Contents

Introduction xxi

Chapter 1 **Overview of Predictive Analytics** 1
- What Is Analytics? 3
- What Is Predictive Analytics? 3
  - Supervised vs. Unsupervised Learning 5
  - Parametric vs. Non-Parametric Models 6
- Business Intelligence 6
- Predictive Analytics vs. Business Intelligence 8
  - Do Predictive Models Just State the Obvious? 9
  - Similarities between Business Intelligence and Predictive Analytics 9
- Predictive Analytics vs. Statistics 10
  - Statistics and Analytics 11
  - Predictive Analytics and Statistics Contrasted 12
- Predictive Analytics vs. Data Mining 13
- Who Uses Predictive Analytics? 13
- Challenges in Using Predictive Analytics 14
  - Obstacles in Management 14
  - Obstacles with Data 14
  - Obstacles with Modeling 15
  - Obstacles in Deployment 16
- What Educational Background Is Needed to Become a Predictive Modeler? 16

Chapter 2 **Setting Up the Problem** 19
- Predictive Analytics Processing Steps: CRISP-DM 19
- Business Understanding 21
  - The Three-Legged Stool 22
  - Business Objectives 23

Defining Data for Predictive Modeling 25
Defining the Columns as Measures 26
Defining the Unit of Analysis 27
Which Unit of Analysis? 28
Defining the Target Variable 29
Temporal Considerations for Target Variable 31
Defining Measures of Success for Predictive Models 32
Success Criteria for Classification 32
Success Criteria for Estimation 33
Other Customized Success Criteria 33
Doing Predictive Modeling Out of Order 34
Building Models First 34
Early Model Deployment 35
Case Study: Recovering Lapsed Donors 35
Overview 36
Business Objectives 36
Data for the Competition 36
The Target Variables 36
Modeling Objectives 37
Model Selection and Evaluation Criteria 38
Model Deployment 39
Case Study: Fraud Detection 39
Overview 39
Business Objectives 39
Data for the Project 40
The Target Variables 40
Modeling Objectives 41
Model Selection and Evaluation Criteria 41
Model Deployment 41
Summary 42

Chapter 3 Data Understanding 43
What the Data Looks Like 44
Single Variable Summaries 44
Mean 45
Standard Deviation 45
The Normal Distribution 45
Uniform Distribution 46
Applying Simple Statistics in Data Understanding 47
Skewness 49
Kurtosis 51
Rank-Ordered Statistics 52
Categorical Variable Assessment 55
Data Visualization in One Dimension 58
Histograms 59
Multiple Variable Summaries 64

Hidden Value in Variable Interactions: Simpson's Paradox 64
The Combinatorial Explosion of Interactions 65
Correlations 66
Spurious Correlations 66
Back to Correlations 67
Crosstabs 68
Data Visualization, Two or Higher Dimensions 69
Scatterplots 69
Anscombe's Quartet 71
Scatterplot Matrices 75
Overlaying the Target Variable in Summary 76
Scatterplots in More Than Two Dimensions 78
The Value of Statistical Significance 80
Pulling It All Together into a Data Audit 81
Summary 82

Chapter 4 Data Preparation 83
Variable Cleaning 84
Incorrect Values 84
Consistency in Data Formats 85
Outliers 85
Multidimensional Outliers 89
Missing Values 90
Fixing Missing Data 91
Feature Creation 98
Simple Variable Transformations 98
Fixing Skew 99
Binning Continuous Variables 103
Numeric Variable Scaling 104
Nominal Variable Transformation 107
Ordinal Variable Transformations 108
Date and Time Variable Features 109
ZIP Code Features 110
Which Version of a Variable Is Best? 110
Multidimensional Features 112
Variable Selection Prior to Modeling 117
Sampling 123
Example: Why Normalization Matters for K-Means Clustering 139
Summary 143

Chapter 5 Itemsets and Association Rules 145
Terminology 146
Condition 147
Left-Hand-Side, Antecedent(s) 148
Right-Hand-Side, Consequent, Output, Conclusion 148
Rule (Item Set) 148

Support 149
Antecedent Support 149
Confidence, Accuracy 150
Lift 150
Parameter Settings 151
How the Data Is Organized 151
Standard Predictive Modeling Data Format 151
Transactional Format 152
Measures of Interesting Rules 154
Deploying Association Rules 156
Variable Selection 157
Interaction Variable Creation 157
Problems with Association Rules 158
Redundant Rules 158
Too Many Rules 158
Too Few Rules 159
Building Classification Rules from Association Rules 159
Summary 161

Chapter 6 Descriptive Modeling 163
Data Preparation Issues with Descriptive Modeling 164
Principal Component Analysis 165
The PCA Algorithm 165
Applying PCA to New Data 169
PCA for Data Interpretation 171
Additional Considerations before Using PCA 172
The Effect of Variable Magnitude on PCA Models 174
Clustering Algorithms 177
The K-Means Algorithm 178
Data Preparation for K-Means 183
Selecting the Number of Clusters 185
The Kohonen SOM Algorithm 192
Visualizing Kohonen Maps 194
Similarities with K-Means 196
Summary 197

Chapter 7 Interpreting Descriptive Models 199
Standard Cluster Model Interpretation 199
Problems with Interpretation Methods 202
Identifying Key Variables in Forming Cluster Models 203
Cluster Prototypes 209
Cluster Outliers 210
Summary 212

Chapter 8 Predictive Modeling 213
Decision Trees 214
The Decision Tree Landscape 215
Building Decision Trees 218

Decision Tree Splitting Metrics 221
Decision Tree Knobs and Options 222
Reweighting Records: Priors 224
Reweighting Records: Misclassification Costs 224
Other Practical Considerations for Decision Trees 229
Logistic Regression 230
Interpreting Logistic Regression Models 233
Other Practical Considerations for Logistic Regression 235
Neural Networks 240
Building Blocks: The Neuron 242
Neural Network Training 244
The Flexibility of Neural Networks 247
Neural Network Settings 249
Neural Network Pruning 251
Interpreting Neural Networks 252
Neural Network Decision Boundaries 253
Other Practical Considerations for Neural Networks 253
K-Nearest Neighbor 254
The k-NN Learning Algorithm 254
Distance Metrics for k-NN 258
Other Practical Considerations for k-NN 259
Naïve Bayes 264
Bayes' Theorem 264
The Naïve Bayes Classifier 268
Interpreting Naïve Bayes Classifiers 268
Other Practical Considerations for Naïve Bayes 269
Regression Models 270
Linear Regression 271
Linear Regression Assumptions 274
Variable Selection in Linear Regression 276
Interpreting Linear Regression Models 278
Using Linear Regression for Classification 279
Other Regression Algorithms 280
Summary 281

Chapter 9 Assessing Predictive Models 283
Batch Approach to Model Assessment 284
Percent Correct Classification 284
Rank-Ordered Approach to Model Assessment 293
Assessing Regression Models 301
Summary 304

Chapter 10 Model Ensembles 307
Motivation for Ensembles 307
The Wisdom of Crowds 308
Bias Variance Tradeoff 309
Bagging 311

Boosting 316
Improvements to Bagging and Boosting 320
Random Forests 320
Stochastic Gradient Boosting 321
Heterogeneous Ensembles 321
Model Ensembles and Occam's Razor 323
Interpreting Model Ensembles 323
Summary 326

Chapter 11 Text Mining 327
Motivation for Text Mining 328
A Predictive Modeling Approach to Text Mining 329
Structured vs. Unstructured Data 329
Why Text Mining Is Hard 330
Text Mining Applications 332
Data Sources for Text Mining 333
Data Preparation Steps 333
POS Tagging 333
Tokens 336
Stop Word and Punctuation Filters 336
Character Length and Number Filters 337
Stemming 337
Dictionaries 338
The Sentiment Polarity Movie Data Set 339
Text Mining Features 340
Term Frequency 341
Inverse Document Frequency 344
TF-IDF 344
Cosine Similarity 346
Multi-Word Features: N-Grams 346
Reducing Keyword Features 347
Grouping Terms 347
Modeling with Text Mining Features 347
Regular Expressions 349
Uses of Regular Expressions in Text Mining 351
Summary 352

Chapter 12 Model Deployment 353
General Deployment Considerations 354
Deployment Steps 355
Summary 375

Chapter 13 Case Studies 377
Survey Analysis Case Study: Overview 377
Business Understanding: Defining the Problem 378
Data Understanding 380
Data Preparation 381
Modeling 385

Deployment: "What-If" Analysis 391
Revisit Models 392
Deployment 401
Summary and Conclusions 401
Help Desk Case Study 402
Data Understanding: Defining the Data 403
Data Preparation 403
Modeling 405
Revisit Business Understanding 407
Deployment 409
Summary and Conclusions 411
Index 413