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

with website

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

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

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

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).

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

[2] See `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 Tribhuwan, 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 Birckhead for invaluable datasets. We also thank MBA students Rob Whitener and Daniel Curtis for the heatmap and map charts.

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

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

Machine Learning for Business Analytics: Concepts, Techniques, and Applications with JMP Pro®,
Second Edition. Galit Shmueli, Peter C. Bruce, Mia L. Stephens, Muralidhara Anandamurthy, and Nitin R. Patel.

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.