

MACHINE LEARNING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES AND APPLICATIONS WITH JMP PRO®

SECOND EDITION

GALIT SHMUELI • PETER C. BRUCE • MIA L. STEPHENS
MURALIDHARA ANANDAMURTHY • NITIN R. PATEL



WILEY

MACHINE LEARNING FOR BUSINESS ANALYTICS

MACHINE LEARNING FOR BUSINESS ANALYTICS

Concepts, Techniques, and Applications with JMP Pro®

Second Edition

GALIT SHMUELI
National Tsing Hua University
Taipei, Taiwan

PETER C. BRUCE
statistics.com
Arlington, USA

MIA L. STEPHENS
JMP Statistical Discovery LLC
Cary, USA

MURALIDHARA ANANDAMURTHY
SAS Institute Inc
Mumbai, India

NITIN R. PATEL
Cytel, Inc.
Cambridge, USA

WILEY

Copyright 2023 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

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: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

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

Hardback: 9781119903833

Cover Design: Wiley

Cover Image: © AdobeLibrary/Adobe Stock Photos

Set in 10/12pt TimesLTStd by Straive, Chennai, India

To our families

Boaz and Noa

Liz, Lisa, and Allison

Michael, Jade Ann, and Audrey L

Seetha and Ananda

Tehmi, Arjun, and in memory of Aneesh

CONTENTS

Foreword	xix
Preface	xx
Acknowledgments	xxiii

PART I PRELIMINARIES

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
Statistical Modeling vs. Machine Learning	6
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	8
1.8 Road Maps to This Book	10
Order of Topics	12
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	19
2.3	The Steps in A Machine Learning Project	21
2.4	Preliminary Steps	22
	Organization of Data	22
	Sampling from a Database	22
	Oversampling Rare Events in Classification Tasks	23
	Preprocessing and Cleaning the Data	23
2.5	Predictive Power and Overfitting	29
	Overfitting	29
	Creation and Use of Data Partitions	31
2.6	Building a Predictive Model with JMP Pro	34
	Predicting Home Values in a Boston Neighborhood	34
	Modeling Process	36
2.7	Using JMP Pro for Machine Learning	42
2.8	Automating Machine Learning Solutions	43
	Predicting Power Generator Failure	44
	Uber's Michelangelo	45
2.9	Ethical Practice in Machine Learning	47
	Machine Learning Software: The State of the Market by Herb Edelstein	47
	Problems	52

PART II DATA EXPLORATION AND DIMENSION REDUCTION

3 Data Visualization

59

3.1	Introduction	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 Graphs, and Scatter Plots	62
	Distribution Plots: Boxplots and Histograms	64
	Heatmaps	67
3.4	Multidimensional Visualization	70
	Adding Variables: Color, Hue, Size, Shape, Multiple Panels, Animation	70
	Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering	73
	Reference: Trend Line and Labels	77
	Scaling Up: Large Datasets	79
	Multivariate Plot: Parallel Coordinates Plot	80
	Interactive Visualization	80

3.5	Specialized Visualizations 82	
	Visualizing Networked Data 82	
	Visualizing Hierarchical Data: More on Treemaps 83	
	Visualizing Geographical Data: Maps 84	
3.6	Summary: Major Visualizations and Operations, According to Machine Learning Goal 87	
	Prediction 87	
	Classification 87	
	Time Series Forecasting 87	
	Unsupervised Learning 88	
	Problems 89	
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 92	
4.4	Data Summaries 93	
	Summary Statistics 94	
	Tabulating Data 96	
4.5	Correlation Analysis 97	
4.6	Reducing the Number of Categories in Categorical Variables 98	
4.7	Converting a Categorical Variable to a Continuous Variable 100	
4.8	Principal Component Analysis 100	
	Example 2: Breakfast Cereals 101	
	Principal Components 106	
	Standardizing the Data 107	
	Using Principal Components for Classification and Prediction 110	
4.9	Dimension Reduction Using Regression Models 110	
4.10	Dimension Reduction Using Classification and Regression Trees 111	
	Problems 112	
 PART III PERFORMANCE EVALUATION		
5	Evaluating Predictive Performance	117
5.1	Introduction 118	
5.2	Evaluating Predictive Performance 118	
	Naive Benchmark: The Average 118	
	Prediction Accuracy Measures 119	
	Comparing Training and Validation Performance 120	
5.3	Judging Classifier Performance 121	
	Benchmark: The Naive Rule 121	
	Class Separation 121	
	The Classification (Confusion) Matrix 122	
	Using the Validation Data 123	
	Accuracy Measures 123	

- Propensities and Threshold for Classification 124
- Performance in Unequal Importance of Classes 127
- Asymmetric Misclassification Costs 130
- Generalization to More Than Two Classes 132
- 5.4 Judging Ranking Performance 133
 - Lift Curves for Binary Data 133
 - Beyond Two Classes 135
 - Lift Curves Incorporating Costs and Benefits 136
- 5.5 Oversampling 137
 - Creating an Over-sampled Training Set 139
 - Evaluating Model Performance Using a Nonoversampled Validation Set 139
 - Evaluating Model Performance If Only Oversampled Validation Set Exists 140
 - Problems 142

PART IV PREDICTION AND CLASSIFICATION METHODS

6 Multiple Linear Regression 147

- 6.1 Introduction 147
- 6.2 Explanatory vs. Predictive Modeling 148
- 6.3 Estimating the Regression Equation and Prediction 149
 - Example: Predicting the Price of Used Toyota Corolla Automobiles 150
- 6.4 Variable Selection in Linear Regression 155
 - Reducing the Number of Predictors 155
 - How to Reduce the Number of Predictors 156
 - Manual Variable Selection 156
 - Automated Variable Selection 157
 - Regularization (Shrinkage Models) 164
 - Problems 170

7 *k*-Nearest Neighbors (*k*-NN) 175

- 7.1 The *k*-NN Classifier (Categorical Outcome) 175
 - Determining Neighbors 175
 - Classification Rule 176
 - Example: Riding Mowers 176
 - Choosing Parameter *k* 178
 - Setting the Threshold Value 179
 - Weighted *k*-NN 181
 - k*-NN with More Than Two Classes 182
 - Working with Categorical Predictors 182
- 7.2 *k*-NN for a Numerical Response 184
- 7.3 Advantages and Shortcomings of *k*-NN Algorithms 184
- Problems 186

8	The Naive Bayes Classifier	189
8.1	Introduction 189	
	Threshold Probability Method 190	
	Conditional Probability 190	
	Example 1: Predicting Fraudulent Financial Reporting 190	
8.2	Applying the Full (Exact) Bayesian Classifier 191	
	Using the “Assign to the Most Probable Class” Method 191	
	Using the Threshold Probability Method 191	
	Practical Difficulty with the Complete (Exact) Bayes Procedure 192	
8.3	Solution: Naive Bayes 192	
	The Naive Bayes Assumption of Conditional Independence 193	
	Using the Threshold Probability Method 194	
	Example 2: Predicting Fraudulent Financial Reports 194	
	Example 3: Predicting Delayed Flights 195	
	Evaluating the Performance of Naive Bayes Output from JMP 198	
	Working with Continuous Predictors 199	
8.4	Advantages and Shortcomings of the Naive Bayes Classifier 201	
	Problems 203	
9	Classification and Regression Trees	205
9.1	Introduction 206	
	Tree Structure 206	
	Decision Rules 207	
	Classifying a New Record 207	
9.2	Classification Trees 207	
	Recursive Partitioning 207	
	Example 1: Riding Mowers 208	
	Categorical Predictors 210	
	Standardization 210	
9.3	Growing a Tree for Riding Mowers Example 210	
	Choice of First Split 211	
	Choice of Second Split 212	
	Final Tree 212	
	Using a Tree to Classify New Records 213	
9.4	Evaluating the Performance of a Classification Tree 215	
	Example 2: Acceptance of Personal Loan 215	
9.5	Avoiding Overfitting 219	
	Stopping Tree Growth: CHAID 220	
	Growing a Full Tree and Pruning It Back 220	
	How JMP Pro Limits Tree Size 221	
9.6	Classification Rules from Trees 222	
9.7	Classification Trees for More Than Two Classes 224	
9.8	Regression Trees 224	
	Prediction 224	
	Evaluating Performance 225	
9.9	Advantages and Weaknesses of a Single Tree 227	

9.10 Improving Prediction: Random Forests and Boosted Trees 229
 Random Forests 229
 Boosted Trees 230
 Problems 233

10 Logistic Regression 237

10.1 Introduction 237
 10.2 The Logistic Regression Model 239
 10.3 Example: Acceptance of Personal Loan 240
 Model with a Single Predictor 241
 Estimating the Logistic Model from Data: Multiple Predictors 243
 Interpreting Results in Terms of Odds (for a Profiling Goal) 246
 10.4 Evaluating Classification Performance 247
 10.5 Variable Selection 249
 10.6 Logistic Regression for Multi-class Classification 250
 Logistic Regression for Nominal Classes 250
 Logistic Regression for Ordinal Classes 251
 Example: Accident Data 252
 10.7 Example of Complete Analysis: Predicting Delayed Flights 253
 Data Preprocessing 255
 Model Fitting, Estimation, and Interpretation—A Simple Model 256
 Model Fitting, Estimation and Interpretation—The Full Model 257
 Model Performance 257
 Problems 264

11 Neural Nets 267

11.1 Introduction 267
 11.2 Concept and Structure of a Neural Network 268
 11.3 Fitting a Network to Data 269
 Example 1: Tiny Dataset 269
 Computing Output of Nodes 269
 Preprocessing the Data 272
 Training the Model 273
 Using the Output for Prediction and Classification 279
 Example 2: Classifying Accident Severity 279
 Avoiding Overfitting 281
 11.4 User Input in *JMP Pro* 282
 11.5 Exploring the Relationship Between Predictors and Outcome 284
 11.6 Deep Learning 285
 Convolutional Neural Networks (CNNs) 285
 Local Feature Map 287
 A Hierarchy of Features 287
 The Learning Process 287
 Unsupervised Learning 288
 Conclusion 289
 11.7 Advantages and Weaknesses of Neural Networks 289
 Problems 290

12 Discriminant Analysis 293

- 12.1 Introduction 293
 - Example 1: Riding Mowers 294
 - Example 2: Personal Loan Acceptance 294
- 12.2 Distance of an Observation from a Class 295
- 12.3 From Distances to Propensities and Classifications 297
- 12.4 Classification Performance of Discriminant Analysis 300
- 12.5 Prior Probabilities 301
- 12.6 Classifying More Than Two Classes 303
 - Example 3: Medical Dispatch to Accident Scenes 303
- 12.7 Advantages and Weaknesses 306
 - Problems 307

13 Generating, Comparing, and Combining Multiple Models 311

- 13.1 Ensembles 311
 - Why Ensembles Can Improve Predictive Power 312
 - Simple Averaging or Voting 313
 - Bagging 314
 - Boosting 315
 - Stacking 316
 - Advantages and Weaknesses of Ensembles 317
- 13.2 Automated Machine Learning (AutoML) 317
 - AutoML: Explore and Clean Data 317
 - AutoML: Determine Machine Learning Task 318
 - AutoML: Choose Features and Machine Learning Methods 318
 - AutoML: Evaluate Model Performance 320
 - AutoML: Model Deployment 321
 - Advantages and Weaknesses of Automated Machine Learning 322
- 13.3 Summary 322
 - Problems 323

PART V INTERVENTION AND USER FEEDBACK**14 Interventions: Experiments, Uplift Models, and Reinforcement Learning 327**

- 14.1 Introduction 327
- 14.2 A/B Testing 328
 - Example: Testing a New Feature in a Photo Sharing App 329
 - The Statistical Test for Comparing Two Groups (*T*-Test) 329
 - Multiple Treatment Groups: *A/B/n* Tests 333
 - Multiple A/B Tests and the Danger of Multiple Testing 333
- 14.3 Uplift (Persuasion) Modeling 333
 - Getting the Data 334
 - A Simple Model 336
 - Modeling Individual Uplift 336
 - Creating Uplift Models in **JMP Pro** 337
 - Using the Results of an Uplift Model 338

- 14.4 Reinforcement Learning 340
 - Explore-Exploit: Multi-armed Bandits 340
 - Markov Decision Process (MDP) 341
- 14.5 Summary 344
- Problems 345

PART VI MINING RELATIONSHIPS AMONG RECORDS

15 Association Rules and Collaborative Filtering 349

- 15.1 Association Rules 349
 - Discovering Association Rules in Transaction Databases 350
 - Example 1: Synthetic Data on Purchases of Phone Faceplates 350
 - Data Format 350
 - Generating Candidate Rules 352
 - The Apriori Algorithm 353
 - Selecting Strong Rules 353
 - The Process of Rule Selection 356
 - Interpreting the Results 358
 - Rules and Chance 359
 - Example 2: Rules for Similar Book Purchases 361
- 15.2 Collaborative Filtering 362
 - Data Type and Format 363
 - Example 3: Netflix Prize Contest 363
 - User-Based Collaborative Filtering: “People Like You” 365
 - Item-Based Collaborative Filtering 366
 - Evaluating Performance 367
 - Advantages and Weaknesses of Collaborative Filtering 368
 - Collaborative Filtering vs. Association Rules 369
- 15.3 Summary 370
- Problems 372

16 Cluster Analysis 375

- 16.1 Introduction 375
 - Example: Public Utilities 377
- 16.2 Measuring Distance Between Two Records 378
 - Euclidean Distance 379
 - Standardizing Numerical Measurements 379
 - Other Distance Measures for Numerical Data 379
 - Distance Measures for Categorical Data 382
 - Distance Measures for Mixed Data 382
- 16.3 Measuring Distance Between Two Clusters 383
 - Minimum Distance 383
 - Maximum Distance 383

- Average Distance 383
- Centroid Distance 383
- 16.4 Hierarchical (Agglomerative) Clustering 385
 - Single Linkage 385
 - Complete Linkage 386
 - Average Linkage 386
 - Centroid Linkage 386
 - Ward's Method 387
 - Dendrograms: Displaying Clustering Process and Results 387
 - Validating Clusters 391
 - Two-Way Clustering 393
 - Limitations of Hierarchical Clustering 393
- 16.5 Nonhierarchical Clustering: The K -Means Algorithm 394
 - Choosing the Number of Clusters (k) 396
- Problems 403

PART VII FORECASTING TIME SERIES

17 Handling Time Series 409

- 17.1 Introduction 409
- 17.2 Descriptive vs. Predictive Modeling 410
- 17.3 Popular Forecasting Methods in Business 411
 - Combining Methods 411
- 17.4 Time Series Components 411
 - Example: Ridership on Amtrak Trains 412
- 17.5 Data Partitioning and Performance Evaluation 415
 - Benchmark Performance: Naive Forecasts 417
 - Generating Future Forecasts 417
- Problems 419

18 Regression-Based Forecasting 423

- 18.1 A Model with Trend 424
 - Linear Trend 424
 - Exponential Trend 427
 - Polynomial Trend 429
- 18.2 A Model with Seasonality 430
 - Additive vs. Multiplicative Seasonality 432
- 18.3 A Model with Trend and Seasonality 433
- 18.4 Autocorrelation and ARIMA Models 433
 - Computing Autocorrelation 433
 - Improving Forecasts by Integrating Autocorrelation Information 437
 - Fitting AR Models to Residuals 439
 - Evaluating Predictability 441
- Problems 444

19 Smoothing and Deep Learning Methods for Forecasting 455

- 19.1 Introduction 455
- 19.2 Moving Average 456
 - Centered Moving Average for Visualization 456
 - Trailing Moving Average for Forecasting 457
 - Choosing Window Width (w) 460
- 19.3 Simple Exponential Smoothing 461
 - Choosing Smoothing Parameter α 462
 - Relation Between Moving Average and Simple Exponential Smoothing 465
- 19.4 Advanced Exponential Smoothing 465
 - Series With a Trend 465
 - Series With a Trend and Seasonality 466
- 19.5 Deep Learning for Forecasting 470
 - Problems 472

PART VIII DATA ANALYTICS**20 Text Mining 483**

- 20.1 Introduction 483
- 20.2 The Tabular Representation of Text: Document–Term Matrix and “Bag-of-Words” 484
- 20.3 Bag-of-Words vs. Meaning Extraction at Document Level 486
- 20.4 Preprocessing the Text 486
 - Tokenization 487
 - Text Reduction 488
 - Presence/Absence vs. Frequency (Occurrences) 489
 - Term Frequency-Inverse Document Frequency (TF-IDF) 489
 - From Terms to Topics: Latent Semantic Analysis and Topic Analysis 490
 - Extracting Meaning 491
 - From Terms to High Dimensional Word Vectors: Word2Vec 491
- 20.5 Implementing Machine Learning Methods 492
- 20.6 Example: Online Discussions on Autos and Electronics 492
 - Importing the Records 493
 - Text Preprocessing in JMP 494
 - Using Latent Semantic Analysis and Topic Analysis 496
 - Fitting a Predictive Model 499
 - Prediction 499
- 20.7 Example: Sentiment Analysis of Movie Reviews 500
 - Data Preparation 500
 - Latent Semantic Analysis and Fitting a Predictive Model 500
- 20.8 Summary 502
 - Problems 503

21 Responsible Data Science 505

- 21.1 Introduction 505
 - Example: Predicting Recidivism 506
- 21.2 Unintentional Harm 506
- 21.3 Legal Considerations 508
 - The General Data Protection Regulation (GDPR) 508
 - Protected Groups 508
- 21.4 Principles of Responsible Data Science 508
 - Non-maleficence 509
 - Fairness 509
 - Transparency 510
 - Accountability 511
 - Data Privacy and Security 511
- 21.5 A Responsible Data Science Framework 511
 - Justification 511
 - Assembly 512
 - Data Preparation 513
 - Modeling 513
 - Auditing 513
- 21.6 Documentation Tools 514
 - Impact Statements 514
 - Model Cards 515
 - Datasheets 516
 - Audit Reports 516
- 21.7 Example: Applying the RDS Framework to the COMPAS Example 517
 - Unanticipated Uses 518
 - Ethical Concerns 518
 - Protected Groups 518
 - Data Issues 518
 - Fitting the Model 519
 - Auditing the Model 520
 - Bias Mitigation 526
- 21.8 Summary 526
 - Problems 528

PART IX CASES**22 Cases 533**

- 22.1 Charles Book Club 533
 - The Book Industry 533
 - Database Marketing at Charles 534
 - Machine Learning Techniques 535
 - Assignment 537
- 22.2 German Credit 541
 - Background 541
 - Data 541
 - Assignment 544

22.3	Tayko Software Cataloger 545	
	Background 545	
	The Mailing Experiment 545	
	Data 545	
	Assignment 546	
22.4	Political Persuasion 548	
	Background 548	
	Predictive Analytics Arrives in US Politics 548	
	Political Targeting 548	
	Uplift 549	
	Data 549	
	Assignment 550	
22.5	Taxi Cancellations 552	
	Business Situation 552	
	Assignment 552	
22.6	Segmenting Consumers of Bath Soap 554	
	Business Situation 554	
	Key Problems 554	
	Data 555	
	Measuring Brand Loyalty 556	
	Assignment 556	
22.7	Catalog Cross-Selling 557	
	Background 557	
	Assignment 557	
22.8	Direct-Mail Fundraising 559	
	Background 559	
	Data 559	
	Assignment 559	
22.9	Time Series Case: Forecasting Public Transportation Demand 562	
	Background 562	
	Problem Description 562	
	Available Data 562	
	Assignment Goal 562	
	Assignment 563	
	Tips and Suggested Steps 563	
22.10	Loan Approval 564	
	Background 564	
	Regulatory Requirements 564	
	Getting Started 564	
	Assignment 564	
	References	567
	Data Files Used in the Book	571
	Index	573

FOREWORD

When I began my career back in the last century, most corporate computing took place on mainframe computers, data was scarce, organizations were far more hierarchical, and managerial decision-making was often driven by the loudest person in the room or the “golden gut” of an experienced executive. By contrast, today’s business world features a wide variety of digitally connected professionals who interact with their customers and their colleagues through software applications (many of them web- and cloud-based), remarkably powerful personal computers, and always-connected smart phones. Data is everywhere, though useful data is often still elusive. And more and more of the systems that companies and individuals rely upon are utilizing techniques from machine learning to deliver data-driven insights, make predictions, and drive decision making.

For the past decade, I have been teaching courses in machine learning and predictive analytics to business students at the University of San Francisco. My students have a wide variety of academic backgrounds and professional interests. My goal is to prepare them for careers in this rapidly evolving, digitally enabled, and increasingly data- and algorithmically-driven business world.

I was fortunate enough to find *Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro* several years ago. This book provides a clear roadmap to the fundamentals of machine learning as well as a number of pathways to explore a variety of specific machine learning methods for business analytics including prediction, classification, and clustering. In addition, this textbook and the JMP Pro software combine to provide a great platform for interactive learning. The textbook utilizes the JMP Pro software to illustrate machine learning fundamentals, exploratory data analysis methods, and data visualization concepts, and a broad range of supervised and unsupervised machine learning methods. The book also provides exercises that also enable you to utilize JMP Pro to learn and master machine learning techniques.

I was very excited when I learned that the next edition of this book was ready to be released. Now entitled *Machine Learning for Business Analytics: Concepts, Techniques, and Applications with JMP Pro*, this 2nd edition is based on the most recent version of the JMP Pro software, and both the text and the software have been significantly expanded and updated. This new edition includes all the first edition material (supervised learning, unsupervised methods, visualization, and time series), as well as a number of new topics: recommendation systems, text mining, ethical issues in data science, deep learning, and interventions and reinforcement learning.

Along with the JMP Pro software, this book will provide you with a foundation of knowledge about machine learning. Its lessons and insights will serve you well in today’s dynamic and data-intensive business world. Welcome aboard!

VIJAY MEHROTRA
University of San Francisco

PREFACE

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 Pro[®] edition, two R editions, a Python edition, a RapidMiner edition, and now this second JMP Pro edition, with its companion website, www.jmp.com/dataminingbook. JMP Pro is a desktop statistical package from JMP Statistical Discovery that runs natively on Mac and Windows machines.¹

The first JMP Pro edition was the first edition to fully integrate JMP Pro. As in the previous JMP edition, the focus in this new edition is on machine learning concepts and how to implement the associated algorithms in JMP Pro. All examples, special topics boxes, instructions, and exercises presented in this book are based on **JMP Pro 17**, the professional version of JMP, which has a rich array of built-in tools for interactive data visualization, analysis, and modeling.²

For this new JMP Pro edition, a new co-author, Muralidhara Anandamurthy, comes on board bringing extensive experience in analytics and data science at Genpact, Target, and Danske, and as a member of the JMP Academic Team.

The new edition provides significant updates both in terms of JMP Pro and in terms of new topics and content. In addition to updating software routines and outputs that have changed or become available since the first 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. Importantly, this edition includes several new topics:

- A new chapter on *Responsible Data Science* (Chapter 21) 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.
- A new chapter on recommendations, covering association rules and collaborative filtering (Chapter 15).
- A new chapter on Text Mining covering main approaches to the analysis of text data (Chapter 20).
- The *Performance Evaluation* exposition in Chapter 5 was expanded to include further metrics (precision and recall, F1).

¹JMP Statistical Discovery LLC, 100 SAS Campus Drive Cary, NC 27513.

²See <https://www.jmp.com/pro>

- 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 first JMP Pro 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 “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 VII might be considered, and we recommend introducing the new 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 the new part on “Data Analytics.”

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

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, and reinforcement learning. Some instructors choose to use the cases (Chapter 22) 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 BRUCE, MIA STEPHENS, MURALIDHARA ANANDAMURTHY, AND NITIN 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, RapidMiner, and JMP.

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.

Kuber Deokar, instructional operations supervisor at Statistics.com, has been unstinting in his assistance, support, and detailed attention. We also thank Anuja Kulkarni, Poonam Tribhuvan, 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.

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 Margret Bjarnadottir have provided suggestions and feedback on numerous occasions. 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 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.

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, helped us navigate through the publishing process with great speed, and together with Curt Hinrichs's encouragement and support helped make this JMP Pro[®] edition possible. Jon Gurstelle, Kathleen Pagliaro, Allison McGinniss, Sari Friedman, and Katrina Maceda at Wiley, and Shikha Pahuja from Thomson Digital, were all helpful and responsive as we finalized the first JMP Pro edition. 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 new JMP Pro edition. We are also especially grateful to Amy Hendrickson, who assisted with typesetting and making this book beautiful.

Finally, we'd like to thank the reviewers of the first JMP Pro edition for their feedback and suggestions, and members of the JMP Documentation, Education and Development teams, for their support, patience, and responsiveness to our endless questions and requests. We thank L. Allison Jones-Farmer, Maria Weese, Ian Cox, Di Michelson, Marie Gaudard, Curt Hinrichs, Rob Carver, Jim Grayson, Brady Brady, Jian Cao, Elizabeth Claassen, Peng Liu, Chris Gotwalt, Russ Wolfinger, and Fang Chen. Most important, we thank John Sall, whose innovation, inspiration, and continued dedication to providing accessible and user-friendly desktop statistical software made JMP, and this book, possible.

PART I

PRELIMINARIES

1

INTRODUCTION

1.1 WHAT IS BUSINESS ANALYTICS?

Business analytics (BA) is the practice and art of bringing quantitative data to bear on decision-making. The term means different things to different organizations.

Consider the role of analytics in helping newspapers survive the transition to a digital world. One tabloid newspaper with a working-class readership in Britain had launched a web version of the paper, and did tests on its home page to determine which images produced more hits: cats, dogs, or monkeys. This simple application, for this company, was considered analytics. By contrast, the *Washington Post* has a highly influential audience that is of interest to big defense contractors: it is perhaps the only newspaper where you routinely see advertisements for aircraft carriers. In the digital environment, the *Post* can track readers by time of day, location, and user subscription information. In this fashion the display of the aircraft carrier advertisement in the online paper may be focused on a very small group of individuals—say, the members of the House and Senate Armed Services Committees who will be voting on the Pentagon’s budget.

Business analytics, or more generically, *analytics*, includes a range of data analysis methods.

Many powerful applications involve little more than counting, rule checking, and basic arithmetic. For some organizations, this is what is meant by analytics.

The next level of business analytics, now termed *business intelligence* (BI), refers to the use of data visualization and reporting for becoming aware and understanding “what happened and what is happening.” This is done by use of charts, tables, and dashboards to display, examine, and explore data. Business intelligence, which earlier consisted mainly of generating static reports, has evolved into more user-friendly and effective tools and practices, such as creating interactive dashboards that allow the user not only to access real-time data, but also to directly interact with it. Effective dashboards are those that tie directly to company data, and give managers a tool to see quickly what might not readily be apparent in a large complex database. One such tool for industrial operations managers

displays customer orders in one two-dimensional display using color and bubble size as added variables. The resulting 2 by 2 matrix shows customer name, type of product, size of order, and length of time to produce.

Business analytics now typically includes BI as well as sophisticated data analysis methods, such as statistical models and machine learning algorithms used for exploring data, quantifying and explaining relationships between measurements, and predicting new records. Methods like regression models are used to describe and quantify “on average” relationships (e.g., between advertising and sales), to predict new records (e.g., whether a new patient will react positively to a medication), and to forecast future values (e.g., next week’s web traffic).

Readers familiar with the earlier edition of this book might have noticed that the book title changed from *Data Mining for Business Analytics* to *Machine Learning for Business Analytics*. The change reflects the more recent term BA, which overtook the earlier term BI to denote advanced analytics. Today, BI is used to refer to data visualization and reporting. The change from *data mining* to *machine learning* reflects today’s common use of *machine learning* to refer to algorithms that learn from data. This book uses primarily the term *machine learning*.

WHO USES PREDICTIVE ANALYTICS?

The widespread adoption of predictive analytics, coupled with the accelerating availability of data, has increased organizations’ capabilities throughout the economy. A few examples:

Credit scoring: One long-established use of predictive modeling techniques for business prediction is credit scoring. A credit score is not some arbitrary judgement of creditworthiness; it is based mainly on a predictive model that uses prior data to predict repayment behavior.

Future purchases: A more recent (and controversial) example is Target’s use of predictive modeling to classify sales prospects as “pregnant” or “not-pregnant.” Those classified as pregnant could then be sent sales promotions at an early stage of pregnancy, giving Target a head start on a significant purchase stream.

Tax evasion: The US Internal Revenue Service found it was 25 times more likely to find tax evasion when enforcement activity was based on predictive models, allowing agents to focus on the most likely tax cheats (Siegel, 2013).

The business analytics toolkit also includes statistical experiments, the most common of which is known to marketers as A/B testing. These are often used for pricing decisions:

- Orbitz, the travel site, has found that it could price hotel options higher for Mac users than Windows users.
- Staples online store found that it could charge more for staplers if a customer lived far from a Staples store.