Performance	250
9.8 Advantages and Weaknesses of a Tree	250
9.9 Improving Prediction: Random Forests and Boosted Trees	252
Random Forests	252
Boosted Trees	254
Problems	257
CHAPTER 10 Logistic Regression	261
10.1 Introduction	261
10.2 The Logistic Regression Model	263
10.3 Example: Acceptance of Personal Loan	264
Model with a Single Predictor	265
Estimating the Logistic Model from Data: Computing Parameter Estimates	267
Interpreting Results in Terms of Odds (for a Profiling Goal)	270
10.4 Evaluating Classification Performance	271
10.5 Variable Selection	273
10.6 Logistic Regression for Multi-Class Classification	274
Ordinal Classes	275
Nominal Classes	276
10.7 Example of Complete Analysis: Predicting Delayed Flights	277
Data Preprocessing	282
Model-Fitting and Estimation	282
Model Interpretation	282
Model Performance	284
Variable Selection	285
Problems	289
CHAPTER 11 Neural Nets	293
11.1 Introduction	293
11.2 Concept and Structure of a Neural Network	294
11.3 Fitting a Network to Data	295
Example 1: Tiny Dataset	295
Computing Output of Nodes	296
Preprocessing the Data	299
Training the Model	300
Example 2: Classifying Accident Severity	304
Avoiding Overfitting	305
Using the Output for Prediction and Classification	305
11.4 Required User Input	307
11.5 Exploring the Relationship Between Predictors and Outcome	308
11.6 Deep Learning	309
Convolutional Neural Networks (CNNs)	310
Local Feature Map	311

A Hierarchy of Features	311
The Learning Process	312
Unsupervised Learning	312
Example: Classification of Fashion Images	313
Conclusion	320
11.7 Advantages and Weaknesses of Neural Networks	320
Problems	322

CHAPTER 12 Discriminant Analysis 325

12.1 Introduction	325
Example 1: Riding Mowers	326
Example 2: Personal Loan Acceptance	327
12.2 Distance of a Record from a Class	327
12.3 Fisher’s Linear Classification Functions	329
12.4 Classification Performance of Discriminant Analysis	333
12.5 Prior Probabilities	334
12.6 Unequal Misclassification Costs	334
12.7 Classifying More Than Two Classes	336
Example 3: Medical Dispatch to Accident Scenes	336
12.8 Advantages and Weaknesses	339
Problems	341

CHAPTER 13 Generating, Comparing, and Combining Multiple Models 345

13.1 Ensembles	346
Why Ensembles Can Improve Predictive Power	346
Simple Averaging or Voting	348
Bagging	349
Boosting	349
Bagging and Boosting in R	349
Stacking	350
Advantages and Weaknesses of Ensembles	351
13.2 Automated Machine Learning (AutoML)	352
AutoML: Explore and Clean Data	352
AutoML: Determine Machine Learning Task	353
AutoML: Choose Features and Machine Learning Methods	354
AutoML: Evaluate Model Performance	354
AutoML: Model Deployment	356
Advantages and Weaknesses of Automated Machine Learning	357
13.3 Explaining Model Predictions	358
13.4 Summary	360
Problems	362

PART V INTERVENTION AND USER FEEDBACK

CHAPTER 14 Interventions: Experiments, Uplift Models, and Reinforcement Learning 367

14.1 A/B Testing	368
Example: Testing a New Feature in a Photo Sharing App	369

	The Statistical Test for Comparing Two Groups (T-Test)	370
	Multiple Treatment Groups: A/B/n Tests	372
	Multiple A/B Tests and the Danger of Multiple Testing	372
14.2	Uplift (Persuasion) Modeling	373
	Gathering the Data	374
	A Simple Model	376
	Modeling Individual Uplift	376
	Computing Uplift with R	378
	Using the Results of an Uplift Model	378
14.3	Reinforcement Learning	380
	Explore-Exploit: Multi-armed Bandits	380
	Example of Using a Contextual Multi-Arm Bandit for Movie Recommendations	382
	Markov Decision Process (MDP)	383
14.4	Summary	388
	Problems	390

PART VI MINING RELATIONSHIPS AMONG RECORDS

CHAPTER 15 Association Rules and Collaborative Filtering 393

15.1	Association Rules	394
	Discovering Association Rules in Transaction Databases	394
	Example 1: Synthetic Data on Purchases of Phone Faceplates	394
	Generating Candidate Rules	395
	The Apriori Algorithm	397
	Selecting Strong Rules	397
	Data Format	399
	The Process of Rule Selection	400
	Interpreting the Results	401
	Rules and Chance	403
	Example 2: Rules for Similar Book Purchases	405
15.2	Collaborative Filtering	407
	Data Type and Format	407
	Example 3: Netflix Prize Contest	408
	User-Based Collaborative Filtering: “People Like You”	409
	Item-Based Collaborative Filtering	411
	Evaluating Performance	412
	Example 4: Predicting Movie Ratings with MovieLens Data	413
	Advantages and Weaknesses of Collaborative Filtering	416
	Collaborative Filtering vs. Association Rules	417
15.3	Summary	419
	Problems	421

CHAPTER 16 Cluster Analysis 425

16.1	Introduction	426
	Example: Public Utilities	427
16.2	Measuring Distance Between Two Records	429
	Euclidean Distance	429
	Normalizing Numerical Variables	430
	Other Distance Measures for Numerical Data	432
	Distance Measures for Categorical Data	433
	Distance Measures for Mixed Data	434

16.3	Measuring Distance Between Two Clusters	434
	Minimum Distance	434
	Maximum Distance	435
	Average Distance	435
	Centroid Distance	435
16.4	Hierarchical (Agglomerative) Clustering	437
	Single Linkage	437
	Complete Linkage	438
	Average Linkage	438
	Centroid Linkage	438
	Ward’s Method	438
	Dendrograms: Displaying Clustering Process and Results	439
	Validating Clusters	441
	Limitations of Hierarchical Clustering	443
16.5	Non-Hierarchical Clustering: The k -Means Algorithm	444
	Choosing the Number of Clusters (k)	445
	Problems	450

PART VII FORECASTING TIME SERIES

CHAPTER 17 Handling Time Series 455

17.1	Introduction	455
17.2	Descriptive vs. Predictive Modeling	457
17.3	Popular Forecasting Methods in Business	457
	Combining Methods	457
17.4	Time Series Components	458
	Example: Ridership on Amtrak Trains	458
17.5	Data Partitioning and Performance Evaluation	463
	Benchmark Performance: Naive Forecasts	463
	Generating Future Forecasts	465
	Problems	466

CHAPTER 18 Regression-Based Forecasting 469

18.1	A Model with Trend	469
	Linear Trend	469
	Exponential Trend	473
	Polynomial Trend	474
18.2	A Model with Seasonality	476
18.3	A Model with Trend and Seasonality	478
18.4	Autocorrelation and ARIMA Models	479
	Computing Autocorrelation	480
	Improving Forecasts by Integrating Autocorrelation Information	483
	Evaluating Predictability	486
	Problems	489

CHAPTER 19 Smoothing and Deep Learning Methods for Forecasting 499

19.1	Smoothing Methods: Introduction	500
19.2	Moving Average	500
	Centered Moving Average for Visualization	500

	Trailing Moving Average for Forecasting	501
	Choosing Window Width (w)	504
19.3	Simple Exponential Smoothing	505
	Choosing Smoothing Parameter α	506
	Relation Between Moving Average and Simple Exponential Smoothing	506
19.4	Advanced Exponential Smoothing	507
	Series with a Trend	508
	Series with a Trend and Seasonality	508
	Series with Seasonality (No Trend)	509
19.5	Deep Learning for Forecasting	511
	Problems	516

PART VIII DATA ANALYTICS

CHAPTER 20 Social Network Analytics 527

20.1	Introduction	527
20.2	Directed vs. Undirected Networks	529
20.3	Visualizing and Analyzing Networks	530
	Plot Layout	530
	Edge List	533
	Adjacency Matrix	533
	Using Network Data in Classification and Prediction	534
20.4	Social Data Metrics and Taxonomy	534
	Node-Level Centrality Metrics	535
	Egocentric Network	536
	Network Metrics	536
20.5	Using Network Metrics in Prediction and Classification	538
	Link Prediction	538
	Entity Resolution	540
	Collaborative Filtering	542
20.6	Collecting Social Network Data with R	545
20.7	Advantages and Disadvantages	545
	Problems	548

CHAPTER 21 Text Mining 549

21.1	Introduction	549
21.2	The Tabular Representation of Text	550
21.3	Bag-of-Words vs. Meaning Extraction at Document Level	551
21.4	Preprocessing the Text	552
	Tokenization	553
	Text Reduction	555
	Presence/Absence vs. Frequency	556
	Term Frequency–Inverse Document Frequency (TF-IDF)	557
	From Terms to Concepts: Latent Semantic Indexing	558
	Extracting Meaning	559
	From Terms to High-Dimensional Word Vectors: Word2Vec or GloVe	559
21.5	Implementing Machine Learning Methods	560
21.6	Example: Online Discussions on Autos and Electronics	560
	Importing and Labeling the Records	561
	Text Preprocessing in R	561

Producing a Concept Matrix	561
Fitting a Predictive Model	562
Prediction	564
21.7 Example: Sentiment Analysis of Movie Reviews	564
Data Loading, Preparation, and Partitioning	565
Generating and Applying the GloVe Model	565
Fitting a Predictive Model	566
21.8 Summary	568
Problems	570

CHAPTER 22 Responsible Data Science 573

22.1 Introduction	573
22.2 Unintentional Harm	574
22.3 Legal Considerations	576
22.4 Principles of Responsible Data Science	577
Non-maleficence	578
Fairness	578
Transparency	579
Accountability	580
Data Privacy and Security	580
22.5 A Responsible Data Science Framework	580
Justification	581
Assembly	581
Data Preparation	582
Modeling	583
Auditing	583
22.6 Documentation Tools	584
Impact Statements	584
Model Cards	585
Datasheets	586
Audit Reports	586
22.7 Example: Applying the RDS Framework to the COMPAS Example	588
Unanticipated Uses	588
Ethical Concerns	588
Protected Groups	588
Data Issues	589
Fitting the Model	589
Auditing the Model	591
Bias Mitigation	596
22.8 Summary	598
Problems	599

PART IX CASES

CHAPTER 23 Cases 603

23.1 Charles Book Club	603
The Book Industry	603
Database Marketing at Charles	604
Machine Learning Techniques	606
Assignment	608

23.2	German Credit	610
	Background	610
	Data	610
	Assignment	614
23.3	Tayko Software Cataloger	615
	Background	615
	The Mailing Experiment	615
	Data	615
	Assignment	617
23.4	Political Persuasion	619
	Background	619
	Predictive Analytics Arrives in US Politics	619
	Political Targeting	619
	Uplift	620
	Data	621
	Assignment	621
23.5	Taxi Cancellations	623
	Business Situation	623
	Assignment	623
23.6	Segmenting Consumers of Bath Soap	625
	Business Situation	625
	Key Problems	625
	Data	626
	Measuring Brand Loyalty	626
	Assignment	626
23.7	Direct-Mail Fundraising	629
	Background	629
	Data	629
	Assignment	629
23.8	Catalog Cross-Selling	632
	Background	632
	Assignment	632
23.9	Time Series Case: Forecasting Public Transportation Demand	634
	Background	634
	Problem Description	634
	Available Data	634
	Assignment Goal	634
	Assignment	635
	Tips and Suggested Steps	635
23.10	Loan Approval	636
	Background	636
	Regulatory Requirements	636
	Getting Started	636
	Assignment	637
	References	639
	R Packages Used in the Book	641
	Data Files Used in the Book	645
	Index	647



Foreword by Ravi Bapna

Converting data into an asset is the new business imperative facing modern managers. Each day the gap between what analytics capabilities make possible and companies' absorptive capacity of creating value from such capabilities increases. In many ways, 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 three billion plus people on the online social grid, and the close to five 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, and 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).

The updated R version is much awaited. R is used by a wide variety of instructors in our MS-Business Analytics program. The open-innovation paradigm used by R is one key part of the analytics perfect storm, the other components being the advances in computing and the business appetite for data-driven decision-making.

The new addition also covers causal analytics as experimentation (often called A/B testing in the industry), which is now becoming mainstream in the tech companies. Further, the authors have added a new chapter on Responsible Data Science, a new part on AutoML, more on deep learning and beefed up deep learning examples in the text mining and forecasting chapters. These updates make this new edition "state of the art" with respect to modern business analytics and AI.

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, 2022



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. Fifteen 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 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 an important new dimension in the form of the R software package. Easily the most widely used and influential open source statistical software, R has become the go-to tool for such purposes. With literally hundreds of freely available add-on packages, R can be used for almost any business analytics related problem. The book provides detailed descriptions and code involving applications of R in numerous business settings, ensuring that the reader will actually be able to apply their knowledge to real-life problems.

I would strongly recommend this book. I'm confident that it will be an indispensable tool for any MBA or business analytics course.

GARETH JAMES

Goizueta Business School, Emory University, 2022



Preface to the Second R 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, a Python edition, a RapidMiner edition, and now this new R edition, with its companion website, www.dataminingbook.com.

This new R edition, which relies on the free and open source R software, presents output from R, as well as the code used to produce that output, including specification of a variety of packages and functions. Unlike computer-science or statistics-oriented textbooks, the focus in this book is on machine learning concepts and how to implement the associated algorithms in R. We assume a basic familiarity with R.

For this new R edition, a new co-author, Peter Gedeck, comes on board bringing extensive data science experience in business.

The new edition provides significant updates both in terms of R and in terms of new topics and content. In addition to updating R code and routines that have changed or become available since the first edition, the new edition provides the following:

- A stronger focus on model selection using cross-validation with the use of the `caret` package
- Streamlined data preprocessing using tidyverse style
- Data visualization using `ggplot`
- Names of R packages, functions, and arguments are highlighted in the text, for easy readability.

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. Importantly, this edition includes several new topics:

- A dedicated section on *deep learning* in Chapter 11, with additional deep learning examples in text mining (Chapter 21) and time series forecasting (Chapter 19).
- A new chapter on *Responsible Data Science* (Chapter 22) covering topics of fairness, transparency, model cards and datasheets, legal considerations, and more, with an illustrative example.
- The *Performance Evaluation* exposition in Chapter 5 was expanded to include further metrics (precision and recall, F1).
- A new chapter on *Generating, Comparing, and Combining Multiple Models* (Chapter 13) that covers ensembles, AutoML, and explaining model predictions.
- A new chapter dedicated to *Interventions and User Feedback* (Chapter 14), that covers A/B tests, uplift modeling, and reinforcement learning.
- A new case (Loan Approval) that touches on regulatory and ethical issues.

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. Later editions were therefore renamed *Data Mining for Business Analytics*. However, the recent AI transformation has made the term *machine learning* more popularly associated with the methods in this textbook. In this new edition, we therefore use the updated terms *Machine Learning* and *Business Analytics*.

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 postgraduate 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 “machine learning” 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 VII might be considered, and we recommend introducing Part VIII (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–III, and instructors typically choose a subset of methods from Part IV according to the course length. We recommend including Part VIII: Data Analytics.

Predictive Analytics—Unsupervised Learning: This course introduces data exploration and visualization, dimension reduction, mining relationships, and clustering (Parts II and VI). 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 Part VIII 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) can focus on Part VIII: Data Analytics, where social network analytics and text mining are introduced, and responsible data science is discussed. Such a course might also include Chapter 13, Generating, Comparing, and Combining Multiple Models from Part IV, as well as Part V, which covers experiments, uplift modeling, and reinforcement learning. Some instructors choose to use the cases (Chapter 23) 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 machine learning 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 machine learning and the challenges that arise in the process.

GALIT SHMUELI, PETER C. BRUCE, PETER GEDECK, INBAL YAHAV, AND NITIN R. PATEL
2022



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), its reincarnation as *Data Mining for Business Analytics*, and now *Machine Learning for Business Analytics*, including translations in Chinese and Korean and versions supporting Analytic Solver Data Mining, R, Python, SAS JMP, and RapidMiner.

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.

Several professors have been especially helpful with the first R edition: Hayri Tongarlak, Prashant Joshi (UKA Tarsadia University), Jay Annadatha, Roger Bohn, Sridhar Vaithianathan, Travis Greene, and Dianne Cook provided detailed comments and/or R code files for the companion website; 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, Poonam Patil, and Shweta Jadhav, assistant teachers. 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. Vijay Kamble at UIC and Travis Greene at NTHU have provided valuable help with the section on reinforcement learning. 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. 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, Professor in the College of Information Studies 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.

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 has Margret Bjarnadottir. 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, was generous with his time and counsel in the early years of this project.

We thank colleagues at the Sloan School of Management at MIT 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 are grateful to colleagues at UMass Lowell's Manning School of Business for their encouragement and support in developing data analytics courses at the undergraduate and graduate levels that led to the development of this edition: Luvai Motiwalla, Harry Zhu, Thomas Sloan, Bob Li, and Sandra Richtermeyer. We also thank Michael Goul (late), Dan Power (late), Ramesh Sharda, Babita Gupta, Ashish Gupta, and Haya Ajjan from the Association for Information System's Decision Support and Analytics (SIGDSA) community for ideas and advice that helped the development of the book.

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 machine learning.

Lastly, we thank the folks at Wiley for this successful journey of nearly two decades. Steve Quigley at Wiley showed confidence in this book from the beginning and helped us navigate through the publishing process with great speed. Curt Hinrichs' vision, tips, and encouragement helped bring the first edition of this book to the starting gate. Jon Gurstelle guided us through additional editions and translations. Brett Kurzman has taken over the reins and is now shepherding the project. Becky Cowan, Sarah Lomore, Kathleen Pagliaro, Katrina Maceda, and Kavya Ramu greatly assisted us in pushing ahead and finalizing this and the earlier R edition. We are also especially grateful to Amy Hendrickson, who assisted with typesetting and making this book beautiful.