



Galit Shmueli
Peter C. Bruce
Kuber R. Deokar
Nitin R. Patel

FOURTH EDITION

MACHINE LEARNING FOR
BUSINESS ANALYTICS

Concepts, Techniques, and
Applications with **Analytic Solver® Data Mining**

WILEY

**MACHINE LEARNING
FOR BUSINESS ANALYTICS**

MACHINE LEARNING FOR BUSINESS ANALYTICS

**Concepts, Techniques, and Applications with
Analytic Solver[®] Data Mining**

Fourth Edition

GALIT SHMUELI

National Tsing Hua University
Taipei, Taiwan

PETER C. BRUCE

statistics.com
Arlington, USA

KUBER R. DEOKAR

UpThink Edutech Services Pvt. Ltd.
Pune, India

NITIN R. PATEL

Cytel, Inc.
Cambridge, USA

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, Kuber R. Deokar, and Nitin R. Patel to be identified as the authors of this work has been asserted in accordance with law.

Registered Office

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

In view of ongoing 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. While the publisher and authors have used their best efforts in preparing this work, they 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 merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. 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 your situation. You should consult with a specialist where appropriate. 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. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data Applied for:

Hardback: 9781119829836

Cover Design: Wiley

Cover Image: © Suthat_Chaitaweasap/Getty Images

Set in 11.5/14.5pt Bembo Std by Straive, Chennai, India

ANALYTIC SOLVER DATA MINING (ASDM) FOR EDUCATION

Your new textbook, *Machine Learning for Business Analytics: Concepts, Techniques, and Applications with Analytic Solver[®] Data Mining*, Fourth Edition, uses this software throughout. Here's how to get it for your course.

For Instructors: Setting Up the Course Code

Students in your course will need to purchase their license for \$25 (140 days in length), or the school itself may pre-purchase student licenses by contacting sales@solver.com and provide students with the course code. Course codes **MUST** be renewed each time you teach your course.

A pre-purchase will work in a similar fashion to the students purchasing themselves, course code and all, but at the purchase page students will see a charge of \$0 and not need to enter any payment information, and will enable Frontline Systems to assist students with installation, and provide technical support to you during the course. Please give the course code, plus the instructions, to your students.

If you're evaluating the book for adoption, you can use the course code yourself to download and install the software as described below.

For Students: Installing Analytic Solver Data Mining (ASDM) for Education

1. To download and install ASDM for Education from Frontline Systems, to work with Microsoft Excel for Windows, please visit <https://www.solver.com/welcome-students>
2. **Fill out the registration form on this page**, supplying your name, school, email address (key information will be sent to this address), Course Code (obtain this from your instructor), and Textbook Code (enter **SDMBI4**).
3. **Click the download button**, and save the downloaded file SolverSetup.
4. **Close any Excel windows you have open.**
5. **Run SolverSetup to install the software.**

If you have problems downloading or installing, please email support@solver.com or call **775-831-0300** and press 4 (tech support). Say that you have Analytic Solver Data Mining for Education, and have your course code and textbook code available.

If you have problems setting up or solving your model, or interpreting the results, please ask your instructor for assistance. Frontline Systems cannot help you with homework problems.

- *If you purchase this textbook but you aren't enrolled in a course, call 775-831-0300 and press 0 for assistance with the software.*
- If you have a Mac, the best option is to use "Analytic Solver Cloud" or use Excel online via an office 365 by inserting the add-in from the Microsoft Store (see <https://solver.zendesk.com/hc/en-us/articles/360024207754-Inserting-Analytic-Solver-Cloud-from-the-Microsoft-Store>).

To our families

Boaz and Noa

Liz, Lisa, and Allison

Komal and Yash

Tehmi, Arjun, and in

memory of Aneesh



Contents

Foreword	xix
Preface to the Fourth Edition	xxi
Acknowledgments	xxv

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?	9
1.7 Terminology and Notation	9
1.8 Road Maps to This Book	11
Order of Topics	12
CHAPTER 2 Overview of the Machine Learning Process	15
2.1 Introduction	15
2.2 Core Ideas in Machine Learning	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 a Machine Learning Project	19
2.4 Preliminary Steps	21
Organization of Data	21

VIII CONTENTS

	Sampling from a Database	21
	Overampling Rare Events in Classification Tasks	22
	Preprocessing and Cleaning the Data	22
2.5	Predictive Power and Overfitting	27
	Creation and Use of Data Partitions	27
	Overfitting	30
2.6	Building a Predictive Model with ASDM	32
	Predicting Home Values in the West Roxbury Neighborhood	32
	Modeling Process	34
	Machine Learning Workflow	41
2.7	Using Excel for Machine Learning	43
2.8	Automating Machine Learning Solutions	43
	Predicting Power Generator Failure	45
	Uber's Michelangelo	47
2.9	Ethical Practice in Machine Learning	49
	Machine Learning Software: The State of the Market (by Herb Edelstein)	49
	Problems	54

PART II DATA EXPLORATION AND DIMENSION REDUCTION

CHAPTER 3 Data Visualization 59

3.1	Uses of Data Visualization	59
3.2	Data Examples	61
	Example 1: Boston Housing Data	61
	Example 2: Ridership on Amtrak Trains	62
3.3	Basic Charts: Bar Charts, Line Charts, and Scatter Plots	62
	Distribution Plots	64
	Heatmaps: Visualizing Correlations and Missing Values	67
3.4	Multidimensional Visualization	68
	Adding Variables	69
	Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering	71
	Reference: Trend Line and Labels	74
	Scaling up to Large Datasets	75
	Multivariate Plot	76
	Interactive Visualization	78
3.5	Specialized Visualizations	81
	Visualizing Networked Data	81
	Visualizing Hierarchical Data: Treemaps	82
	Visualizing Geographical Data: Map Charts	84
3.6	Summary: Major Visualizations and Operations	86
	Prediction	86
	Classification	86
	Time Series Forecasting	86
	Unsupervised Learning	87
	Problems	88

CHAPTER 4 Dimension Reduction	91
4.1 Introduction	91
4.2 Curse of Dimensionality	92
4.3 Practical Considerations	92
Example 1: House Prices in Boston	93
4.4 Data Summaries	94
4.5 Correlation Analysis	96
4.6 Reducing the Number of Categories in Categorical Variables	97
4.7 Converting a Categorical Variable to a Numerical Variable	98
4.8 Principal Component Analysis	99
Example 2: Breakfast Cereals	99
Principal Components	103
Normalizing the Data	105
Using Principal Components for Classification and Prediction	107
4.9 Dimension Reduction Using Regression Models	109
4.10 Dimension Reduction Using Classification and Regression Trees	110
Problems	111

PART III PERFORMANCE EVALUATION

CHAPTER 5 Evaluating Predictive Performance	115
5.1 Introduction	115
5.2 Evaluating Predictive Performance	116
Benchmark: The Average	117
Prediction Accuracy Measures	117
5.3 Judging Classifier Performance	121
Benchmark: The Naive Rule	121
Class Separation	121
The Classification Matrix	122
Using the Validation Data	123
Accuracy Measures	123
Cutoff for Classification	124
Performance in Unequal Importance of Classes	128
Asymmetric Misclassification Costs	131
5.4 Judging Ranking Performance	134
5.5 Oversampling	139
Problems	145

PART IV PREDICTION AND CLASSIFICATION METHODS

CHAPTER 6 Multiple Linear Regression	151
6.1 Introduction	151
6.2 Explanatory vs. Predictive Modeling	152

6.3	Estimating the Regression Equation and Prediction	154
	Example: Predicting the Price of Used Toyota Corolla Cars	155
6.4	Variable Selection in Linear Regression	158
	Reducing the Number of Predictors	158
	Problems	165
CHAPTER 7 <i>k</i>-Nearest-Neighbors (<i>k</i>-NN)		169
7.1	The <i>k</i> -NN Classifier (categorical outcome)	169
	Determining Neighbors	170
	Classification Rule	170
	Example: Riding Mowers	171
	Choosing <i>k</i>	172
	Setting the Cutoff Value	173
	<i>k</i> -NN with More Than Two Classes	174
	Converting Categorical Variables to Binary Dummies	174
7.2	<i>k</i> -NN for a Numerical Response	175
7.3	Machine Learning Workflow	175
7.4	Advantages and Shortcomings of <i>k</i> -NN Algorithms	175
	Problems	179
CHAPTER 8 The Naive Bayes Classifier		181
8.1	Introduction	181
	Example 1: Predicting Fraudulent Financial Reporting	182
8.2	Applying the Full (Exact) Bayesian Classifier	183
	Using the “Assign to the Most Probable Class” Method	183
	Using the Cutoff Probability Method	184
	Practical Difficulty with the Complete (Exact) Bayes Procedure	184
8.3	Solution: Naive Bayes	185
8.4	Advantages and Shortcomings of the Naive Bayes Classifier	193
	Problems	195
CHAPTER 9 Classification and Regression Trees		197
9.1	Introduction	197
	Tree Structure	199
	Decision Rules	199
9.2	Classification Trees	200
	Example 1: Riding Mowers	200
	Measures of Impurity	203
9.3	Evaluating the Performance of a Classification Tree	206
	Example 2: Acceptance of Personal Loan	207
9.4	Avoiding Overfitting	211
	Stopping Tree Growth: CHAID	211
	Pruning the Tree	212

9.5	Classification Rules from Trees	216
9.6	Classification Trees for More Than Two Classes	217
9.7	Regression Trees	217
	Prediction	218
	Measuring Impurity	218
	Evaluating Performance	220
9.8	Advantages and Weaknesses of Single Trees	220
9.9	Improving Prediction: Random Forests and Boosted Trees	222
	Random Forests	222
	Boosted Trees	222
	Problems	225

CHAPTER 10 Logistic Regression 229

10.1	Introduction	229
10.2	The Logistic Regression Model	231
	Example: Acceptance of Personal Loan	232
	Model with a Single Predictor	233
	Estimating the Logistic Model from Data	234
	Interpreting Results in Terms of Odds	238
	Evaluating Classification Performance	239
	Variable Selection	241
10.3	Example of Complete Analysis: Predicting Delayed Flights	242
	Data Visualization	243
	Data Preprocessing	244
	Model Fitting and Estimation	245
	Model Interpretation	246
	Model Performance	246
	Variable Selection	247
10.4	Appendix: Logistic Regression for More Than Two Classes	250
	Problems	253

CHAPTER 11 Neural Nets 257

11.1	Introduction	257
11.2	Concept and Structure of a Neural Network	258
11.3	Fitting a Network to Data	259
	Example 1: Tiny Dataset	259
	Computing Output of Nodes	260
	Preprocessing the Data	263
	Training the Model	264
11.4	Required User Input for Training a Network	267
	Example 2: Classifying Accident Severity	269
11.5	Model Validation and Use	272
	Avoiding Overfitting	272
	Using the Output for Prediction and Classification	273

XII CONTENTS

11.6 Deep Learning 273
Convolutional Neural Networks (CNNs) 274
Local Feature Map 276
A Hierarchy of Features 276
The Learning Process 276
Unsupervised Learning 277
Conclusion 278
11.7 Advantages and Weaknesses of Neural Networks 279
Problems 280

CHAPTER 12 Discriminant Analysis 283

12.1 Introduction 283
Example 1: Riding Mowers 284
Example 2: Personal Loan Acceptance 284
12.2 Distance of an Observation from a Class 286
12.3 Fisher’s Linear Classification Functions 287
12.4 Classification Performance of Discriminant Analysis 291
12.5 Prior Probabilities 292
12.6 Unequal Misclassification Costs 293
12.7 Classifying More Than Two Classes 293
Example 3: Medical Dispatch to Accident Scenes 293
12.8 Advantages and Weaknesses 297
Problems 299

CHAPTER 13 Generating, Comparing, and Combining Multiple Models 303

13.1 Ensembles 304
Why Ensembles Can Improve Predictive Power 304
Simple Averaging or Voting 306
Bagging 307
Boosting 307
Bagging and Boosting in ASDM 307
Advantages and Weaknesses of Ensembles 308
13.2 Automated Machine Learning (AutoML) 309
AutoML: Explore and Clean Data 310
AutoML: Determine Machine Learning Task 310
AutoML: Choose Features and Machine Learning Methods 310
AutoML: Evaluate Model Performance 312
AutoML: Model Deployment 313
Advantages and Weaknesses of Automated Machine Learning 313
13.3 Summary 314
Problems 315

PART V INTERVENTION AND USER FEEDBACK

CHAPTER 14 Experiments, Uplift Modeling, and Reinforcement Learning 319

14.1	A/B Testing	319
	Example: Testing a New Feature in a Photo Sharing App	321
	The Statistical Test for Comparing Two Groups (t -test)	322
	Multiple Treatment Groups: A/B/ n tests	324
	Multiple A/B Tests and the Danger of Multiple Testing	324
14.2	Uplift (Persuasion) Modeling	325
	Gathering the Data	326
	A Simple Model	327
	Modeling Individual Uplift	328
	Using the Results of an Uplift Model	330
14.3	Reinforcement Learning	330
	Explore-Exploit: Multi-Armed Bandits	331
	Markov Decision Process (MDP)	333
14.4	Summary	335
	Problems	337

PART VI MINING RELATIONSHIPS AMONG RECORDS

CHAPTER 15 Association Rules and Collaborative Filtering 341

15.1	Association Rules	341
	Discovering Association Rules in Transaction Databases	342
	Example 1: Synthetic Data on Purchases of Phone Faceplates	342
	Generating Candidate Rules	343
	The Apriori Algorithm	345
	Selecting Strong Rules	345
	Data Format	347
	The Process of Rule Selection	348
	Interpreting the Results	350
	Rules and Chance	350
	Example 2: Rules for Similar Book Purchases	352
15.2	Collaborative Filtering	354
	Data Type and Format	355
	Example 3: Netflix Prize Contest	355
	User-Based Collaborative Filtering: “People Like You”	357
	Item-Based Collaborative Filtering	359
	Advantages and Weaknesses of Collaborative Filtering	360
	Collaborative Filtering vs. Association Rules	361
15.3	Summary	362
	Problems	364

CHAPTER 16 Cluster Analysis 369

- 16.1 Introduction 369
 - Example: Public Utilities 371
- 16.2 Measuring Distance Between Two Observations 373
 - Euclidean Distance 373
 - Normalizing Numerical Variables 373
 - Other Distance Measures for Numerical Data 375
 - Distance Measures for Categorical Data 376
 - Distance Measures for Mixed Data 377
- 16.3 Measuring Distance Between Two Clusters 377
- 16.4 Hierarchical (Agglomerative) Clustering 380
 - Single Linkage 380
 - Complete Linkage 381
 - Average Linkage 381
 - Centroid Linkage 382
 - Dendrograms: Displaying Clustering Process and Results 383
 - Validating Clusters 385
 - Limitations of Hierarchical Clustering 387
- 16.5 Non-hierarchical Clustering: The k -Means Algorithm 389
 - Initial Partition into k Clusters 391
- Problems 395

PART VII FORECASTING TIME SERIES

CHAPTER 17 Handling Time Series 401

- 17.1 Introduction 401
- 17.2 Descriptive vs. Predictive Modeling 403
- 17.3 Popular Forecasting Methods in Business 403
 - Combining Methods 403
- 17.4 Time Series Components 404
 - Example: Ridership on Amtrak Trains 404
- 17.5 Data Partitioning and Performance Evaluation 408
 - Benchmark Performance: Naive Forecasts 409
 - Generating Future Forecasts 410
- Problems 412

CHAPTER 18 Regression-Based Forecasting 415

- 18.1 A Model with Trend 415
 - Linear Trend 415
 - Exponential Trend 418
 - Polynomial Trend 419

18.2 A Model with Seasonality 420

18.3 A Model with Trend and Seasonality 423

18.4 Autocorrelation and ARIMA Models 425

Computing Autocorrelation 425

Improving Forecasts by Integrating Autocorrelation Information 428

Evaluating Predictability 431

Problems 434

CHAPTER 19 Smoothing Methods 445

19.1 Introduction 445

19.2 Moving Average 446

Centered Moving Average for Visualization 446

Trailing Moving Average for Forecasting 447

Choosing Window Width (w) 449

19.3 Simple Exponential Smoothing 451

Choosing Smoothing Parameter α 452

Relation Between Moving Average and Simple Exponential Smoothing 453

19.4 Advanced Exponential Smoothing 453

Series with a Trend 454

Series with a Trend and Seasonality 454

Series with Seasonality (No Trend) 455

Problems 457

PART VIII DATA ANALYTICS

CHAPTER 20 Social Network Analytics 467

20.1 Introduction 467

20.2 Directed vs. Undirected Networks 468

20.3 Visualizing and Analyzing Networks 469

Plot Layout 470

Adjacency List 472

Adjacency Matrix 472

Using Network Data in Classification and Prediction 473

20.4 Social Data Metrics and Taxonomy 473

Node-Level Centrality Metrics 474

Egocentric Network 475

Network Metrics 475

20.5 Using Network Metrics in Prediction and Classification 478

Link Prediction 478

Entity Resolution 479

Collaborative Filtering 481

20.6 Advantages and Disadvantages 484

Problems 486

CHAPTER 21 Text Mining 487

- 21.1 Introduction 487
- 21.2 The Spreadsheet Representation of Text: Term–Document Matrix and “Bag-of-Words ” 488
- 21.3 Bag-of-Words vs. Meaning Extraction at Document Level 489
- 21.4 Preprocessing the Text 490
 - Tokenization 490
 - Text Reduction 491
 - Presence/Absence vs. Frequency 494
 - Term Frequency - Inverse Document Frequency (TF-IDF) 494
 - From Terms to Concepts: Latent Semantic Indexing 495
 - Extracting Meaning 497
 - From Terms to High Dimensional Word Vectors: Word2Vec 497
- 21.5 Implementing Machine Learning Methods 497
- 21.6 Example: Online Discussions on Autos and Electronics 498
 - Importing and Labeling the Records 498
 - Tokenization 499
 - Text Processing and Reduction 499
 - Producing a Concept Matrix 500
 - Labeling the Documents 500
 - Fitting a Model 501
 - Prediction 502
- 21.7 Summary 502
- Problems 504

CHAPTER 22 Responsible Data Science 507

- 22.1 Introduction 507
- 22.2 Unintentional Harm 508
- 22.3 Legal Considerations 509
- 22.4 Principles of Responsible Data Science 511
 - Non-maleficence 511
 - Fairness 512
 - Transparency 513
 - Accountability 514
 - Data Privacy and Security 514
- 22.5 A Responsible Data Science Framework 514
 - Justification 514
 - Assembly 515
 - Data Preparation 516
 - Modeling 517
 - Auditing 517
- 22.6 Documentation Tools 518
 - Impact Statements 518
 - Model Cards 519
 - Datasheets 520

Audit Reports 520

22.7 Example: Applying the RDS Framework to the COMPAS Example 522

 Unanticipated Uses 522

 Ethical Concerns 522

 Protected Groups 522

 Data Issues 523

 Fitting the Model 523

 Auditing the Model 524

 Bias Mitigation 530

22.8 Summary 531

Problems 532

PART IX CASES

CHAPTER 23 Cases 537

23.1 Charles Book Club 537

 The Book Industry 537

 Database Marketing at Charles 538

 Machine Learning Techniques 540

 Assignment 544

23.2 German Credit 546

 Background 546

 Data 546

 Assignment 549

23.3 Tayko Software Cataloger 551

 Background 551

 The Mailing Experiment 551

 Data 551

 Assignment 553

23.4 Political Persuasion 555

 Background 555

 Predictive Analytics Arrives in US Politics 555

 Political Targeting 555

 Uplift 556

 Data 557

 Assignment 557

23.5 Taxi Cancellations 559

 Business Situation 559

 Assignment 559

23.6 Segmenting Consumers of Bath Soap 561

 Business Situation 561

 Key Problems 561

 Data 562

 Measuring Brand Loyalty 562

 Assignment 562

XVIII CONTENTS

23.7	Direct-Mail Fundraising	565
	Background	565
	Data	565
	Assignment	565
23.8	Catalog Cross-Selling	568
	Background	568
	Assignment	568
23.9	Time Series Case: Forecasting Public Transportation Demand	570
	Background	570
	Problem Description	570
	Available Data	570
	Assignment Goal	570
	Assignment	571
	Tips and Suggested Steps	571
23.10	Loan Approval	572
	Background	572
	Regulatory Requirements	572
	Getting Started	572
	Assignment	573
	References	575
	Data Files Used in the Book	577
	Index	579



Foreword

I was tasked to develop a course on statistical and machine learning for our new business analytics program at a major public university almost a decade ago. I quickly discovered that, unlike business statistics and management science, a legacy of textbooks on this subject appropriate for students without an extensive mathematical background did not exist. Fortunately, a colleague at another institution pointed me toward *Data Mining for Business Intelligence*, the 2nd edition of this book (now called *Machine Learning for Business Analytics*), and it has been the core text for my predictive analytics course ever since. My initial choice was validated, and the book is still the best choice for our students.

Universities are now offering a wide range of degrees, concentrations, and certificates in business analytics. Success with analytics is grounded in employees with the requisite skills to meet industry needs. The market for business analytics continues to expand rapidly as companies adopt new technologies to manage and understand business processes using enterprise, e-commerce, and sensor data. Numerous media and vendor reports have documented the contribution of predictive analytics to improved business outcomes. More recently, analytical solutions have been extended to incorporate artificial intelligence using advanced machine learning techniques, thereby reducing labor costs, enhancing customer experiences, and improving cybersecurity. Negotiating such a new constellation of developing frontiers requires good integrated learning resources. This book is certainly one of them.

The genesis of this book, now in its 4th edition with a revised title, was to make machine learning accessible to students outside of traditional STEM disciplines. It uses the Analytic Solver® add-in for Excel from Frontline Systems, Inc., to quickly introduce students to the machine learning process in a spreadsheet environment without the need for writing computer code. With the Data Mining module of Analytic Solver, students can gain hands-on skills with a broad range of data visualization, data handling utilities, and statistical and machine learning techniques, augmented by tools for data partitioning, variable transformation, feature selection, and model performance evaluation. As a result, both instructors and students are free to focus on conceptual learning using data

sets that illustrate a variety of use cases for machine learning in the major functional areas of business.

While this book is also available in other editions based on alternative software platforms, I can speak firsthand as to how its content when used in conjunction with Analytic Solver facilitates a concrete understanding of how statistical and machine learning algorithms work. Students can interactively experiment with the effects of changing algorithmic settings and explore the detailed reports and charts generated by Analytic Solver to review intermediate calculations. By the end of each chapter, students can describe the primary mechanism being used by an algorithm for classification or prediction, articulate its requirements in terms of input variables and parameters, interpret its results, and identify its strengths and weaknesses. They are also exposed to carefully crafted business scenarios and associated real data for each topic. With this foundation, they will be well-prepared for a data analyst role in their respective domains, as well as further study of advanced analytics.

The authors of this book bring together a combination of extensive academic and industry experience that makes a unique contribution to the textbook landscape. In this new edition, they have updated the text and figures to correspond to the latest version of Analytic Solver, expanded coverage of metrics for assessing model performance, and added contemporary topics including deep learning and ethical considerations in the use of data science.

If you are a newcomer to teaching in this area of predictive analytics, *Machine Learning for Business Analytics* will support you and your students while you gain traction in the classroom. With more background, it will serve as a useful resource to guide your own discovery and acquisition of new knowledge and skills for machine learning.

SUSAN W. PALOCSAY
James Madison University 2023



Preface to the Fourth 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 (ASDM, previously XLMiner), was followed by two more Analytic Solver editions, a JMP edition, two R editions, a Python edition, a RapidMiner edition, and now this 4th edition with Analytic Solver, with its companion website, www.dataminingbook.com.

As in the previous Analytic Solver editions, the focus in this new edition is on machine learning concepts and how to implement the associated algorithms in Analytic Solver Data Mining.

For this new ASDM edition, a new co-author, Kuber Deokar, comes on board bringing extensive experience in online course design, development, and delivery, including using ASDM. He has taught courses using various editions of this book for over a decade and has helped shape the student experience in many ways.

The new edition provides significant updates both in terms of ASDM and in terms of new topics and content. In addition to updating software routines and outputs that have changed or become available since the 3rd edition, 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 several new topics:

- 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.
- A dedicated section on *deep learning* in Chapter 11.
- 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 and AutoML.

- 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 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 “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 and AutoML from Part IV, as well as Part V, which covers experiments, uplift, 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, KUBER R. DEOKAR, AND NITIN R. PATEL
2023



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.

At Statistics.com, Valerie Troiano has shepherded many instructors and students through the courses that have helped nurture the development of these books, and Janet Dobbins has helped bring them to the wider machine learning community. We also thank the many students who have used and commented on this text at Statistics.com. We thank assistant teachers (from UpThink Edutech Services Pvt Ltd) Anuja Kulkarni and Poonam Tribhuwan, and especially Shweta Jadhav for her hours of double checking the software (ASDM).

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 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 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 Lemore, and Kavya Ramu greatly assisted us in pushing ahead and finalizing this 4th edition. We are also especially grateful to Amy Hendrickson, who assisted with typesetting and making this book beautiful.

