CONTENTS

PREFACE xxi
ACKNOWLEDGMENTS xxix

PART I

DATA PREPARATION 1

CHAPTER 1 *AN INTRODUCTION TO DATA MINING AND PREDICTIVE ANALYTICS* 3

1.1 What is Data Mining? What is Predictive Analytics? 3
1.2 Wanted: Data Miners 5
1.3 The Need for Human Direction of Data Mining 6
1.4 The Cross-Industry Standard Process for Data Mining: CRISP-DM 6
1.4.1 CRISP-DM: The Six Phases 7
1.5 Fallacies of Data Mining 9
1.6 What Tasks Can Data Mining Accomplish 10
1.6.1 Description 10
1.6.2 Estimation 11
1.6.3 Prediction 12
1.6.4 Classification 12
1.6.5 Clustering 15
1.6.6 Association 16
The R Zone 17
R References 18
Exercises 18

CHAPTER 2 *DATA PREPROCESSING* 20

2.1 Why do We Need to Preprocess the Data? 20
2.2 Data Cleaning 21
2.3 Handling Missing Data 22
2.4 Identifying Misclassifications 25
2.5 Graphical Methods for Identifying Outliers 26
2.6 Measures of Center and Spread 27
2.7 Data Transformation 30
2.8 Min–Max Normalization 30
2.9 Z-Score Standardization 31
2.10 Decimal Scaling 32
2.11 Transformations to Achieve Normality 32

2.12 Numerical Methods for Identifying Outliers 38
2.13 Flag Variables 39
2.14 Transforming Categorical Variables into Numerical Variables 40
2.15 Binning Numerical Variables 41
2.16 Reclassifying Categorical Variables 42
2.17 Adding an Index Field 43
2.18 Removing Variables that are not Useful 43
2.19 Variables that Should Probably not be Removed 43
2.20 Removal of Duplicate Records 44
2.21 A Word About ID Fields 45
The R Zone 45
R Reference 51
Exercises 51

CHAPTER 3 *EXPLORATORY DATA ANALYSIS* 54

3.1 Hypothesis Testing Versus Exploratory Data Analysis 54
3.2 Getting to Know the Data Set 54
3.3 Exploring Categorical Variables 56
3.4 Exploring Numeric Variables 64
3.5 Exploring Multivariate Relationships 69
3.6 Selecting Interesting Subsets of the Data for Further Investigation 70
3.7 Using EDA to Uncover Anomalous Fields 71
3.8 Binning Based on Predictive Value 72
3.9 Deriving New Variables: Flag Variables 75
3.10 Deriving New Variables: Numerical Variables 77
3.11 Using EDA to Investigate Correlated Predictor Variables 78
3.12 Summary of Our EDA 81
The R Zone 82
R References 89
Exercises 89

CHAPTER 4 *DIMENSION-REDUCTION METHODS* 92

4.1 Need for Dimension-Reduction in Data Mining 92
4.2 Principal Components Analysis 93
4.3 Applying PCA to the *Houses* Data Set 96
4.4 How Many Components Should We Extract? 102
4.4.1 The Eigenvalue Criterion 102
4.4.2 The Proportion of Variance Explained Criterion 103
4.4.3 The Minimum Communality Criterion 103
4.4.4 The Scree Plot Criterion 103
4.5 Profiling the Principal Components 105
4.6 Communalities 108
4.6.1 Minimum Communality Criterion 109
4.7 Validation of the Principal Components 110
4.8 Factor Analysis 110
4.9 Applying Factor Analysis to the *Adult* Data Set 111
4.10 Factor Rotation 114
4.11 User-Defined Composites 117

4.12 An Example of a User-Defined Composite 118
The R Zone 119
R References 124
Exercises 124

PART II
STATISTICAL ANALYSIS 129

CHAPTER 5 *UNIVARIATE STATISTICAL ANALYSIS* 131

5.1 Data Mining Tasks in Discovering Knowledge in Data 131
5.2 Statistical Approaches to Estimation and Prediction 131
5.3 Statistical Inference 132
5.4 How Confident are We in Our Estimates? 133
5.5 Confidence Interval Estimation of the Mean 134
5.6 How to Reduce the Margin of Error 136
5.7 Confidence Interval Estimation of the Proportion 137
5.8 Hypothesis Testing for the Mean 138
5.9 Assessing the Strength of Evidence Against the Null Hypothesis 140
5.10 Using Confidence Intervals to Perform Hypothesis Tests 141
5.11 Hypothesis Testing for the Proportion 143
Reference 144
The R Zone 144
R Reference 145
Exercises 145

CHAPTER 6 *MULTIVARIATE STATISTICS* 148

6.1 Two-Sample t-Test for Difference in Means 148
6.2 Two-Sample Z-Test for Difference in Proportions 149
6.3 Test for the Homogeneity of Proportions 150
6.4 Chi-Square Test for Goodness of Fit of Multinomial Data 152
6.5 Analysis of Variance 153
Reference 156
The R Zone 157
R Reference 158
Exercises 158

CHAPTER 7 *PREPARING TO MODEL THE DATA* 160

7.1 Supervised Versus Unsupervised Methods 160
7.2 Statistical Methodology and Data Mining Methodology 161
7.3 Cross-Validation 161
7.4 Overfitting 163
7.5 Bias–Variance Trade-Off 164
7.6 Balancing the Training Data Set 166
7.7 Establishing Baseline Performance 167
The R Zone 168

R Reference 169
Exercises 169

CHAPTER 8 *SIMPLE LINEAR REGRESSION* 171

8.1 An Example of Simple Linear Regression 171
8.1.1 The Least-Squares Estimates 174
8.2 Dangers of Extrapolation 177
8.3 How Useful is the Regression? The Coefficient of Determination, r^2 178
8.4 Standard Error of the Estimate, s 183
8.5 Correlation Coefficient r 184
8.6 Anova Table for Simple Linear Regression 186
8.7 Outliers, High Leverage Points, and Influential Observations 186
8.8 Population Regression Equation 195
8.9 Verifying the Regression Assumptions 198
8.10 Inference in Regression 203
8.11 t-Test for the Relationship Between x and y 204
8.12 Confidence Interval for the Slope of the Regression Line 206
8.13 Confidence Interval for the Correlation Coefficient ρ 208
8.14 Confidence Interval for the Mean Value of y Given x 210
8.15 Prediction Interval for a Randomly Chosen Value of y Given x 211
8.16 Transformations to Achieve Linearity 213
8.17 Box–Cox Transformations 220
The R Zone 220
R References 227
Exercises 227

CHAPTER 9 *MULTIPLE REGRESSION AND MODEL BUILDING* 236

9.1 An Example of Multiple Regression 236
9.2 The Population Multiple Regression Equation 242
9.3 Inference in Multiple Regression 243
9.3.1 The t-Test for the Relationship Between y and x_i 243
9.3.2 t-Test for Relationship Between Nutritional *Rating* and *Sugars* 244
9.3.3 t-Test for Relationship Between Nutritional *Rating* and *Fiber* Content 244
9.3.4 The F-Test for the Significance of the Overall Regression Model 245
9.3.5 F-Test for Relationship between Nutritional Rating and {Sugar and Fiber}, Taken Together 247
9.3.6 The Confidence Interval for a Particular Coefficient, β_i 247
9.3.7 The Confidence Interval for the Mean Value of y, Given $x_1, x_2, \ldots, x_m$ 248
9.3.8 The Prediction Interval for a Randomly Chosen Value of y, Given $x_1, x_2, \ldots, x_m$ 248
9.4 Regression with Categorical Predictors, Using Indicator Variables 249
9.5 Adjusting R^2: Penalizing Models for Including Predictors that are not Useful 256
9.6 Sequential Sums of Squares 257
9.7 Multicollinearity 258
9.8 Variable Selection Methods 266
9.8.1 The Partial F-Test 266

9.8.2 The Forward Selection Procedure 268
9.8.3 The Backward Elimination Procedure 268
9.8.4 The Stepwise Procedure 268
9.8.5 The Best Subsets Procedure 269
9.8.6 The All-Possible-Subsets Procedure 269
9.9 Gas Mileage Data Set 270
9.10 An Application of Variable Selection Methods 271
9.10.1 Forward Selection Procedure Applied to the *Gas Mileage* Data Set 271
9.10.2 Backward Elimination Procedure Applied to the *Gas Mileage* Data Set 273
9.10.3 The Stepwise Selection Procedure Applied to the *Gas Mileage* Data Set 273
9.10.4 Best Subsets Procedure Applied to the *Gas Mileage* Data Set 274
9.10.5 Mallows' C_p Statistic 275
9.11 Using the Principal Components as Predictors in Multiple Regression 279
The R Zone 284
R References 292
Exercises 293

PART III

CLASSIFICATION 299

CHAPTER 10 *k-NEAREST NEIGHBOR ALGORITHM* 301

10.1 Classification Task 301
10.2 *k*-Nearest Neighbor Algorithm 302
10.3 Distance Function 305
10.4 Combination Function 307
10.4.1 Simple Unweighted Voting 307
10.4.2 Weighted Voting 308
10.5 Quantifying Attribute Relevance: Stretching the Axes 309
10.6 Database Considerations 310
10.7 *k*-Nearest Neighbor Algorithm for Estimation and Prediction 310
10.8 Choosing *k* 311
10.9 Application of *k*-Nearest Neighbor Algorithm Using IBM/SPSS Modeler 312
The R Zone 312
R References 315
Exercises 315

CHAPTER 11 *DECISION TREES* 317

11.1 What is a Decision Tree? 317
11.2 Requirements for Using Decision Trees 319
11.3 Classification and Regression Trees 319
11.4 C4.5 Algorithm 326
11.5 Decision Rules 332
11.6 Comparison of the C5.0 and CART Algorithms Applied to Real Data 332
The R Zone 335

R References 337
Exercises 337

CHAPTER 12 *NEURAL NETWORKS* 339

12.1 Input and Output Encoding 339
12.2 Neural Networks for Estimation and Prediction 342
12.3 Simple Example of a Neural Network 342
12.4 Sigmoid Activation Function 344
12.5 Back-Propagation 345
12.6 Gradient-Descent Method 346
12.7 Back-Propagation Rules 347
12.8 Example of Back-Propagation 347
12.9 Termination Criteria 349
12.10 Learning Rate 350
12.11 Momentum Term 351
12.12 Sensitivity Analysis 353
12.13 Application of Neural Network Modeling 353
The R Zone 356
R References 357
Exercises 357

CHAPTER 13 *LOGISTIC REGRESSION* 359

13.1 Simple Example of Logistic Regression 359
13.2 Maximum Likelihood Estimation 361
13.3 Interpreting Logistic Regression Output 362
13.4 Inference: are the Predictors Significant? 363
13.5 Odds Ratio and Relative Risk 365
13.6 Interpreting Logistic Regression for a Dichotomous Predictor 367
13.7 Interpreting Logistic Regression for a Polychotomous Predictor 370
13.8 Interpreting Logistic Regression for a Continuous Predictor 374
13.9 Assumption of Linearity 378
13.10 Zero-Cell Problem 382
13.11 Multiple Logistic Regression 384
13.12 Introducing Higher Order Terms to Handle Nonlinearity 388
13.13 Validating the Logistic Regression Model 395
13.14 WEKA: Hands-On Analysis Using Logistic Regression 399
The R Zone 404
R References 409
Exercises 409

CHAPTER 14 *NAÏVE BAYES AND BAYESIAN NETWORKS* 414

14.1 Bayesian Approach 414
14.2 Maximum a Posteriori (Map) Classification 416
14.3 Posterior Odds Ratio 420

14.4 Balancing the Data 422
14.5 Naïve Bayes Classification 423
14.6 Interpreting the Log Posterior Odds Ratio 426
14.7 Zero-Cell Problem 428
14.8 Numeric Predictors for Naïve Bayes Classification 429
14.9 WEKA: Hands-on Analysis Using Naïve Bayes 432
14.10 Bayesian Belief Networks 436
14.11 Clothing Purchase Example 436
14.12 Using the Bayesian Network to Find Probabilities 439
14.12.1 WEKA: Hands-on Analysis Using Bayes Net 441
The R Zone 444
R References 448
Exercises 448

CHAPTER 15 *MODEL EVALUATION TECHNIQUES* 451

15.1 Model Evaluation Techniques for the Description Task 451
15.2 Model Evaluation Techniques for the Estimation and Prediction Tasks 452
15.3 Model Evaluation Measures for the Classification Task 454
15.4 Accuracy and Overall Error Rate 456
15.5 Sensitivity and Specificity 457
15.6 False-Positive Rate and False-Negative Rate 458
15.7 Proportions of True Positives, True Negatives, False Positives, and False Negatives 458
15.8 Misclassification Cost Adjustment to Reflect Real-World Concerns 460
15.9 Decision Cost/Benefit Analysis 462
15.10 Lift Charts and Gains Charts 463
15.11 Interweaving Model Evaluation with Model Building 466
15.12 Confluence of Results: Applying a Suite of Models 466
The R Zone 467
R References 468
Exercises 468

CHAPTER 16 *COST-BENEFIT ANALYSIS USING DATA-DRIVEN COSTS* 471

16.1 Decision Invariance Under Row Adjustment 471
16.2 Positive Classification Criterion 473
16.3 Demonstration of the Positive Classification Criterion 474
16.4 Constructing the Cost Matrix 474
16.5 Decision Invariance Under Scaling 476
16.6 Direct Costs and Opportunity Costs 478
16.7 Case Study: Cost-Benefit Analysis Using Data-Driven Misclassification Costs 478
16.8 Rebalancing as a Surrogate for Misclassification Costs 483
The R Zone 485
R References 487
Exercises 487

CHAPTER 17 *COST-BENEFIT ANALYSIS FOR TRINARY AND k-NARY CLASSIFICATION MODELS* 491

17.1 Classification Evaluation Measures for a Generic Trinary Target 491
17.2 Application of Evaluation Measures for Trinary Classification to the Loan Approval Problem 494
17.3 Data-Driven Cost-Benefit Analysis for Trinary Loan Classification Problem 498
17.4 Comparing Cart Models with and without Data-Driven Misclassification Costs 500
17.5 Classification Evaluation Measures for a Generic *k*-Nary Target 503
17.6 Example of Evaluation Measures and Data-Driven Misclassification Costs for *k*-Nary Classification 504
The R Zone 507
R References 508
Exercises 508

CHAPTER 18 *GRAPHICAL EVALUATION OF CLASSIFICATION MODELS* 510

18.1 Review of Lift Charts and Gains Charts 510
18.2 Lift Charts and Gains Charts Using Misclassification Costs 510
18.3 Response Charts 511
18.4 Profits Charts 512
18.5 Return on Investment (ROI) Charts 514
The R Zone 516
R References 517
Exercises 518

PART IV

CLUSTERING 521

CHAPTER 19 *HIERARCHICAL AND k-MEANS CLUSTERING* 523

19.1 The Clustering Task 523
19.2 Hierarchical Clustering Methods 525
19.3 Single-Linkage Clustering 526
19.4 Complete-Linkage Clustering 527
19.5 *k*-Means Clustering 529
19.6 Example of *k*-Means Clustering at Work 530
19.7 Behavior of MSB, MSE, and Pseudo-*F* as the *k*-Means Algorithm Proceeds 533
19.8 Application of *k*-Means Clustering Using SAS Enterprise Miner 534
19.9 Using Cluster Membership to Predict Churn 537
The R Zone 538
R References 540
Exercises 540

CHAPTER 20 *KOHONEN NETWORKS* 542

20.1 Self-Organizing Maps 542
20.2 Kohonen Networks 544
20.3 Example of a Kohonen Network Study 545
20.4 Cluster Validity 549
20.5 Application of Clustering Using Kohonen Networks 549
20.6 Interpreting The Clusters 551
20.6.1 Cluster Profiles 554
20.7 Using Cluster Membership as Input to Downstream Data Mining Models 556
The R Zone 557
R References 558
Exercises 558

CHAPTER 21 *BIRCH CLUSTERING* 560

21.1 Rationale for Birch Clustering 560
21.2 Cluster Features 561
21.3 Cluster Feature Tree 562
21.4 Phase 1: Building the CF Tree 562
21.5 Phase 2: Clustering the Sub-Clusters 564
21.6 Example of Birch Clustering, Phase 1: Building the CF Tree 565
21.7 Example of Birch Clustering, Phase 2: Clustering the Sub-Clusters 570
21.8 Evaluating the Candidate Cluster Solutions 571
21.9 Case Study: Applying Birch Clustering to the Bank Loans Data Set 571
21.9.1 Case Study Lesson One: Avoid Highly Correlated Inputs to Any Clustering Algorithm 572
21.9.2 Case Study Lesson Two: Different Sortings May Lead to Different Numbers of Clusters 577
The R Zone 579
R References 580
Exercises 580

CHAPTER 22 *MEASURING CLUSTER GOODNESS* 582

22.1 Rationale for Measuring Cluster Goodness 582
22.2 The Silhouette Method 583
22.3 Silhouette Example 584
22.4 Silhouette Analysis of the *IRIS* Data Set 585
22.5 The Pseudo-*F* Statistic 590
22.6 Example of the Pseudo-*F* Statistic 591
22.7 Pseudo-*F* Statistic Applied to the *IRIS* Data Set 592
22.8 Cluster Validation 593
22.9 Cluster Validation Applied to the Loans Data Set 594
The R Zone 597
R References 599
Exercises 599

PART V

ASSOCIATION RULES 601

CHAPTER 23 *ASSOCIATION RULES* 603

23.1 Affinity Analysis and Market Basket Analysis 603
23.1.1 Data Representation for Market Basket Analysis 604
23.2 Support, Confidence, Frequent Itemsets, and the a Priori Property 605
23.3 How Does the a Priori Algorithm Work (Part 1)? Generating Frequent Itemsets 607
23.4 How Does the a Priori Algorithm Work (Part 2)? Generating Association Rules 608
23.5 Extension from Flag Data to General Categorical Data 611
23.6 Information-Theoretic Approach: Generalized Rule Induction Method 612
23.6.1 *J*-Measure 613
23.7 Association Rules are Easy to do Badly 614
23.8 How can we Measure the Usefulness of Association Rules? 615
23.9 Do Association Rules Represent Supervised or Unsupervised Learning? 616
23.10 Local Patterns Versus Global Models 617
The R Zone 618
R References 618
Exercises 619

PART VI

ENHANCING MODEL PERFORMANCE 623

CHAPTER 24 *SEGMENTATION MODELS* 625

24.1 The Segmentation Modeling Process 625
24.2 Segmentation Modeling Using EDA to Identify the Segments 627
24.3 Segmentation Modeling using Clustering to Identify the Segments 629
The R Zone 634
R References 635
Exercises 635

CHAPTER 25 *ENSEMBLE METHODS: BAGGING AND BOOSTING* 637

25.1 Rationale for Using an Ensemble of Classification Models 637
25.2 Bias, Variance, and Noise 639
25.3 When to Apply, and not to apply, Bagging 640
25.4 Bagging 641
25.5 Boosting 643
25.6 Application of Bagging and Boosting Using IBM/SPSS Modeler 647
References 648
The R Zone 649
R Reference 650
Exercises 650

CHAPTER 26 *MODEL VOTING AND PROPENSITY AVERAGING* 653

26.1 Simple Model Voting 653
26.2 Alternative Voting Methods 654
26.3 Model Voting Process 655
26.4 An Application of Model Voting 656
26.5 What is Propensity Averaging? 660
26.6 Propensity Averaging Process 661
26.7 An Application of Propensity Averaging 661
The R Zone 665
R References 666
Exercises 666

PART VII

FURTHER TOPICS 669

CHAPTER 27 *GENETIC ALGORITHMS* 671

27.1 Introduction To Genetic Algorithms 671
27.2 Basic Framework of a Genetic Algorithm 672
27.3 Simple Example of a Genetic Algorithm at Work 673
27.3.1 First Iteration 674
27.3.2 Second Iteration 675
27.4 Modifications and Enhancements: Selection 676
27.5 Modifications and Enhancements: Crossover 678
27.5.1 Multi-Point Crossover 678
27.5.2 Uniform Crossover 678
27.6 Genetic Algorithms for Real-Valued Variables 679
27.6.1 Single Arithmetic Crossover 680
27.6.2 Simple Arithmetic Crossover 680
27.6.3 Whole Arithmetic Crossover 680
27.6.4 Discrete Crossover 681
27.6.5 Normally Distributed Mutation 681
27.7 Using Genetic Algorithms to Train a Neural Network 681
27.8 WEKA: Hands-On Analysis Using Genetic Algorithms 684
The R Zone 692
R References 693
Exercises 693

CHAPTER 28 *IMPUTATION OF MISSING DATA* 695

28.1 Need for Imputation of Missing Data 695
28.2 Imputation of Missing Data: Continuous Variables 696
28.3 Standard Error of the Imputation 699
28.4 Imputation of Missing Data: Categorical Variables 700
28.5 Handling Patterns in Missingness 701
Reference 701
The R Zone 702

R References 704
Exercises 704

PART VIII

CASE STUDY: PREDICTING RESPONSE TO DIRECT-MAIL MARKETING 705

CHAPTER 29 *CASE STUDY, PART 1: BUSINESS UNDERSTANDING, DATA PREPARATION, AND EDA* 707

29.1 Cross-Industry Standard Practice for Data Mining 707
29.2 Business Understanding Phase 709
29.3 Data Understanding Phase, Part 1: Getting a Feel for the Data Set 710
29.4 Data Preparation Phase 714
29.4.1 Negative Amounts Spent? 714
29.4.2 Transformations to Achieve Normality or Symmetry 716
29.4.3 Standardization 717
29.4.4 Deriving New Variables 719
29.5 Data Understanding Phase, Part 2: Exploratory Data Analysis 721
29.5.1 Exploring the Relationships between the Predictors and the Response 722
29.5.2 Investigating the Correlation Structure among the Predictors 727
29.5.3 Importance of De-Transforming for Interpretation 730

CHAPTER 30 *CASE STUDY, PART 2: CLUSTERING AND PRINCIPAL COMPONENTS ANALYSIS* 732

30.1 Partitioning the Data 732
30.1.1 Validating the Partition 732
30.2 Developing the Principal Components 733
30.3 Validating the Principal Components 737
30.4 Profiling the Principal Components 737
30.5 Choosing the Optimal Number of Clusters Using Birch Clustering 742
30.6 Choosing the Optimal Number of Clusters Using k-Means Clustering 744
30.7 Application of k-Means Clustering 745
30.8 Validating the Clusters 745
30.9 Profiling the Clusters 745

CHAPTER 31 *CASE STUDY, PART 3: MODELING AND EVALUATION FOR PERFORMANCE AND INTERPRETABILITY* 749

31.1 Do you Prefer the Best Model Performance, or a Combination of Performance and Interpretability? 749
31.2 Modeling and Evaluation Overview 750
31.3 Cost-Benefit Analysis Using Data-Driven Costs 751
31.3.1 Calculating Direct Costs 752
31.4 Variables to be Input to the Models 753

31.5 Establishing the Baseline Model Performance 754
31.6 Models that use Misclassification Costs 755
31.7 Models that Need Rebalancing as a Surrogate for Misclassification Costs 756
31.8 Combining Models Using Voting and Propensity Averaging 757
31.9 Interpreting the Most Profitable Model 758

CHAPTER 32 *CASE STUDY, PART 4: MODELING AND EVALUATION FOR HIGH PERFORMANCE ONLY* 762

32.1 Variables to be Input to the Models 762
32.2 Models that use Misclassification Costs 762
32.3 Models that Need Rebalancing as a Surrogate for Misclassification Costs 764
32.4 Combining Models using Voting and Propensity Averaging 765
32.5 Lessons Learned 766
32.6 Conclusions 766

APPENDIX A *DATA SUMMARIZATION AND VISUALIZATION* 768

Part 1: Summarization 1: Building Blocks of Data Analysis 768
Part 2: Visualization: Graphs and Tables for Summarizing and Organizing Data 770
Part 3: Summarization 2: Measures of Center, Variability, and Position 774
Part 4: Summarization and Visualization of Bivariate Relationships 777

INDEX 781