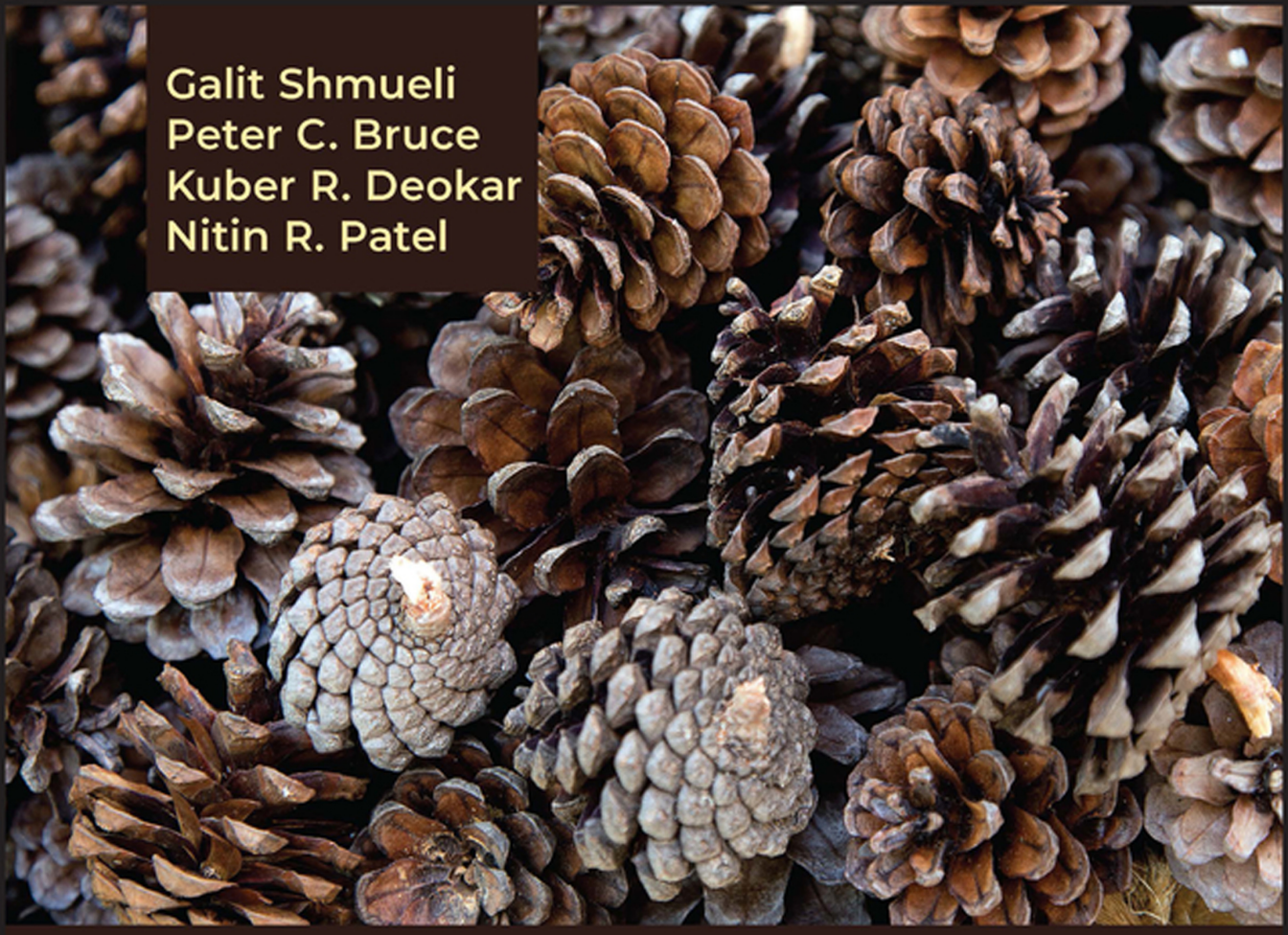
Galit Shmueli
Peter C. Bruce
Kuber R. Deokar
Nitin R. Patel



FOURTH EDITION

MACHINE LEARNING FOR BUSINESS ANALYTICS

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

WILEY

MACHINE LEARNING FOR BUSINESS ANALYTICS

MACHINE LEARNING FOR BUSINESS ANALYTICS

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

Fourth Edition

GALIT SHMUELI
National Tsing Hua University
Taipei, Taiwan

PETER C. BRUCE
statistics.com
Arlington, USA

KUBER R. DEOKAR
UpThink Edutech Services Pvt. Ltd.
Pune, India

NITIN R. PATEL
Cytel, Inc.
Cambridge, USA

WILEY

This edition first published 2023

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

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

Hardback: 9781119829836

Cover Design: Wiley
Cover Image: © Suthat_Chaitaweesap/Getty Images

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

Analytic Solver Data Mining (ASDM) for Education

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

For Instructors: Setting Up the Course Code

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

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

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

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

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

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

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

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

To our families

Boaz and Noa

Liz, Lisa, and Allison

Komal and Yash

Tehmi, Arjun, and in

memory of Aneesh

Contents

Foreword

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

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

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

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

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

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

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

SUSAN W. PALOCSAY
James Madison University 2023

Preface to the Fourth Edition

This textbook first appeared in early 2007 and has been used by numerous students and practitioners and in many courses, including our own experience teaching this material both online and in person for more than 15 years. The first edition, based on the Excel add-in Analytic Solver Data Mining (ASDM, previously XLMiner), was followed by two more Analytic Solver editions, a JMP edition, two R editions, a Python edition, a RapidMiner edition, and now this 4th edition with Analytic Solver, with its companion website, www.dataminingbook.com.

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

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

The new edition provides significant updates both in terms of ASDM and in terms of new topics and content. In addition to updating software routines and outputs that have changed or become available since the 3rd edition, this edition also incorporates updates and new material based on feedback from instructors teaching MBA, MS, undergraduate, diploma, and executive courses, and from their students as well. Importantly, this edition includes several new topics:

- A new chapter on *Responsible Data Science* (Chapter 22) covering topics of fairness, transparency, model cards and datasheets, legal considerations, and more, with an illustrative example.
- A dedicated section on *deep learning* in Chapter 11.
- The *Performance Evaluation* exposition in Chapter 5 was expanded to include further metrics (precision and recall, F1).
- A new chapter on *Generating, Comparing, and Combining Multiple Models* (Chapter 13) that covers ensembles and AutoML.

- A new chapter dedicated to *Interventions and User Feedback* (Chapter 14) that covers A/B tests, uplift modeling, and reinforcement learning.
- A new case (Loan Approval) that touches on regulatory and ethical issues.

A note about the book's title: The first two editions of the book used the title *Data Mining for Business Intelligence*. Business Intelligence today refers mainly to reporting and data visualization ("what is happening now"), while Business Analytics has taken over the "advanced analytics," which include predictive analytics and data mining. Later editions were therefore renamed *Data Mining for Business Analytics*. However, the recent AI transformation has made the term *machine learning* more popularly associated with the methods in this textbook. In this new edition, we therefore use the updated terms *Machine Learning* and *Business Analytics*.

Since the appearance of the (Analytic Solver based) second edition, the landscape of the courses using the textbook has greatly expanded: whereas initially, the book was used mainly in semester-long elective MBA-level courses, it is now used in a variety of courses in Business Analytics degrees and certificate programs, ranging from undergraduate programs, to post-graduate and executive education programs. Courses in such programs also vary in their duration and coverage. In many cases, this textbook is used across multiple courses. The book is designed to continue supporting the general "Predictive Analytics" or "Data Mining" course as well as supporting a set of courses in dedicated business analytics programs.

A general "Business Analytics," "Predictive Analytics," or "Machine Learning" course, common in MBA and undergraduate programs as a one-semester elective, would cover Parts I–III, and choose a subset of methods from Parts IV and V. Instructors can choose to use cases as team assignments, class discussions, or projects. For a two-semester course, Part VII might be considered, and we recommend introducing Part VIII (Data Analytics).

For a set of courses in a dedicated business analytics program, here are a few courses that have been using our book:

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

Predictive Analytics—Unsupervised Learning: This course introduces data exploration and visualization, dimension reduction, mining relationships, and clustering (Parts II and VI). If this course follows the Predictive Analytics: Supervised Learning course, then it is useful to examine examples and approaches that integrate unsupervised and supervised learning, such as Part VIII on Data Analytics.

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

Advanced Analytics: A course that integrates the learnings from Predictive Analytics (supervised and unsupervised learning) can focus on Part VIII: Data Analytics, where social network analytics and text mining are introduced, and responsible data science is discussed. Such a course might also include Chapter 13, Generating, Comparing, and Combining Multiple Models and AutoML from Part IV, as well as Part V, which covers experiments, uplift, and reinforcement learning. Some instructors choose to use the Cases (Chapter 23) in such a course.

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

GALIT SHMUELI, PETER C. BRUCE, KUBER R. DEOKAR, AND NITIN R. PATEL
2023

Acknowledgments

We thank the many people who assisted us in improving the book from its inception as *Data Mining for Business Intelligence* in 2006 (using XLMiner, now Analytic Solver), its reincarnation as *Data Mining for Business Analytics*, and now *Machine Learning for Business Analytics*, including translations in Chinese and Korean and versions supporting Analytic Solver Data Mining, R, Python, SAS JMP, and RapidMiner.

Anthony Babinec, who has been using earlier editions of this book for years in his data mining courses at Statistics.com, provided us with detailed and expert corrections. Dan Toy and John Elder IV greeted our project with early enthusiasm and provided detailed and useful comments on initial drafts. Ravi Bapna, who used an early draft in a data mining course at the Indian School of Business and later at University of Minnesota, has provided invaluable comments and helpful suggestions since the book's start.

Many of the instructors, teaching assistants, and students using earlier editions of the book have contributed invaluable feedback both directly and indirectly, through fruitful discussions, learning journeys, and interesting data mining projects that have helped shape and improve the book. These include MBA students from the University of Maryland, MIT, the Indian School of Business, National Tsing Hua University, and Statistics.com. Instructors from many universities and teaching programs, too numerous to list, have supported and helped improve the book since its inception.

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

Colleagues and family members have been providing ongoing feedback and assistance with this book project. Vijay Kamble at UIC and Travis Greene at NTHU have provided valuable help with the section on reinforcement learning. Boaz Shmueli and Raquelle Azran gave detailed editorial comments

and suggestions on the first two editions; Bruce McCullough and Adam Hughes did the same for the first edition. Noa Shmueli provided careful proofs of the third edition. Ran Shenberger offered design tips. Ken Strasma, founder of the microtargeting firm HaystaqDNA and director of targeting for the 2004 Kerry campaign and the 2008 Obama campaign, provided the scenario and data for the section on uplift modeling. We also thank Jen Golbeck, Professor in the College of Information Studies at the University of Maryland and author of *Analyzing the Social Web*, whose book inspired our presentation in the chapter on social network analytics. Randall Pruim contributed extensively to the chapter on visualization.

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

Susan Palocsay and Mia Stephens have provided suggestions and feedback on numerous occasions, as has Margret Bjarnadottir. We also thank Catherine Plaisant at the University of Maryland's Human–Computer Interaction Lab, who helped out in a major way by contributing exercises and illustrations to the data visualization chapter. Gregory Piatetsky-Shapiro, founder of KDNuggets.com, was generous with his time and counsel in the early years of this project.

We thank colleagues at the Sloan School of Management at MIT for their support during the formative stage of this book—Dimitris Bertsimas, James Orlin, Robert Freund, Roy Welsch, Gordon Kaufmann, and Gabriel Bitran. As teaching assistants for the data mining course at Sloan, Adam Mersereau gave detailed comments on the notes and cases that were the genesis of this book, Romy Shioda helped with the preparation of several cases and exercises used here, and Mahesh Kumar helped with the material on clustering.

Colleagues at the University of Maryland's Smith School of Business: Shrivardhan Lele, Wolfgang Jank, and Paul Zantek provided practical advice and comments. We thank Robert Windle, and University of Maryland MBA students Timothy Roach, Pablo Macouzet, and Nathan Birckhead for invaluable datasets. We also thank MBA students Rob Whitener and Daniel Curtis for the heatmap and map charts.

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

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

System's Decision Support and Analytics (SIGDSA) community for ideas and advice that helped the development of the book.

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

Lastly, we thank the folks at Wiley for this successful journey of nearly two decades. Steve Quigley at Wiley showed confidence in this book from the beginning and helped us navigate through the publishing process with great speed. Curt Hinrichs' vision, tips, and encouragement helped bring the first edition of this book to the starting gate. Jon Gurstelle guided us through additional editions and translations. Brett Kurzman has taken over the reins and is now shepherding the project. Becky Cowan, Sarah Lemore, and Kavya Ramu greatly assisted us in pushing ahead and finalizing this 4th edition. We are also especially grateful to Amy Hendrickson, who assisted with typesetting and making this book beautiful.