

Foreword by Gareth James	xix
Foreword by Ravi Bapna	xxix
Preface to the R Edition	xxiii
Acknowledgments	xxvii
PART I PRELIMINARIES	
CHAPTER 1 Introduction	3
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
CHAPTER 2 Overview of the Data Mining Process	15
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

	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

	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

	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

CHAPTER 5	Evaluating Predictive Performance	117
5.1	Introduction	117
5.2	Evaluating Predictive Performance	118
	Naive 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 Naive 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 k -NN Classifier (Categorical Outcome)	173
Determining Neighbors	173
Classification Rule	174
Example: Riding Mowers	175
Choosing k	176
Setting the Cutoff Value	179
k -NN with More Than Two Classes	180
Converting Categorical Variables to Binary Dummies	180
7.2 k -NN for a Numerical Outcome	180
7.3 Advantages and Shortcomings of k -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

CHAPTER 9 Classification and Regression Trees	205
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
CHAPTER 10 Logistic Regression	237
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	287
11.6 Advantages and Weaknesses of Neural Networks	288
Problems	290
CHAPTER 12 Discriminant Analysis	293
12.1 Introduction	293
Example 1: Riding Mowers	294
Example 2: Personal Loan Acceptance	294
12.2 Distance of a Record from a Class	296
12.3 Fisher's Linear Classification Functions	297
12.4 Classification Performance of Discriminant Analysis	300
12.5 Prior Probabilities	302
12.6 Unequal Misclassification Costs	302
12.7 Classifying More Than Two Classes	303
Example 3: Medical Dispatch to Accident Scenes	303
12.8 Advantages and Weaknesses	306
Problems	307
CHAPTER 13 Combining Methods: Ensembles and Uplift Modeling	311
13.1 Ensembles	311
Why Ensembles Can Improve Predictive Power	312
Simple Averaging	314
Bagging	315
Boosting	315
Bagging and Boosting in R	315
Advantages and Weaknesses of Ensembles	315
13.2 Uplift (Persuasion) Modeling	317
A-B Testing	318

Uplift 318

Gathering the Data 319

13.1 A Simple Model 320

13.2 Modeling Individual Uplift 321

Computing Uplift with R 322

Using the Results of an Uplift Model 322

13.3 Summary 324

Problems 325

PART V MINING RELATIONSHIPS AMONG RECORDS

CHAPTER 14 Association Rules and Collaborative Filtering 329

14.1 Association Rules 329

Discovering Association Rules in Transaction Databases 330

Example 1: Synthetic Data on Purchases of Phone Faceplates 330

Generating Candidate Rules 330

The Apriori Algorithm 333

Selecting Strong Rules 333

Data Format 335

The Process of Rule Selection 336

Interpreting the Results 337

Rules and Chance 339

Example 2: Rules for Similar Book Purchases 340

14.2 Collaborative Filtering 342

Data Type and Format 343

Example 3: Netflix Prize Contest 343

User-Based Collaborative Filtering: “People Like You” 344

Item-Based Collaborative Filtering 347

Advantages and Weaknesses of Collaborative Filtering 348

Collaborative Filtering vs. Association Rules 349

14.3 Summary 351

Problems 352

CHAPTER 15 Cluster Analysis 357

15.1 Introduction 357

Example: Public Utilities 359

15.2 Measuring Distance Between Two Records 361

Euclidean Distance 361

Normalizing Numerical Measurements 362

Other Distance Measures for Numerical Data 362

Distance Measures for Categorical Data 365

Distance Measures for Mixed Data 366

15.3 Measuring Distance Between Two Clusters 366

Minimum Distance 366

Maximum Distance 366

	Average Distance	367
	Centroid Distance	367
15.4	Hierarchical (Agglomerative) Clustering	368
	Single Linkage	369
	Complete Linkage	370
	Average Linkage	370
	Centroid Linkage	370
	Ward's Method	370
	Dendrograms: Displaying Clustering Process and Results	371
	Validating Clusters	373
	Limitations of Hierarchical Clustering	375
15.5	Non-Hierarchical Clustering: The k -Means Algorithm	376
	Choosing the Number of Clusters (k)	377
	Problems	382

PART VI FORECASTING TIME SERIES

CHAPTER 16	Handling Time Series	387
16.1	Introduction	387
16.2	Descriptive vs. Predictive Modeling	389
16.3	Popular Forecasting Methods in Business	389
	Combining Methods	389
16.4	Time Series Components	390
	Example: Ridership on Amtrak Trains	390
16.5	Data-Partitioning and Performance Evaluation	395
	Benchmark Performance: Naive Forecasts	395
	Generating Future Forecasts	396
	Problems	398
CHAPTER 17	Regression-Based Forecasting	401
17.1	A Model with Trend	401
	Linear Trend	401
	Exponential Trend	405
	Polynomial Trend	407
17.2	A Model with Seasonality	407
17.3	A Model with Trend and Seasonality	411
17.4	Autocorrelation and ARIMA Models	412
	Computing Autocorrelation	413
	Improving Forecasts by Integrating Autocorrelation Information	416
	Evaluating Predictability	420
	Problems	422

CHAPTER 18 Smoothing Methods	433
18.1 Introduction	433
18.2 Moving Average	434
Centered Moving Average for Visualization	434
Trailing Moving Average for Forecasting	435
Choosing Window Width (w)	439
18.3 Simple Exponential Smoothing	439
Choosing Smoothing Parameter α	440
Relation Between Moving Average and Simple Exponential Smoothing	440
18.4 Advanced Exponential Smoothing	442
Series with a Trend	442
Series with a Trend and Seasonality	443
Series with Seasonality (No Trend)	443
Problems	446

PART VII DATA ANALYTICS

CHAPTER 19 Social Network Analytics	455
19.1 Introduction	455
19.2 Directed vs. Undirected Networks	457
19.3 Visualizing and Analyzing Networks	458
Graph Layout	458
Edge List	460
Adjacency Matrix	461
Using Network Data in Classification and Prediction	461
19.4 Social Data Metrics and Taxonomy	462
Node-Level Centrality Metrics	463
Egocentric Network	463
Network Metrics	465
19.5 Using Network Metrics in Prediction and Classification	467
Link Prediction	467
Entity Resolution	467
Collaborative Filtering	468
19.6 Collecting Social Network Data with R	471
19.7 Advantages and Disadvantages	474
Problems	476
CHAPTER 20 Text Mining	479
20.1 Introduction	479
20.2 The Tabular Representation of Text: Term-Document Matrix and “Bag-of-Words”	480
20.3 Bag-of-Words vs. Meaning Extraction at Document Level	481
20.4 Preprocessing the Text	482
Tokenization	484
Text Reduction	485

487	Presence/Absence vs. Frequency	487
487	Term Frequency–Inverse Document Frequency (TF-IDF)	487
488	From Terms to Concepts: Latent Semantic Indexing	488
489	Extracting Meaning	489
20.5	Implementing Data Mining Methods	489
20.6	Example: Online Discussions on Autos and Electronics	490
	Importing and Labeling the Records	490
	Text Preprocessing in R	491
	Producing a Concept Matrix	491
	Fitting a Predictive Model	492
	Prediction	492
20.7	Summary	494
	Problems	495

PART VIII CASES

CHAPTER 21 Cases 499

21.1	Charles Book Club	499
	The Book Industry	499
	Database Marketing at Charles	500
	Data Mining Techniques	502
	Assignment	504
21.2	German Credit	505
	Background	505
	Data	506
	Assignment	507
21.3	Tayko Software Cataloger	510
	Background	510
	The Mailing Experiment	510
	Data	510
	Assignment	512
21.4	Political Persuasion	513
	Background	513
	Predictive Analytics Arrives in US Politics	513
	Political Targeting	514
	Uplift	514
	Data	515
	Assignment	516
21.5	Taxi Cancellations	517
	Business Situation	517
	Assignment	517
21.6	Segmenting Consumers of Bath Soap	518
	Business Situation	518
	Key Problems	519
	Data	519

	Measuring Brand Loyalty	519
	Assignment	521
21.7	Direct-Mail Fundraising	521
	Background	521
	Data	522
	Assignment	523
21.8	Catalog Cross-Selling	524
	Background	524
	Assignment	524
21.9	Predicting Bankruptcy	525
	Predicting Corporate Bankruptcy	525
	Assignment	526
21.10	Time Series Case: Forecasting Public Transportation Demand	528
	Background	528
	Problem Description	528
	Available Data	528
	Assignment Goal	528
	Assignment	529
	Tips and Suggested Steps	529
	References	531
	Data Files Used in the Book	533
	Index	535