DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES, AND APPLICATIONS WITH JMP PRO®

Galit Shmueli • Peter C. Bruce • Mia L. Stephens • Nitin R. Patel

with website

WILEY

DATA MINING FOR BUSINESS ANALYTICS

DATA MINING FOR BUSINESS ANALYTICS

Concepts, Techniques, and Applications with JMP Pro®

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

WILEY

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

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:

Names: Shmueli, Galit, 1971- author. | Bruce, Peter C., 1953- author. |
Stephens, Mia L., author. | Patel, Nitin R. (Nitin Ratilal), author.
Title: Data mining for business analytics : concepts, techniques, and
applications with JMP Pro / Galit Shmueli, Peter C. Bruce, Mia L. Stephens,
Nitin R. Patel.
Description: Hoboken, New Jersey : John Wiley & Sons, 2017. | Includes index.
Identifiers: LCCN 2015048305| ISBN 9781118877432 (cloth) | ISBN 9781118877524
(epub)
Subjects: LCSH: Business mathematics–Computer programs. | Business–Data
processing. | JMP (Computer file) | Data mining.
Classification: LCC HF5691 .S43245 2016 | DDC 006.3/12–dc23 LC record available at http://lccn.loc.gov/2015048305

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To our families

Boaz and Noa
Liz, Lisa, and Allison
Michael, Madi, Olivia, and in memory of E.C. Jr.
Tehmi, Arjun, and in memory of Aneesh

CONTENTS

PART I PRELIMINARIES

PART II DATA EXPLORATION AND DIMENSION REDUCTION

PART III PERFORMANCE EVALUATION

PART IV PREDICTION AND CLASSIFICATION METHODS

PART V MINING RELATIONSHIPS AMONG RECORDS

PART VI FORECASTING TIME SERIES

PART VII CASES

FOREWORD

No matter what your chosen profession or place of work, your future will almost certainly be saturated with data. The modern world is defined by the bits of data pulsing from billions of keyboards and trillions of card swipes—emanating from every manner of electronic device and system—transmitted instantaneously around the globe. The sheer amount of data is measured in volumes difficult to comprehend. But it's not about how much data you have; it's what you do with it, and how quickly, that counts most. Grappling with this messy world of data and putting it to good use will be key to productive and well-functioning organizations and successful managerial careers, not just in the obvious places circling Silicon Valley such as Google and Facebook but in insurance companies, banks, auto manufacturers, airlines, hospitals, and indeed nearly everywhere.

That's where *Data Mining for Business Analytics: Concepts, Techniques and Applications with JMP Pro®* can help. Professor Shmueli and her coauthors provide a very useful guide for students of business to learn the important concepts and methods for navigating complex datasets. Born out of the authors' years of experience teaching the subject, the book has evolved from earlier editions to keep pace with the changing landscape of business analytics in graduate and undergraduate education. Most important, new with this edition is the integration of JMP Pro®, a statistical tool from SAS Institute, which is provided as the vehicle for working with data in problem sets. Learning analytics is ultimately about doing things to and with data to generate insights. Mastering one's dexterity with powerful statistical tools is a necessary and critical step in the learning process.

If you've set your sights on leading in a digital world, this book is a great place to start preparing yourself for the future.

Michael Rappa
Institute for Advanced Analytics
North Carolina State University

PREFACE

The textbook *Data Mining for Business Intelligence* first appeared in early 2007. Since then, it has been used by numerous practitioners and in many courses, ranging from dedicated data mining classes to more general business analytics courses (including our own experience teaching this material both online and in person for more than 10 years). Following feedback from instructors teaching MBA, undergraduate, and executive courses, and from students, the second edition saw revisions to some of the existing chapters and included two new topics: data visualization and time series forecasting.

This book is the first edition to fully integrate JMP Pro®[1] rather than the Microsoft Office Excel add-in, XLMiner. JMP Pro® is a desktop statistical package from SAS Institute that runs natively on Mac and Windows machines. All examples, special topics boxes, instructions, and exercises presented in this book are based on JMP 12 Pro, the professional version of JMP, which has a rich array of built- in tools for interactive data visualization, analysis, and modeling.[2]

There are other important changes in this edition. The first noticeable change is the title: other than the addition of JMP Pro®, we now use *Business Analytics* in place of *Business Intelligence*. This update reflects the change in terminology since the second edition: BI today refers mainly to reporting and data visualization (what is happening now), while BA has taken over the advanced analytics, which include predictive analytics and data mining. In this new edition we therefore also updated these terms in the book, using them as is currently common.

We added a new chapter, *Combining Methods: Ensembles and Uplift Modeling* (Chapter 13). This chapter, which is the last in Part IV on Prediction and Classification Methods, introduces two important approaches. The first—ensembles—is the combination of multiple models for improving predictive power. Ensembles have routinely proved their

[1] JMP Pro®, Version 12. SAS Institute Inc., Cary, NC 27513. See Chapter 1 for information on how to get JMP Pro®.

[2] Relevant new features in JMP Pro 13 are noted in the chapters.

usefulness in practical applications and in data mining contests. The second topic—uplift modeling—introduces an improved approach for measuring the impact of an intervention or treatment. Similar to other chapters, this new chapter includes real-world examples and end-of-chapter problems.

Other changes include the addition of two new cases based on real data (one on political persuasion and uplift modeling, and another on taxi cancellations), and the removal of one chapter, Association Rules (association rules is a feature not available in JMP 12 Pro, but will be a new feature in JMP 13 Pro).

Since the second edition's appearance, the landscape of 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 degree 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, our book is used across multiple courses. The book is designed to continue supporting the general Predictive Analytics or Data Mining course as well as supporting a set of courses in dedicated business analytics programs.

A general "Business Analytics," "Predictive Analytics," or Data Mining course, common in MBA and undergraduate programs as a one-semester elective, would cover Parts I–III, and choose a subset of methods from Parts IV and V. Instructors can choose to use cases as team assignments, class discussions, or projects. For a two-semester course, Part VI might be considered. For a set of courses in a dedicated Business Analytics program, here are a few courses that have been using the second edition of *Data Mining for Business Intelligence*:

Predictive Analytics: Supervised Learning In a dedicated Business Analytics program, the topic of Predictive Analytics is typically instructed across a set of courses. The first course would cover Parts I–IV and instructors typically choose a subset of methods from Part IV according to the course length. We recommend including the new Chapter 13 in such a course.

Predictive Analytics: Unsupervised Learning This course introduces data exploration and visualization, dimension reduction, mining relationships, and clustering (Parts III and V). If this course follows the Predictive Analytics: Supervised Learning course, then it is useful to examine examples and approaches that integrate unsupervised and supervised learning.

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

In all courses, we strongly recommend including a project component, where data are either collected by students according to their interest or provided by the instructor (e.g., from the many data mining competition datasets available). From our experience and the experience of other instructors, such projects enhance learning and provide students with an excellent opportunity to understand the strengths of data mining and the challenges that arise in working with data and solving real business problems.

ACKNOWLEDGMENTS

The authors thank the many people who assisted us in improving the first edition and improving it further in the second edition, and now in this JMP edition. Anthony Babinec, who has been using drafts of this book for years in his Data Mining courses at Statistics.com, provided us with detailed and expert corrections. The Statistics.com team has also provided valuable checking, trouble-shooting, and critiquing: Kuber Deokar, Instructional Operations Supervisor, and Shweta Jadhav and Dhanashree Vishwasrao, Assistant Teachers. We also thank the many students who have used and commented on earlier editions of this text at Statistics.com.

Similarly, Dan Toy and John Elder IV greeted our project with enthusiasm and provided detailed and useful comments on earlier drafts. Boaz Shmueli and Raquelle Azran gave detailed editorial comments and suggestions on the first two editions; Noa Shmueli provided edits on the new edition; Bruce McCullough and Adam Hughes did the same for the first edition. Ravi Bapna, who used an early draft in a Data Mining course at the Indian School of Business, provided invaluable comments and helpful suggestions. Useful comments and feedback have also come from the many instructors, too numerous to mention, who have used the book in their classes.

From the Smith School of Business at the University of Maryland, colleagues Shrivardhan Lele, Wolfgang Jank, and Paul Zantek provided practical advice and comments. We thank Robert Windle, and 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. And we thank the many MBA students for fruitful discussions and interesting data mining projects that have helped shape and improve the book.

This book would not have seen the light of day without the nurturing support of the faculty at the Sloan School of Management at MIT. Our special thanks to 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. We are grateful to the MBA students at Sloan for stimulating discussions in the class that led to refinement of the notes.

Chris Albright, Gregory Piatetsky-Shapiro, Wayne Winston, and Uday Karmarkar gave us helpful advice. Anand Bodapati provided both data and advice. 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. They helped us develop a balanced philosophical perspective on the emerging field of data mining.

Our thanks to Ajay Sathe, and his Cytel colleagues who helped launch this project: Suresh Ankolekar, Poonam Baviskar, Kuber Deokar, Rupali Desai, YogeshGajjar, Ajit Ghanekar, Ayan Khare, Bharat Lande, Dipankar Mukhopadhyay, S. V.Sabnis, Usha Sathe, Anurag Srivastava, V. Subramaniam, Ramesh Raman, and Sanhita Yeolkar.

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, Allison McGinniss, Sari Friedman, and Katrina Maceda at Wiley, and Shikha Pahuja from Thomson Digital, were all helpful and responsive as we finalized this new JMP Pro® edition.

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, Marietta Tretter at Texas A&M for her helpful comments and thoughts on the time series chapters, and Stephen Few and Ben Shneiderman for feedback and suggestions on the data visualization chapter and overall design tips.

Gregory Piatetsky-Shapiro, founder of KDNuggets.com, has been generous with his time and counsel over the many years of this project. 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.

Finally, we'd like to thank the reviewers of this 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, Chris Gotwalt, and Fang Chen. Most important, we thank John Sall, whose inovatation, 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 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*, 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

Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro®, First Edition.
Galit Shmueli, Peter C. Bruce, Mia L. Stephens, and Nitin R. Patel.

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 includes more sophisticated data analysis methods, such as statistical models and data mining 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).

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 it could charge more for staplers if a customer lived far from a Staples store.

Beware the organizational setting where analytics is a solution in search of a problem: a manager, knowing that business analytics and data mining are hot areas, decides that her organization must deploy them too, to capture that hidden value that must be lurking somewhere. Successful use of analytics and data mining requires both an understanding of the business context where value is to be captured, and an understanding of exactly what the data mining methods do.

1.2 WHAT IS DATA MINING?

In this book *data mining* refers to business analytics methods that go beyond counts, descriptive techniques, reporting, and methods based on business rules. While we do introduce data visualization, which is commonly the first step into more advanced analytics, the book focuses mostly on the more advanced data analytics tools. Specifically, data mining *includes statistical and machine learning methods that inform decision-making*, often in automated fashion. Prediction is typically an important component, often at the individual level. Rather than "what is the relationship between advertising and sales," we might be interested in "what specific advertisement, or recommended product, should be shown to a given online shopper at this moment?" Or we might be interested in clustering customers into different "personas" that receive different marketing treatment, then assigning each new prospect to one of these personas.

The era of big data has accelerated the use of data mining. Data mining methods, with their power and automaticity, have the ability to cope with huge amounts of data and extract value.

1.3 DATA MINING AND RELATED TERMS

The field of analytics is growing rapidly, both in terms of the breadth of applications, and in terms of the number of organizations using advanced analytics. As a result, there is considerable overlap and inconsistency in terms of definitions.

The term *data mining* itself means different things to different people. To the general public, it may have a general, somewhat hazy and pejorative meaning of digging through vast stores of (often personal) data in search of something interesting. One major consulting firm has a "data mining department," but its responsibilities are in the area of studying and graphing past data in search of general trends. And, to confuse matters, their more advanced predictive models are the responsibility of an "advanced analytics department." Other terms that organizations use are *predictive analytics*, *predictive modeling*, and *machine learning*.

Data mining stands at the confluence of the fields of statistics and machine learning (also known as *artificial intelligence*). A variety of techniques for exploring data and building models have been around for a long time in the world of statistics: linear regression, logistic regression, discriminant analysis, and principal components analysis, for example. But the core tenets of classical statistics—computing is difficult and data are scarce—do not apply in data mining applications where both data and computing power are plentiful.

This is what gives rise to Daryl Pregibon's description of data mining as "statistics at scale and speed" (Pregibon, 1999). Another major difference between the fields of statistics and machine learning is the focus in statistics on inference from a sample to the population regarding an "average effect"—for example, "a $1 price increase will reduce average demand by 2 boxes." In contrast, the focus in machine learning is on predicting individual records— "the predicted demand for person *i* given a $1 price increase is 1 box, while for person *j* it is 3 boxes." The emphasis that classical statistics places on inference (determining whether a pattern or interesting result might have happened by chance in our sample) is missing in data mining.

In comparison to statistics, data mining deals with large datasets in an open-ended fashion, making it impossible to put the strict limits around the question being addressed that inference would require. As a result the general approach to data mining is vulnerable to the danger of *overfitting*, where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.

In this book, we use the term *machine learning* to refer to algorithms that learn directly from data, especially local data, often in layered or iterative fashion. In contrast, we use *statistical models* to refer to methods that apply global structure to the data. A simple example is a linear regression model (statistical) versus a *k*-nearest neighbors algorithm (machine learning). A given record would be treated by linear regression in accord with an overall linear equation that applies to *all* the records. In *k*-nearest neighbors, that record would be classified in accord with the values of a small number of nearby record.

However, many practitioners, particularly those from the IT and computer science communities, use the term *machine learning* to refer to all the methods discussed in this book.

1.4 BIG DATA

Data mining and big data go hand in hand. *Big data* is a relative term—data today are big by reference to the past, and to the methods and devices available to deal with them. The challenge big data presents is often characterized by the four V's - volume, velocity, variety, and veracity. *Volume* refers to the amount of data. *Velocity* refers to the flow rate—the speed at which it is being generated and changed. *Variety* refers to the different types of data being generated (currency, dates, numbers, text, etc.). *Veracity* refers to the fact that data is being generated by organic distributed processes (e.g., millions of people signing up for services or free downloads) and not subject to the controls or quality checks that apply to data collected for a study.

Most large organizations face both the challenge and the opportunity of big data because most routine data processes now generate data that can be stored and, possibly, analyzed. The scale can be visualized by comparing the data in a traditional statistical analysis on the large size (e.g., 15 variables and 5000 records) to the Walmart database. If you consider the traditional statistical study to be the size of a period at the end of a sentence, then the Walmart database is the size of a football field. And that probably does not include other data associated with Walmart—social media data, for example, which comes in the form of unstructured text.

If the analytical challenge is substantial, so can be the reward:

- OKCupid, the dating site, uses statistical models with their data to predict what forms of message content are most likely to produce a response.
- Telenor, a Norwegian mobile phone service company, was able to reduce subscriber turnover 37% by using models to predict which customers were most likely to leave, and then lavishing attention on them.
- Allstate, the insurance company, tripled the accuracy of predicting injury liability in auto claims by incorporating more information about vehicle type.

The examples above are from Eric Siegel's *Predictive Analytics* (2013, Wiley).