



# Contents

---

<b>Foreword</b>	<b>XXI</b>
<b>Preface to the third edition</b>	<b>XXIII</b>
<b>Preface to the first edition</b>	<b>XXVII</b>
<b>Acknowledgments</b>	<b>XXIX</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 . . . . .	7
1.5 Data Science . . . . .	8
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 . . . . .	17
Predictive Analytics . . . . .	17
Data Reduction and Dimension Reduction . . . . .	17
Data Exploration and Visualization . . . . .	18
Supervised and Unsupervised Learning . . . . .	18
2.3 The Steps in Data Mining . . . . .	19
2.4 Preliminary Steps . . . . .	21
Organization of Datasets . . . . .	21
Sampling from a Database . . . . .	21
Oversampling Rare Events in Classification Tasks . . . . .	22

- Preprocessing and Cleaning the Data . . . . . 23
- 2.5 Predictive Power and Overfitting . . . . . 27
  - Creation and Use of Data Partitions . . . . . 28
  - Overfitting . . . . . 30
- 2.6 Building a Predictive Model with XLMiner . . . . . 32
  - Predicting Home Values in the West Roxbury Neighborhood Modeling Process . . . . . 34
- 2.7 Using Excel for Data Mining . . . . . 41
- 2.8 Automating Data Mining Solutions . . . . . 42
  - Data Mining Software Tools: the State of the Market (by Herb Edelstein) . . . . . 43
- Problems . . . . . 47

---

**PART II DATA EXPLORATION AND DIMENSION REDUCTION**

**CHAPTER 3 Data Visualization** **52**

- 3.1 Uses of Data Visualization . . . . . 52
- 3.2 Data Examples . . . . . 54
  - Example 1: Boston Housing Data . . . . . 54
  - Example 2: Ridership on Amtrak Trains . . . . . 55
- 3.3 Basic Charts: Bar Charts, Line Graphs, and Scatter Plots . . . . . 55
  - Distribution Plots: Boxplots and Histograms . . . . . 57
  - Heatmaps: Visualizing Correlations and Missing Values . . . . . 60
- 3.4 Multidimensional Visualization . . . . . 62
  - Adding Variables: Color, Size, Shape, Multiple Panels, and Animation . . . . . 62
  - Manipulations: Re-scaling, Aggregation and Hierarchies, Zooming, Filtering . . . . . 64
  - Reference: Trend Line and Labels . . . . . 67
  - Scaling up to Large Datasets . . . . . 67
  - Multivariate Plot: Parallel Coordinates Plot . . . . . 69
  - Interactive Visualization . . . . . 70
- 3.5 Specialized Visualizations . . . . . 73
  - Visualizing Networked Data . . . . . 73
  - Visualizing Hierarchical Data: Treemaps . . . . . 75
  - Visualizing Geographical Data: Map Charts . . . . . 76
- 3.6 Summary: Major Visualizations and Operations, by Data Mining Goal . . . . . 78
  - Prediction . . . . . 78
  - Classification . . . . . 78
  - Time Series Forecasting . . . . . 78
  - Unsupervised Learning . . . . . 79
- Problems . . . . . 80

<b>CHAPTER 4 Dimension Reduction</b>	<b>82</b>
4.1 Introduction . . . . .	82
4.2 Curse of Dimensionality . . . . .	83
4.3 Practical Considerations . . . . .	83
Example 1: House Prices in Boston . . . . .	84
4.4 Data Summaries . . . . .	84
Summary Statistics . . . . .	86
Pivot Tables . . . . .	87
4.5 Correlation Analysis . . . . .	88
4.6 Reducing the Number of Categories in Categorical Variables . . . . .	89
4.7 Converting a Categorical Variable to a Numerical Variable .	90
4.8 Principal Components Analysis . . . . .	90
Example 2: Breakfast Cereals . . . . .	91
Principal Components . . . . .	95
Normalizing the Data . . . . .	97
Using Principal Components for Classification and Prediction	100
4.9 Dimension Reduction Using Regression Models . . . . .	100
4.10 Dimension Reduction Using Classification and Regression Trees . . . . .	101
Problems . . . . .	102

## **PART III PERFORMANCE EVALUATION**

---

<b>CHAPTER 5 Evaluating Predictive Performance</b>	<b>106</b>
5.1 Introduction . . . . .	107
5.2 Evaluating Predictive Performance . . . . .	107
Benchmark: The Average . . . . .	108
Prediction Accuracy Measures . . . . .	108
Comparing Training and Validation Performance . . . . .	109
Lift Chart . . . . .	110
5.3 Judging Classifier Performance . . . . .	112
Benchmark: The Naive Rule . . . . .	112
Class Separation . . . . .	112
The Classification Matrix . . . . .	113
Using the Validation Data . . . . .	114
Accuracy Measures . . . . .	115
Propensities and Cutoff for Classification . . . . .	115
Performance in Unequal Importance of Classes . . . . .	119
Asymmetric Misclassification Costs . . . . .	121
Generalization to More Than Two Classes . . . . .	124
5.4 Judging Ranking Performance . . . . .	124
Lift Charts for Binary Data . . . . .	125

Decile Lift Charts . . . . .	127
Beyond Two Classes . . . . .	127
Lift Charts Incorporating Costs and Benefits . . . . .	128
Lift as Function of Cutoff . . . . .	129
5.5 Oversampling . . . . .	129
Oversampling the Training Set . . . . .	132
Evaluating Model Performance Using a Non-oversampled Validation Set . . . . .	132
Evaluating Model Performance If Only Oversampled Validation Set Exists . . . . .	132
Problems . . . . .	135

**PART IV PREDICTION AND CLASSIFICATION METHODS**

**CHAPTER 6 Multiple Linear Regression 140**

6.1 Introduction . . . . .	140
6.2 Explanatory vs. Predictive Modeling . . . . .	141
6.3 Estimating the Regression Equation and Prediction . . . . .	143
Example: Predicting the Price of Used Toyota Corolla Cars . . . . .	144
6.4 Variable Selection in Linear Regression . . . . .	147
Reducing the Number of Predictors . . . . .	147
How to Reduce the Number of Predictors . . . . .	148
Problems . . . . .	153

**CHAPTER 7 *k*-Nearest-Neighbors (*k*-NN) 157**

7.1 The <i>k</i> -NN Classifier ( <i>categorical outcome</i> ) . . . . .	157
Determining Neighbors . . . . .	158
Classification Rule . . . . .	158
Example: Riding Mowers . . . . .	159
Choosing <i>k</i> . . . . .	160
Setting the Cutoff Value . . . . .	161
<i>k</i> -NN with More Than Two Classes . . . . .	162
Converting Categorical Variables to Binary Dummies . . . . .	163
7.2 <i>k</i> -NN for a Numerical Response . . . . .	163
7.3 Advantages and Shortcomings of <i>k</i> -NN Algorithms . . . . .	165
Problems . . . . .	167

**CHAPTER 8 The Naive Bayes Classifier 169**

8.1 Introduction . . . . .	169
Cutoff Probability Method . . . . .	170
Conditional Probability . . . . .	170
Example 1: Predicting Fraudulent Financial Reporting . . . . .	170
8.2 Applying the Full (Exact) Bayesian Classifier . . . . .	171
Using the “Assign to the Most Probable Class” Method . . . . .	172

Using the Cutoff Probability Method . . . . .	172
Practical Difficulty with the Complete (Exact) Bayes Procedure	172
Solution: Naive Bayes . . . . .	173
Example 2: Predicting Fraudulent Financial Reports, Two Predictors . . . . .	175
Example 3: Predicting Delayed Flights . . . . .	176
8.3 Advantages and Shortcomings of the Naive Bayes Classifier . . . . .	180
Problems . . . . .	184
<b>CHAPTER 9 Classification and Regression Trees</b>	<b>186</b>
9.1 Introduction . . . . .	187
9.2 Classification Trees . . . . .	188
Recursive Partitioning . . . . .	188
Example 1: Riding Mowers . . . . .	189
Measures of Impurity . . . . .	191
Tree Structure . . . . .	195
Classifying a New Observation . . . . .	195
9.3 Evaluating the Performance of a Classification Tree . . . . .	196
Example 2: Acceptance of Personal Loan . . . . .	196
9.4 Avoiding Overfitting . . . . .	199
Stopping Tree Growth: CHAID . . . . .	199
Pruning the Tree . . . . .	201
9.5 Classification Rules from Trees . . . . .	206
9.6 Classification Trees for More Than two Classes . . . . .	207
9.7 Regression Trees . . . . .	207
Prediction . . . . .	207
Measuring Impurity . . . . .	208
Evaluating Performance . . . . .	208
9.8 Advantages, Weaknesses and Extensions . . . . .	209
9.9 Improving Prediction: Multiple Trees . . . . .	210
Problems . . . . .	214
<b>CHAPTER 10 Logistic Regression</b>	<b>218</b>
10.1 Introduction . . . . .	219
10.2 The Logistic Regression Model . . . . .	220
Example: Acceptance of Personal Loan . . . . .	222
Model with a Single Predictor . . . . .	223
Estimating the Logistic Model from Data: Computing Parameter Estimates . . . . .	225
Interpreting Results in Terms of Odds (for a Profiling Goal)	227
10.3 Evaluating Classification Performance . . . . .	229
Variable Selection . . . . .	231

- 10.4 Example of Complete Analysis: Predicting
  - Delayed Flights . . . . . 231
  - Data Preprocessing . . . . . 234
  - Model Fitting and Estimation . . . . . 234
  - Model Interpretation . . . . . 235
  - Model Performance . . . . . 236
  - Variable Selection . . . . . 237
- 10.5 Appendix: Logistic Regression for Profiling . . . . . 240
  - Appendix A: Why Linear Regression Is Problematic for a Categorical Response . . . . . 240
  - Appendix B: Evaluating Explanatory Power . . . . . 241
  - Appendix C: Logistic Regression for More Than Two Classes Problems . . . . . 247

**CHAPTER 11 Neural Nets** 250

- 11.1 Introduction . . . . . 250
- 11.2 Concept and Structure of a Neural Network . . . . . 251
- 11.3 Fitting a Network to Data . . . . . 252
  - Example 1: Tiny Dataset . . . . . 252
  - Computing Output of Nodes . . . . . 253
  - Preprocessing the Data . . . . . 256
  - Training the Model . . . . . 257
  - Example 2: Classifying Accident Severity . . . . . 261
  - Avoiding Overfitting . . . . . 263
  - Using the Output for Prediction and Classification . . . . . 264
- 11.4 Required User Input . . . . . 266
- 11.5 Exploring the Relationship Between Predictors and Response . . . . . 268
  - Unsupervised Feature Extraction and Deep Learning . . . . . 270
- 11.6 Advantages and Weaknesses of Neural Networks . . . . . 268
- Problems . . . . . 271

**CHAPTER 12 Discriminant Analysis** 273

- 12.1 Introduction . . . . . 273
  - Example 1: Riding Mowers . . . . . 274
  - Example 2: Personal Loan Acceptance . . . . . 275
- 12.2 Distance of an Observation from a Class . . . . . 275
- 12.3 Fisher’s Linear Classification Functions . . . . . 278
- 12.4 Classification Performance of Discriminant Analysis . . . . . 281
- 12.5 Prior Probabilities . . . . . 282
- 12.6 Unequal Misclassification Costs . . . . . 283
- 12.7 Classifying More Than Two Classes . . . . . 284

Example 3: Medical Dispatch to Accident Scenes . . . . .	284
12.8 Advantages and Weaknesses . . . . .	286
Problems . . . . .	289

**CHAPTER 13 Combining Methods: Ensembles and Uplift Modeling** 292

13.1 Ensembles . . . . .	293
Why Ensembles Can Improve Predictive Power . . . . .	293
Simple Averaging . . . . .	295
Bagging . . . . .	296
Boosting . . . . .	296
Advantages and Weaknesses of Ensembles . . . . .	297
13.2 Uplift (Persuasion) Modeling . . . . .	297
A-B Testing . . . . .	298
Uplift . . . . .	298
Gathering the Data . . . . .	299
A Simple Model . . . . .	301
Modeling Individual Uplift . . . . .	301
Using the Results of an Uplift Model . . . . .	303
13.3 Summary . . . . .	303
Problems . . . . .	304

**PART V MINING RELATIONSHIPS AMONG RECORDS**

**CHAPTER 14 Association Rules and Collaborative Filtering** 308

14.1 Association Rules . . . . .	309
Discovering Association Rules in Transaction Databases . . . . .	309
Example 1: Synthetic Data on Purchases of Phone Faceplates . . . . .	309
Generating Candidate Rules . . . . .	311
The Apriori Algorithm . . . . .	312
Selecting Strong Rules . . . . .	312
Data Format . . . . .	314
The Process of Rule Selection . . . . .	315
Interpreting the Results . . . . .	317
Rules and Chance . . . . .	318
Example 2: Rules for Similar Book Purchases . . . . .	320
14.2 Collaborative Filtering . . . . .	322
Data Type and Format . . . . .	322
Example 3: Netflix Prize Contest . . . . .	323
User-Based Collaborative Filtering: “People Like You” . . . . .	324
Item-Based Collaborative Filtering . . . . .	327
Advantages and Weaknesses of Collaborative Filtering . . . . .	328
Collaborative Filtering vs. Association Rules . . . . .	328
14.3 Summary . . . . .	330
Problems . . . . .	332

<b>CHAPTER 15 Cluster Analysis</b>	<b>336</b>
15.1 Introduction . . . . .	337
Example: Public Utilities . . . . .	338
15.2 Measuring Distance Between Two Observations . . . . .	340
Euclidean Distance . . . . .	340
Normalizing Numerical Measurements . . . . .	341
Other Distance Measures for Numerical Data . . . . .	341
Distance Measures for Categorical Data . . . . .	343
Distance Measures for Mixed Data . . . . .	344
15.3 Measuring Distance Between Two Clusters . . . . .	345
Minimum Distance . . . . .	345
Maximum Distance . . . . .	345
Average Distance . . . . .	345
Centroid Distance . . . . .	345
15.4 Hierarchical (Agglomerative) Clustering . . . . .	347
Single Linkage . . . . .	348
Complete Linkage . . . . .	348
Average Linkage (in XLMiner: “Group Average Linkage”) . . . . .	349
Centroid Linkage . . . . .	349
Ward’s Method . . . . .	349
Dendrograms: Displaying Clustering Process and Results . . . . .	350
Validating Clusters . . . . .	352
Limitations of Hierarchical Clustering . . . . .	353
15.5 Non-hierarchical Clustering: The <i>k</i> -Means Algorithm . . . . .	354
Initial Partition into <i>k</i> Clusters . . . . .	356
Problems . . . . .	360

**PART VI FORECASTING TIME SERIES**

---

<b>CHAPTER 16 Handling Time Series</b>	<b>364</b>
16.1 Introduction . . . . .	364
16.2 Descriptive vs. Predictive Modeling . . . . .	366
16.3 Popular Forecasting Methods in Business . . . . .	366
Combining Methods . . . . .	366
16.4 Time Series Components . . . . .	367
Example: Ridership on Amtrak Trains . . . . .	367
16.5 Data Partitioning and Performance Evaluation . . . . .	371
Benchmark Performance: Naive Forecasts . . . . .	372
Generating Future Forecasts . . . . .	372
Problems . . . . .	374



<b>CHAPTER 17 Regression-Based Forecasting</b>	<b>377</b>
17.1 A Model with Trend . . . . .	377
Linear Trend . . . . .	377
Exponential Trend . . . . .	379
Polynomial Trend . . . . .	382
17.2 A Model with Seasonality . . . . .	383
17.3 A Model with Trend and Seasonality . . . . .	384
17.4 Autocorrelation and ARIMA Models . . . . .	387
Computing Autocorrelation . . . . .	387
Improving Forecasts by Integrating Autocorrelation Information . . . . .	389
Evaluating Predictability . . . . .	393
Problems . . . . .	396
<b>CHAPTER 18 Smoothing Methods</b>	<b>407</b>
18.1 Introduction . . . . .	407
18.2 Moving Average . . . . .	408
Centered Moving Average for Visualization . . . . .	408
Trailing Moving Average for Forecasting . . . . .	409
Choosing Window Width ( $w$ ) . . . . .	411
18.3 Simple Exponential Smoothing . . . . .	414
Choosing Smoothing Parameter $\alpha$ . . . . .	415
Relation between Moving Average and Simple Exponential Smoothing . . . . .	415
18.4 Advanced Exponential Smoothing . . . . .	416
Series with a Trend . . . . .	417
Series with a Trend and Seasonality . . . . .	417
Series with Seasonality (No Trend) . . . . .	418
Problems . . . . .	420
<b>PART VII DATA ANALYTICS</b>	
<b>CHAPTER 19 Social Network Analytics</b>	<b>430</b>
19.1 Introduction . . . . .	430
19.2 Directed vs. Undirected Networks . . . . .	431
19.3 Visualizing and Analyzing Networks . . . . .	432
Graph Layout . . . . .	434
Adjacency List . . . . .	435
Adjacency Matrix . . . . .	436
Using Network Data in Classification and Prediction . . . . .	436
19.4 Social Data Metrics and Taxonomy . . . . .	437
Node-Level Centrality Metrics . . . . .	438
Egocentric Network . . . . .	438

Network Metrics . . . . .	439
19.5 Using Network Metrics in Prediction and Classification . . .	440
Link Prediction . . . . .	441
Entity Resolution . . . . .	442
Collaborative Filtering . . . . .	444
19.6 Advantages and Disadvantages . . . . .	446
Problems . . . . .	448

**CHAPTER 20 Text Mining 450**

20.1 Introduction . . . . .	450
20.2 The Spreadsheet Representation of Text: “Bag-of-Words” . .	451
20.3 Bag-of-Words vs. Meaning Extraction at Document Level . .	452
20.4 Preprocessing the Text . . . . .	453
Tokenization . . . . .	453
Text Reduction . . . . .	454
Presence/Absence vs. Frequency . . . . .	454
Term Frequency—Inverse Document Frequency (TF-IDF) . .	455
From Terms to Concepts: Latent Semantic Indexing . . . .	455
Extracting Meaning . . . . .	456
20.5 Implementing Data Mining Methods . . . . .	456
20.6 Example: Online Discussions on Autos and Electronics . . .	457
Importing and Labeling the Records . . . . .	457
Tokenization . . . . .	458
Text Processing and Reduction . . . . .	459
Producing a Concept Matrix . . . . .	460
Labeling the Documents . . . . .	461
Fitting a Model . . . . .	462
Prediction . . . . .	462
20.7 Summary . . . . .	462
Problems . . . . .	464

**PART VIII CASES**

---

**CHAPTER 21 Cases 468**

21.1 Charles Book Club . . . . .	468
The Book Industry . . . . .	468
Database Marketing at Charles . . . . .	469
Data Mining Techniques . . . . .	472
Assignment . . . . .	473
21.2 German Credit . . . . .	477
Background . . . . .	477
Data . . . . .	477
Assignment . . . . .	481
21.3 Tayko Software Cataloger . . . . .	482

Background . . . . .	482
The Mailing Experiment . . . . .	482
Data . . . . .	482
Assignment . . . . .	484
21.4 Political Persuasion . . . . .	486
Background . . . . .	486
Predictive Analytics Arrives in US Politics . . . . .	486
Political Targeting . . . . .	486
Uplift . . . . .	487
Data . . . . .	488
Assignment . . . . .	488
21.5 Taxi Cancellations . . . . .	490
Business Situation . . . . .	490
Assignment . . . . .	490
21.6 Segmenting Consumers of Bath Soap . . . . .	491
Business Situation . . . . .	491
Key Problems . . . . .	492
Data . . . . .	492
Measuring Brand Loyalty . . . . .	492
Assignment . . . . .	494
Appendix . . . . .	494
21.7 Direct-Mail Fundraising . . . . .	495
Background . . . . .	495
Data . . . . .	495
Assignment . . . . .	495
21.8 Catalog Cross-Selling . . . . .	497
Background . . . . .	497
Assignment . . . . .	498
21.9 Predicting Bankruptcy . . . . .	499
Predicting Corporate Bankruptcy . . . . .	499
Assignment . . . . .	500
21.10 Time Series Case: Forecasting Public Transportation Demand	502
Background . . . . .	502
Problem Description . . . . .	502
Available Data . . . . .	502
Assignment Goal . . . . .	502
Assignment . . . . .	503
Tips and Suggested Steps . . . . .	503
<b>References</b>	<b>504</b>
<b>Data Files Used in the Book</b>	<b>506</b>
<b>Index</b>	<b>508</b>

