

# DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES,  
AND APPLICATIONS IN **R**

Galit Shmueli • Peter C. Bruce  
Inbal Yahav • Nitin R. Patel  
Kenneth C. Lichtendahl Jr.



WILEY



**DATA MINING  
FOR BUSINESS ANALYTICS**



---

# **DATA MINING FOR BUSINESS ANALYTICS**

**Concepts, Techniques, and Applications in R**

---

**GALIT SHMUELI**

**PETER C. BRUCE**

**INBAL YAHAV**

**NITIN R. PATEL**

**KENNETH C. LICHTENDAHL, JR.**

**WILEY**

This edition first published 2018

© 2018 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, and Kenneth C. Lichtendahl Jr. to be identified as the authors of this work has been asserted in accordance with law.

*Registered Offices*

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

*Editorial Office*

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](http://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*

The publisher and the authors 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 fitness for a particular purpose. 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 every situation. In view of on-going 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. The fact that an organization or website is referred to in this work as a citation and/or potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. 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. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising here from.

*Library of Congress Cataloging-in-Publication Data applied for*

Hardback: 9781118879368

Cover Design: Wiley

Cover Image: © Achim Mittler, Frankfurt am Main/Gettyimages

Set in 11.5/14.5pt BemboStd by Aptara Inc., New Delhi, India  
Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

*The beginning of wisdom is this:  
Get wisdom, and whatever else you get, get insight.*

ראשית חכמה, קנה חכמה; ובכל-קנייך, קנה בינה.

– Proverbs 4:7







# Contents

---

<b>Foreword by Gareth James</b>	<b>xix</b>
<b>Foreword by Ravi Bapna</b>	<b>xxi</b>
<b>Preface to the R Edition</b>	<b>xxiii</b>
<b>Acknowledgments</b>	<b>xxvii</b>

## **PART I PRELIMINARIES**

---

<b>CHAPTER 1 Introduction</b>	<b>3</b>
1.1 What Is Business Analytics? . . . . .	3
1.2 What Is Data Mining? . . . . .	5
1.3 Data Mining and Related Terms . . . . .	5
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 . . . . .	9
1.8 Road Maps to This Book . . . . .	11
Order of Topics . . . . .	11
<b>CHAPTER 2 Overview of the Data Mining Process</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 Core Ideas in Data Mining . . . . .	16
Classification . . . . .	16
Prediction . . . . .	16
Association Rules and Recommendation Systems . . . . .	16
Predictive Analytics . . . . .	17
Data Reduction and Dimension Reduction . . . . .	17
Data Exploration and Visualization . . . . .	17
Supervised and Unsupervised Learning . . . . .	18
2.3 The Steps in Data Mining . . . . .	19
2.4 Preliminary Steps . . . . .	21
Organization of Datasets . . . . .	21
Predicting Home Values in the West Roxbury Neighborhood . . . . .	21

## VIII CONTENTS

	Loading and Looking at the Data in R . . . . .	22
	Sampling from a Database . . . . .	24
	Oversampling Rare Events in Classification Tasks . . . . .	25
	Preprocessing and Cleaning the Data . . . . .	26
2.5	Predictive Power and Overfitting . . . . .	33
	Overfitting . . . . .	33
	Creation and Use of Data Partitions . . . . .	35
2.6	Building a Predictive Model . . . . .	38
	Modeling Process . . . . .	39
2.7	Using R for Data Mining on a Local Machine . . . . .	43
2.8	Automating Data Mining Solutions . . . . .	43
	Data Mining Software: The State of the Market (by Herb Edelstein) . . . . .	45
	Problems . . . . .	49

## **PART II DATA EXPLORATION AND DIMENSION REDUCTION**

---

### **CHAPTER 3 Data Visualization** **55**

3.1	Uses of Data Visualization . . . . .	55
	Base R or ggplot? . . . . .	57
3.2	Data Examples . . . . .	57
	Example 1: Boston Housing Data . . . . .	57
	Example 2: Ridership on Amtrak Trains . . . . .	59
3.3	Basic Charts: Bar Charts, Line Graphs, and Scatter Plots . . . . .	59
	Distribution Plots: Boxplots and Histograms . . . . .	61
	Heatmaps: Visualizing Correlations and Missing Values . . . . .	64
3.4	Multidimensional Visualization . . . . .	67
	Adding Variables: Color, Size, Shape, Multiple Panels, and Animation . . . . .	67
	Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering . . . . .	70
	Reference: Trend Lines and Labels . . . . .	74
	Scaling up to Large Datasets . . . . .	74
	Multivariate Plot: Parallel Coordinates Plot . . . . .	75
	Interactive Visualization . . . . .	77
3.5	Specialized Visualizations . . . . .	80
	Visualizing Networked Data . . . . .	80
	Visualizing Hierarchical Data: Treemaps . . . . .	82
	Visualizing Geographical Data: Map Charts . . . . .	83
3.6	Summary: Major Visualizations and Operations, by Data Mining Goal . . . . .	86
	Prediction . . . . .	86
	Classification . . . . .	86
	Time Series Forecasting . . . . .	86
	Unsupervised Learning . . . . .	87
	Problems . . . . .	88

### **CHAPTER 4 Dimension Reduction** **91**

4.1	Introduction . . . . .	91
4.2	Curse of Dimensionality . . . . .	92

4.3	Practical Considerations . . . . .	92
	Example 1: House Prices in Boston . . . . .	93
4.4	Data Summaries . . . . .	94
	Summary Statistics . . . . .	94
	Aggregation and Pivot Tables . . . . .	96
4.5	Correlation Analysis . . . . .	97
4.6	Reducing the Number of Categories in Categorical Variables . . . . .	99
4.7	Converting a Categorical Variable to a Numerical Variable . . . . .	99
4.8	Principal Components Analysis . . . . .	101
	Example 2: Breakfast Cereals . . . . .	101
	Principal Components . . . . .	106
	Normalizing the Data . . . . .	107
	Using Principal Components for Classification and Prediction . . . . .	109
4.9	Dimension Reduction Using Regression Models . . . . .	111
4.10	Dimension Reduction Using Classification and Regression Trees . . . . .	111
	Problems . . . . .	112

## PART III PERFORMANCE EVALUATION

---

<b>CHAPTER 5</b>	<b>Evaluating Predictive Performance</b>	<b>117</b>
5.1	Introduction . . . . .	117
5.2	Evaluating Predictive Performance . . . . .	118
	Naïve Benchmark: The Average . . . . .	118
	Prediction Accuracy Measures . . . . .	119
	Comparing Training and Validation Performance . . . . .	121
	Lift Chart . . . . .	121
5.3	Judging Classifier Performance . . . . .	122
	Benchmark: The Naïve Rule . . . . .	124
	Class Separation . . . . .	124
	The Confusion (Classification) Matrix . . . . .	124
	Using the Validation Data . . . . .	126
	Accuracy Measures . . . . .	126
	Propensities and Cutoff for Classification . . . . .	127
	Performance in Case of Unequal Importance of Classes . . . . .	131
	Asymmetric Misclassification Costs . . . . .	133
	Generalization to More Than Two Classes . . . . .	135
5.4	Judging Ranking Performance . . . . .	136
	Lift Charts for Binary Data . . . . .	136
	Decile Lift Charts . . . . .	138
	Beyond Two Classes . . . . .	139
	Lift Charts Incorporating Costs and Benefits . . . . .	139
	Lift as a Function of Cutoff . . . . .	140
5.5	Oversampling . . . . .	140
	Oversampling the Training Set . . . . .	144

Evaluating Model Performance Using a Non-oversampled Validation Set . . . . .	144
Evaluating Model Performance if Only Oversampled Validation Set Exists . . . . .	144
Problems . . . . .	147

**PART IV PREDICTION AND CLASSIFICATION METHODS**

---

**CHAPTER 6 Multiple Linear Regression 153**

6.1 Introduction . . . . .	153
6.2 Explanatory vs. Predictive Modeling . . . . .	154
6.3 Estimating the Regression Equation and Prediction . . . . .	156
Example: Predicting the Price of Used Toyota Corolla Cars . . . . .	156
6.4 Variable Selection in Linear Regression . . . . .	161
Reducing the Number of Predictors . . . . .	161
How to Reduce the Number of Predictors . . . . .	162
Problems . . . . .	169

**CHAPTER 7 *k*-Nearest Neighbors (*k*NN) 173**

7.1 The <i>k</i> -NN Classifier (Categorical Outcome) . . . . .	173
Determining Neighbors . . . . .	173
Classification Rule . . . . .	174
Example: Riding Mowers . . . . .	175
Choosing <i>k</i> . . . . .	176
Setting the Cutoff Value . . . . .	179
<i>k</i> -NN with More Than Two Classes . . . . .	180
Converting Categorical Variables to Binary Dummies . . . . .	180
7.2 <i>k</i> -NN for a Numerical Outcome . . . . .	180
7.3 Advantages and Shortcomings of <i>k</i> -NN Algorithms . . . . .	182
Problems . . . . .	184

**CHAPTER 8 The Naive Bayes Classifier 187**

8.1 Introduction . . . . .	187
Cutoff Probability Method . . . . .	188
Conditional Probability . . . . .	188
Example 1: Predicting Fraudulent Financial Reporting . . . . .	188
8.2 Applying the Full (Exact) Bayesian Classifier . . . . .	189
Using the “Assign to the Most Probable Class” Method . . . . .	190
Using the Cutoff Probability Method . . . . .	190
Practical Difficulty with the Complete (Exact) Bayes Procedure . . . . .	190
Solution: Naive Bayes . . . . .	191
The Naive Bayes Assumption of Conditional Independence . . . . .	192
Using the Cutoff Probability Method . . . . .	192
Example 2: Predicting Fraudulent Financial Reports, Two Predictors . . . . .	193
Example 3: Predicting Delayed Flights . . . . .	194
8.3 Advantages and Shortcomings of the Naive Bayes Classifier . . . . .	199
Problems . . . . .	202

<b>CHAPTER 9</b>	<b>Classification and Regression Trees</b>	<b>205</b>
9.1	Introduction . . . . .	205
9.2	Classification Trees . . . . .	207
	Recursive Partitioning . . . . .	207
	Example 1: Riding Mowers . . . . .	207
	Measures of Impurity . . . . .	210
	Tree Structure . . . . .	214
	Classifying a New Record . . . . .	214
9.3	Evaluating the Performance of a Classification Tree . . . . .	215
	Example 2: Acceptance of Personal Loan . . . . .	215
9.4	Avoiding Overfitting . . . . .	216
	Stopping Tree Growth: Conditional Inference Trees . . . . .	221
	Pruning the Tree . . . . .	222
	Cross-Validation . . . . .	222
	Best-Pruned Tree . . . . .	224
9.5	Classification Rules from Trees . . . . .	226
9.6	Classification Trees for More Than Two Classes . . . . .	227
9.7	Regression Trees . . . . .	227
	Prediction . . . . .	228
	Measuring Impurity . . . . .	228
	Evaluating Performance . . . . .	229
9.8	Improving Prediction: Random Forests and Boosted Trees . . . . .	229
	Random Forests . . . . .	229
	Boosted Trees . . . . .	231
9.9	Advantages and Weaknesses of a Tree . . . . .	232
	Problems . . . . .	234
<b>CHAPTER 10</b>	<b>Logistic Regression</b>	<b>237</b>
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: Computing Parameter Estimates . . . . .	243
	Interpreting Results in Terms of Odds (for a Profiling Goal) . . . . .	244
10.4	Evaluating Classification Performance . . . . .	247
	Variable Selection . . . . .	248
10.5	Example of Complete Analysis: Predicting Delayed Flights . . . . .	250
	Data Preprocessing . . . . .	251
	Model-Fitting and Estimation . . . . .	254
	Model Interpretation . . . . .	254
	Model Performance . . . . .	254
	Variable Selection . . . . .	257
10.6	Appendix: Logistic Regression for Profiling . . . . .	259
	Appendix A: Why Linear Regression Is Problematic for a Categorical Outcome . . . . .	259

Appendix B: Evaluating Explanatory Power . . . . . 261  
Appendix C: Logistic Regression for More Than Two Classes . . . . . 264  
Problems . . . . . 268

**CHAPTER 11 Neural Nets 271**

11.1 Introduction . . . . . 271  
11.2 Concept and Structure of a Neural Network . . . . . 272  
11.3 Fitting a Network to Data . . . . . 273  
    Example 1: Tiny Dataset . . . . . 273  
    Computing Output of Nodes . . . . . 274  
    Preprocessing the Data . . . . . 277  
    Training the Model . . . . . 278  
    Example 2: Classifying Accident Severity . . . . . 282  
    Avoiding Overfitting . . . . . 283  
    Using the Output for Prediction and Classification . . . . . 283  
11.4 Required User Input . . . . . 285  
11.5 Exploring the Relationship Between Predictors and Outcome . . . . . 286  
11.6 Advantages and Weaknesses of Neural Networks . . . . . 288  
Problems . . . . . 289

**CHAPTER 12 Discriminant Analysis 291**

12.1 Introduction . . . . . 291  
    Example 1: Riding Mowers . . . . . 292  
    Example 2: Personal Loan Acceptance . . . . . 292  
12.2 Distance of a Record from a Class . . . . . 294  
12.3 Fisher’s Linear Classification Functions . . . . . 295  
12.4 Classification Performance of Discriminant Analysis . . . . . 298  
12.5 Prior Probabilities . . . . . 300  
12.6 Unequal Misclassification Costs . . . . . 300  
12.7 Classifying More Than Two Classes . . . . . 301  
    Example 3: Medical Dispatch to Accident Scenes . . . . . 301  
12.8 Advantages and Weaknesses . . . . . 304  
Problems . . . . . 305

**CHAPTER 13 Combining Methods: Ensembles and Uplift Modeling 309**

13.1 Ensembles . . . . . 309  
    Why Ensembles Can Improve Predictive Power . . . . . 310  
    Simple Averaging . . . . . 312  
    Bagging . . . . . 313  
    Boosting . . . . . 313  
    Bagging and Boosting in R . . . . . 313  
    Advantages and Weaknesses of Ensembles . . . . . 313  
13.2 Uplift (Persuasion) Modeling . . . . . 315  
    A-B Testing . . . . . 316

Uplift . . . . .	316
Gathering the Data . . . . .	317
A Simple Model . . . . .	318
Modeling Individual Uplift . . . . .	319
Computing Uplift with R . . . . .	320
Using the Results of an Uplift Model . . . . .	320
13.3 Summary . . . . .	322
Problems . . . . .	323

## **PART V MINING RELATIONSHIPS AMONG RECORDS**

---

### **CHAPTER 14 Association Rules and Collaborative Filtering 327**

14.1 Association Rules . . . . .	327
Discovering Association Rules in Transaction Databases . . . . .	328
Example 1: Synthetic Data on Purchases of Phone Faceplates . . . . .	328
Generating Candidate Rules . . . . .	328
The Apriori Algorithm . . . . .	331
Selecting Strong Rules . . . . .	331
Data Format . . . . .	333
The Process of Rule Selection . . . . .	334
Interpreting the Results . . . . .	335
Rules and Chance . . . . .	337
Example 2: Rules for Similar Book Purchases . . . . .	338
14.2 Collaborative Filtering . . . . .	340
Data Type and Format . . . . .	341
Example 3: Netflix Prize Contest . . . . .	341
User-Based Collaborative Filtering: “People Like You” . . . . .	342
Item-Based Collaborative Filtering . . . . .	345
Advantages and Weaknesses of Collaborative Filtering . . . . .	346
Collaborative Filtering vs. Association Rules . . . . .	347
14.3 Summary . . . . .	349
Problems . . . . .	350

### **CHAPTER 15 Cluster Analysis 355**

15.1 Introduction . . . . .	355
Example: Public Utilities . . . . .	357
15.2 Measuring Distance Between Two Records . . . . .	359
Euclidean Distance . . . . .	359
Normalizing Numerical Measurements . . . . .	360
Other Distance Measures for Numerical Data . . . . .	360
Distance Measures for Categorical Data . . . . .	363
Distance Measures for Mixed Data . . . . .	364
15.3 Measuring Distance Between Two Clusters . . . . .	364
Minimum Distance . . . . .	364
Maximum Distance . . . . .	364

Average Distance . . . . .	365
Centroid Distance . . . . .	365
15.4 Hierarchical (Agglomerative) Clustering . . . . .	366
Single Linkage . . . . .	367
Complete Linkage . . . . .	368
Average Linkage . . . . .	368
Centroid Linkage . . . . .	368
Ward’s Method . . . . .	368
Dendrograms: Displaying Clustering Process and Results . . . . .	369
Validating Clusters . . . . .	371
Limitations of Hierarchical Clustering . . . . .	373
15.5 Non-Hierarchical Clustering: The $k$ -Means Algorithm . . . . .	374
Choosing the Number of Clusters ( $k$ ) . . . . .	375
Problems . . . . .	380

## PART VI FORECASTING TIME SERIES

---

### CHAPTER 16 Handling Time Series 385

16.1 Introduction . . . . .	385
16.2 Descriptive vs. Predictive Modeling . . . . .	387
16.3 Popular Forecasting Methods in Business . . . . .	387
Combining Methods . . . . .	387
16.4 Time Series Components . . . . .	388
Example: Ridership on Amtrak Trains . . . . .	388
16.5 Data-Partitioning and Performance Evaluation . . . . .	393
Benchmark Performance: Naïve Forecasts . . . . .	393
Generating Future Forecasts . . . . .	394
Problems . . . . .	396

### CHAPTER 17 Regression-Based Forecasting 399

17.1 A Model with Trend . . . . .	399
Linear Trend . . . . .	399
Exponential Trend . . . . .	403
Polynomial Trend . . . . .	405
17.2 A Model with Seasonality . . . . .	405
17.3 A Model with Trend and Seasonality . . . . .	409
17.4 Autocorrelation and ARIMA Models . . . . .	410
Computing Autocorrelation . . . . .	411
Improving Forecasts by Integrating Autocorrelation Information . . . . .	414
Evaluating Predictability . . . . .	418
Problems . . . . .	420



<b>CHAPTER 18 Smoothing Methods</b>	<b>431</b>
18.1 Introduction . . . . .	431
18.2 Moving Average . . . . .	432
Centered Moving Average for Visualization . . . . .	432
Trailing Moving Average for Forecasting . . . . .	433
Choosing Window Width ( $w$ ) . . . . .	437
18.3 Simple Exponential Smoothing . . . . .	437
Choosing Smoothing Parameter $\alpha$ . . . . .	438
Relation Between Moving Average and Simple Exponential Smoothing . . . . .	438
18.4 Advanced Exponential Smoothing . . . . .	440
Series with a Trend . . . . .	440
Series with a Trend and Seasonality . . . . .	441
Series with Seasonality (No Trend) . . . . .	441
Problems . . . . .	444

## PART VII DATA ANALYTICS

---

<b>CHAPTER 19 Social Network Analytics</b>	<b>453</b>
19.1 Introduction . . . . .	453
19.2 Directed vs. Undirected Networks . . . . .	455
19.3 Visualizing and Analyzing Networks . . . . .	456
Graph Layout . . . . .	456
Edge List . . . . .	458
Adjacency Matrix . . . . .	459
Using Network Data in Classification and Prediction . . . . .	459
19.4 Social Data Metrics and Taxonomy . . . . .	460
Node-Level Centrality Metrics . . . . .	461
Egocentric Network . . . . .	461
Network Metrics . . . . .	463
19.5 Using Network Metrics in Prediction and Classification . . . . .	465
Link Prediction . . . . .	465
Entity Resolution . . . . .	465
Collaborative Filtering . . . . .	466
19.6 Collecting Social Network Data with R . . . . .	469
19.7 Advantages and Disadvantages . . . . .	472
Problems . . . . .	474
<b>CHAPTER 20 Text Mining</b>	<b>477</b>
20.1 Introduction . . . . .	477
20.2 The Tabular Representation of Text: Term-Document Matrix and “Bag-of-Words” . . . . .	478
20.3 Bag-of-Words vs. Meaning Extraction at Document Level . . . . .	479
20.4 Preprocessing the Text . . . . .	480
Tokenization . . . . .	482
Text Reduction . . . . .	483

	Presence/Absence vs. Frequency . . . . .	485
	Term Frequency–Inverse Document Frequency (TF-IDF) . . . . .	485
	From Terms to Concepts: Latent Semantic Indexing . . . . .	486
	Extracting Meaning . . . . .	487
20.5	Implementing Data Mining Methods . . . . .	487
20.6	Example: Online Discussions on Autos and Electronics . . . . .	488
	Importing and Labeling the Records . . . . .	488
	Text Preprocessing in R . . . . .	489
	Producing a Concept Matrix . . . . .	489
	Fitting a Predictive Model . . . . .	490
	Prediction . . . . .	490
20.7	Summary . . . . .	492
	Problems . . . . .	493

## PART VIII CASES

---

<b>CHAPTER 21</b>	<b>Cases</b>	<b>497</b>
21.1	Charles Book Club . . . . .	497
	The Book Industry . . . . .	497
	Database Marketing at Charles . . . . .	498
	Data Mining Techniques . . . . .	500
	Assignment . . . . .	502
21.2	German Credit . . . . .	503
	Background . . . . .	503
	Data . . . . .	504
	Assignment . . . . .	505
21.3	Tayko Software Cataloger . . . . .	508
	Background . . . . .	508
	The Mailing Experiment . . . . .	508
	Data . . . . .	508
	Assignment . . . . .	510
21.4	Political Persuasion . . . . .	511
	Background . . . . .	511
	Predictive Analytics Arrives in US Politics . . . . .	511
	Political Targeting . . . . .	512
	Uplift . . . . .	512
	Data . . . . .	513
	Assignment . . . . .	514
21.5	Taxi Cancellations . . . . .	515
	Business Situation . . . . .	515
	Assignment . . . . .	515
21.6	Segmenting Consumers of Bath Soap . . . . .	516
	Business Situation . . . . .	516
	Key Problems . . . . .	517
	Data . . . . .	517

Measuring Brand Loyalty . . . . .	517
Assignment . . . . .	519
21.7 Direct-Mail Fundraising . . . . .	519
Background . . . . .	519
Data . . . . .	520
Assignment . . . . .	521
21.8 Catalog Cross-Selling . . . . .	522
Background . . . . .	522
Assignment . . . . .	522
21.9 Predicting Bankruptcy . . . . .	523
Predicting Corporate Bankruptcy . . . . .	523
Assignment . . . . .	524
21.10 Time Series Case: Forecasting Public Transportation Demand . . . . .	526
Background . . . . .	526
Problem Description . . . . .	526
Available Data . . . . .	526
Assignment Goal . . . . .	526
Assignment . . . . .	527
Tips and Suggested Steps . . . . .	527
<b>References</b>	<b>529</b>
<b>Data Files Used in the Book</b>	<b>531</b>
<b>Index</b>	<b>533</b>





## Foreword by Gareth James

---

The field of statistics has existed in one form or another for 200 years, and by the second half of the 20th century had evolved into a well-respected and essential academic discipline. However, its prominence expanded rapidly in the 1990s with the explosion of new, and enormous, data sources. For the first part of this century, much of this attention was focused on biological applications, in particular, genetics data generated as a result of the sequencing of the human genome. However, the last decade has seen a dramatic increase in the availability of data in the business disciplines, and a corresponding interest in business-related statistical applications.

The impact has been profound. Ten years ago, when I was able to attract a full class of MBA students to my new statistical learning elective, my colleagues were astonished because our department struggled to fill most electives. Today, we offer a Masters in Business Analytics, which is the largest specialized masters program in the school and has application volume rivaling those of our MBA programs. Our department's faculty size and course offerings have increased dramatically, yet the MBA students are still complaining that the classes are all full. Google's chief economist, Hal Varian, was indeed correct in 2009 when he stated that "the sexy job in the next 10 years will be statisticians."

This demand is driven by a simple, but undeniable, fact. Business analytics solutions have produced significant and measurable improvements in business performance, on multiple dimensions and in numerous settings, and as a result, there is a tremendous demand for individuals with the requisite skill set. However, training students in these skills is challenging given that, in addition to the obvious required knowledge of statistical methods, they need to understand business-related issues, possess strong communication skills, and be comfortable dealing with multiple computational packages. Most statistics texts concentrate on abstract training in classical methods, without much emphasis on practical, let alone business, applications.

This book has by far the most comprehensive review of business analytics methods that I have ever seen, covering everything from classical approaches such as linear and logistic regression, through to modern methods like neural

networks, bagging and boosting, and even much more business specific procedures such as social network analysis and text mining. If not the bible, it is at the least a definitive manual on the subject. However, just as important as the list of topics, is the way that they are all presented in an applied fashion using business applications. Indeed the last chapter is entirely dedicated to 10 separate cases where business analytics approaches can be applied.

In this latest edition, the authors have added an important new dimension in the form of the R software package. Easily the most widely used and influential open source statistical software, R has become the go-to tool for such purposes. With literally hundreds of freely available add-on packages, R can be used for almost any business analytics related problem. The book provides detailed descriptions and code involving applications of R in numerous business settings, ensuring that the reader will actually be able to apply their knowledge to real-life problems.

We recently introduced a business analytics course into our required MBA core curriculum and I intend to make heavy use of this book in developing the syllabus. I'm confident that it will be an indispensable tool for any such course.

GARETH JAMES

*Marshall School of Business, University of Southern California, 2017*



## Foreword by Ravi Bapna

---

Data is the new gold—and mining this gold to create business value in today’s context of a highly networked and digital society requires a skillset that we haven’t traditionally delivered in business or statistics or engineering programs on their own. For those businesses and organizations that feel overwhelmed by today’s Big Data, the phrase *you ain’t seen nothing yet* comes to mind. Yesterday’s three major sources of Big Data—the 20+ years of investment in enterprise systems (ERP, CRM, SCM, ...), the 3 billion plus people on the online social grid, and the close to 5 billion people carrying increasingly sophisticated mobile devices—are going to be dwarfed by tomorrow’s smarter physical ecosystems fueled by the Internet of Things (IoT) movement.

The idea that we can use sensors to connect physical objects such as homes, automobiles, roads, even garbage bins and streetlights, to digitally optimized systems of governance goes hand in glove with bigger data and the need for deeper analytical capabilities. We are not far away from a smart refrigerator sensing that you are short on, say, eggs, populating your grocery store’s mobile app’s shopping list, and arranging a Task Rabbit to do a grocery run for you. Or the refrigerator negotiating a deal with an Uber driver to deliver an evening meal to you. Nor are we far away from sensors embedded in roads and vehicles that can compute traffic congestion, track roadway wear and tear, record vehicle use and factor these into dynamic usage-based pricing, insurance rates, and even taxation. This brave new world is going to be fueled by analytics and the ability to harness data for competitive advantage.

Business Analytics is an emerging discipline that is going to help us ride this new wave. This new Business Analytics discipline requires individuals who are grounded in the fundamentals of business such that they know the right questions to ask, who have the ability to harness, store, and optimally process vast datasets from a variety of structured and unstructured sources, and who can then use an array of techniques from machine learning and statistics to uncover new insights for decision-making. Such individuals are a rare commodity today, but their creation has been the focus of this book for a decade now. This book’s forte is that it relies on explaining the core set of concepts required for today’s business analytics professionals using real-world data-rich cases in a hands-on manner,

without sacrificing academic rigor. It provides a modern day foundation for Business Analytics, the notion of linking the x's to the y's of interest in a predictive sense. I say this with the confidence of someone who was probably the first adopter of the zeroth edition of this book (Spring 2006 at the Indian School of Business).

I can't say enough about the long-awaited R edition. R is my go-to platform for analytics these days. It's also used by a wide variety of instructors in our MS-Business Analytics program. The open-innovation paradigm used by R is one key part of the analytics perfect storm, the other components being the advances in computing and the business appetite for data-driven decision-making.

I look forward to using the book in multiple fora, in executive education, in MBA classrooms, in MS-Business Analytics programs, and in Data Science bootcamps. I trust you will too!

RAVI BAPNA

*Carlson School of Management, University of Minnesota, 2017*





## Preface to the R Edition

---

This textbook first appeared in early 2007 and has been used by numerous students and 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). The first edition, based on the Excel add-in XLMiner, was followed by two more XLMiner editions, a JMP edition, and now this R edition, with its companion website, [www.dataminingbook.com](http://www.dataminingbook.com).

This new R edition, which relies on the free and open-source R software, presents output from R, as well as the code used to produce that output, including specification of a variety of packages and functions. Unlike computer-science or statistics-oriented textbooks, the focus in this book is on data mining concepts, and how to implement the associated algorithms in R. We assume a basic facility with R.

For this R edition, two new co-authors stepped on board—Inbal Yahav and Casey Lichtendahl—bringing both expertise teaching business analytics courses using R and data mining consulting experience in business and government. Such practical experience is important, since the open-source nature of R software makes available a plethora of approaches, packages, and functions available for data mining. Given the main goal of this book—to introduce data mining concepts using R software for illustration—our challenge was to choose an R code cocktail that supports highlighting the important concepts. In addition to providing R code and output, this edition also incorporates updates and new material based on feedback from instructors teaching MBA, undergraduate, diploma, and executive courses, and from their students as well.

One update, compared to the first two editions of the book, is the title: we now use *Business Analytics* in place of *Business Intelligence*. This reflects the change in terminology since the second edition: 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. In this new edition, we therefore use the updated terms.

This R edition includes the material that was recently added in the third edition of the original (XLMiner-based) book:

- Social network analysis
- Text mining
- Ensembles
- Uplift modeling
- Collaborative filtering

Since the appearance of the (XLMiner-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 “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, and we recommend introducing the new Part VII (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–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, as well as the new “Part VII: Data Analytics.”

**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, such as the new part 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). Such a course can focus on Part VII: Data Analytics, where social network analytics and text mining are introduced. Some instructors choose to use the Cases (Chapter 21) 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 data mining 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 data mining and the challenges that arise in the process.





# Acknowledgments

---

We thank the many people who assisted us in improving the first three editions of the initial XLMiner version of this book and the JMP edition, as well as those who helped with comments on early drafts of this R edition. 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, 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.

Several professors have been especially helpful with this R edition: Hayri Tongarlak, Prashant Joshi (UKA Tarsadia University), Jay Annadatha, Roger Bohn, and Sridhar Vaithianathan provided detailed comments and R code files for the companion website; Scott Nestler has been a helpful friend of this book project from the beginning.

Kuber Deokar, instructional operations supervisor at Statistics.com, has been unstinting in his assistance, support, and detailed attention. We also thank Shweta Jadhav and Dhanashree Vishwasrao, 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. 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, director of the Social Intelligence Lab 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, has been generous with his time and counsel over the many years of this project.

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.

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 a decade ago, they helped us develop a balanced philosophical perspective on the emerging field of data mining.