



# Contents

|   |              |
|---|--------------|
| <b>Acknowledgments</b>                                    | <b>xix</b>   |
| <b>About the Authors</b>                                  | <b>xxi</b>   |
| <b>Introduction</b>                                       | <b>xxiii</b> |
| <b>Chapter 1 Why and What Is Data Mining?</b>             | <b>1</b>     |
| Analytic Customer Relationship Management                 | 2            |
| The Role of Transaction Processing Systems                | 3            |
| The Role of Data Warehousing                              | 4            |
| The Role of Data Mining                                   | 5            |
| The Role of the Customer Relationship Management Strategy | 6            |
| What Is Data Mining?                                      | 7            |
| What Tasks Can Be Performed with Data Mining?             | 8            |
| Classification  | 8            |
| Estimation  | 9            |
| Prediction  | 10           |
| Affinity Grouping or Association Rules                    | 11           |
| Clustering  | 11           |
| Profiling   | 12           |
| Why Now?  | 12           |
| Data Is Being Produced                                    | 12           |
| Data Is Being Warehoused                                  | 13           |
| Computing Power Is Affordable                             | 13           |
| Interest in Customer Relationship Management Is Strong    | 13           |
| Every Business Is a Service Business                      | 14           |
| Information Is a Product                                  | 14           |
| Commercial Data Mining Software Products                  |              |
| Have Become Available                                     | 15           |

|                  |  |           |
|------------------|--|-----------|
|                  | How Data Mining Is Being Used Today                              | 15        |
|                  | A Supermarket Becomes an Information Broker                      | 15        |
|                  | A Recommendation-Based Business                                  | 16        |
|                  | Cross-Selling  | 17        |
|                  | Holding on to Good Customers                                     | 17        |
|                  | Weeding out Bad Customers  | 18        |
|                  | Revolutionizing an Industry                                      | 18        |
|                  | And Just about Anything Else                                     | 19        |
|                  | Lessons Learned  | 19        |
| <b>Chapter 2</b> | <b>The Virtuous Cycle of Data Mining</b>                         | <b>21</b> |
|                  | A Case Study in Business Data Mining                             | 22        |
|                  | Identifying the Business Challenge                               | 23        |
|                  | Applying Data Mining   | 24        |
|                  | Acting on the Results  | 25        |
|                  | Measuring the Effects  | 25        |
|                  | What Is the Virtuous Cycle?                                      | 26        |
|                  | Identify the Business Opportunity                                | 27        |
|                  | Mining Data  | 28        |
|                  | Take Action  | 30        |
|                  | Measuring Results  | 30        |
|                  | Data Mining in the Context of the Virtuous Cycle                 | 32        |
|                  | A Wireless Communications Company Makes<br>the Right Connections | 34        |
|                  | The Opportunity  | 34        |
|                  | How Data Mining Was Applied                                      | 35        |
|                  | Defining the Inputs  | 37        |
|                  | Derived Inputs   | 37        |
|                  | The Actions  | 38        |
|                  | Completing the Cycle   | 39        |
|                  | Neural Networks and Decision Trees Drive SUV Sales               | 39        |
|                  | The Initial Challenge  | 39        |
|                  | How Data Mining Was Applied                                      | 40        |
|                  | The Data   | 40        |
|                  | Down the Mine Shaft  | 40        |
|                  | The Resulting Actions  | 41        |
|                  | Completing the Cycle   | 42        |
|                  | Lessons Learned  | 42        |
| <b>Chapter 3</b> | <b>Data Mining Methodology and Best Practices</b>                | <b>43</b> |
|                  | Why Have a Methodology?  | 44        |
|                  | Learning Things That Aren't True                                 | 44        |
|                  | Patterns May Not Represent Any Underlying Rule                   | 45        |
|                  | The Model Set May Not Reflect the Relevant Population            | 46        |
|                  | Data May Be at the Wrong Level of Detail                         | 47        |

|  |    |
|--|----|
| Learning Things That Are True, but Not Useful                | 48 |
| Learning Things That Are Already Known                       | 49 |
| Learning Things That Can't Be Used                           | 49 |
| Hypothesis Testing   | 50 |
| Generating Hypotheses  | 51 |
| Testing Hypotheses   | 51 |
| Models, Profiling, and Prediction                            | 51 |
| Profiling  | 53 |
| Prediction   | 54 |
| The Methodology  | 54 |
| Step One: Translate the Business Problem                     |    |
| into a Data Mining Problem                                   | 56 |
| What Does a Data Mining Problem Look Like?                   | 56 |
| How Will the Results Be Used?                                | 57 |
| How Will the Results Be Delivered?                           | 58 |
| The Role of Business Users and Information Technology        | 58 |
| Step Two: Select Appropriate Data                            | 60 |
| What Is Available?   | 61 |
| How Much Data Is Enough?                                     | 62 |
| How Much History Is Required?                                | 63 |
| How Many Variables?  | 63 |
| What Must the Data Contain?                                  | 64 |
| Step Three: Get to Know the Data                             | 64 |
| Examine Distributions  | 65 |
| Compare Values with Descriptions                             | 66 |
| Validate Assumptions   | 67 |
| Ask Lots of Questions  | 67 |
| Step Four: Create a Model Set                                | 68 |
| Assembling Customer Signatures                               | 68 |
| Creating a Balanced Sample                                   | 68 |
| Including Multiple Timeframes                                | 70 |
| Creating a Model Set for Prediction                          | 70 |
| Partitioning the Model Set                                   | 71 |
| Step Five: Fix Problems with the Data                        | 72 |
| Categorical Variables with Too Many Values                   | 73 |
| Numeric Variables with Skewed Distributions and Outliers     | 73 |
| Missing Values   | 73 |
| Values with Meanings That Change over Time                   | 74 |
| Inconsistent Data Encoding                                   | 74 |
| Step Six: Transform Data to Bring Information to the Surface | 74 |
| Capture Trends   | 75 |
| Create Ratios and Other Combinations of Variables            | 75 |
| Convert Counts to Proportions                                | 75 |
| Step Seven: Build Models                                     | 77 |

|   |               |
|---|---------------|
| Step Eight: Assess Models   | 78            |
| Assessing Descriptive Models  | 78            |
| Assessing Directed Models   | 78            |
| Assessing Classifiers and Predictors  | 79            |
| Assessing Estimators  | 79            |
| Comparing Models Using Lift   | 81            |
| Problems with Lift  | 83            |
| Step Nine: Deploy Models  | 84            |
| Step Ten: Assess Results  | 85            |
| Step Eleven: Begin Again  | 85            |
| Lessons Learned   | 86            |
| <br><b>Chapter 4 Data Mining Applications in Marketing and Customer Relationship Management</b> | <br><b>87</b> |
| Prospecting   | 87            |
| Identifying Good Prospects  | 88            |
| Choosing a Communication Channel  | 89            |
| Picking Appropriate Messages  | 89            |
| Data Mining to Choose the Right Place to Advertise  | 90            |
| Who Fits the Profile?   | 90            |
| Measuring Fitness for Groups of Readers   | 93            |
| Data Mining to Improve Direct Marketing Campaigns   | 95            |
| Response Modeling   | 96            |
| Optimizing Response for a Fixed Budget  | 97            |
| Optimizing Campaign Profitability   | 100           |
| How the Model Affects Profitability   | 103           |
| Reaching the People Most Influenced by the Message  | 106           |
| Differential Response Analysis  | 107           |
| Using Current Customers to Learn About Prospects  | 108           |
| Start Tracking Customers before They Become Customers   | 109           |
| Gather Information from New Customers   | 109           |
| Acquisition-Time Variables Can Predict Future Outcomes  | 110           |
| Data Mining for Customer Relationship Management  | 110           |
| Matching Campaigns to Customers   | 110           |
| Segmenting the Customer Base  | 111           |
| Finding Behavioral Segments   | 111           |
| Tying Market Research Segments to Behavioral Data   | 113           |
| Reducing Exposure to Credit Risk  | 113           |
| Predicting Who Will Default   | 113           |
| Improving Collections   | 114           |
| Determining Customer Value  | 114           |
| Cross-selling, Up-selling, and Making Recommendations   | 115           |
| Finding the Right Time for an Offer   | 115           |
| Making Recommendations  | 116           |
| Retention and Churn   | 116           |
| Recognizing Churn   | 116           |
| Why Churn Matters   | 117           |
| Different Kinds of Churn  | 118           |



|                  |   |            |
|------------------|---|------------|
|                  | Different Kinds of Churn Model                                  | 119        |
|                  | Predicting Who Will Leave                                       | 119        |
|                  | Predicting How Long Customers Will Stay                         | 119        |
|                  | Lessons Learned   | 120        |
| <b>Chapter 5</b> | <b>The Lure of Statistics: Data Mining Using Familiar Tools</b> | <b>123</b> |
|                  | Occam's Razor   | 124        |
|                  | The Null Hypothesis   | 125        |
|                  | P-Values  | 126        |
|                  | A Look at Data  | 126        |
|                  | Looking at Discrete Values                                      | 127        |
|                  | Histograms  | 127        |
|                  | Time Series   | 128        |
|                  | Standardized Values   | 129        |
|                  | From Standardized Values to Probabilities                       | 133        |
|                  | Cross-Tabulations   | 136        |
|                  | Looking at Continuous Variables                                 | 136        |
|                  | Statistical Measures for Continuous Variables                   | 137        |
|                  | Variance and Standard Deviation                                 | 138        |
|                  | A Couple More Statistical Ideas                                 | 139        |
|                  | Measuring Response  | 139        |
|                  | Standard Error of a Proportion                                  | 139        |
|                  | Comparing Results Using Confidence Bounds                       | 141        |
|                  | Comparing Results Using Difference of Proportions               | 143        |
|                  | Size of Sample  | 145        |
|                  | What the Confidence Interval Really Means                       | 146        |
|                  | Size of Test and Control for an Experiment                      | 147        |
|                  | Multiple Comparisons  | 148        |
|                  | The Confidence Level with Multiple Comparisons                  | 148        |
|                  | Bonferroni's Correction   | 149        |
|                  | Chi-Square Test   | 149        |
|                  | Expected Values   | 150        |
|                  | Chi-Square Value  | 151        |
|                  | Comparison of Chi-Square to Difference of Proportions           | 153        |
|                  | An Example: Chi-Square for Regions and Starts                   | 155        |
|                  | Data Mining and Statistics                                      | 158        |
|                  | No Measurement Error in Basic Data                              | 159        |
|                  | There Is a Lot of Data  | 160        |
|                  | Time Dependency Pops Up Everywhere                              | 160        |
|                  | Experimentation is Hard   | 160        |
|                  | Data Is Censored and Truncated                                  | 161        |
|                  | Lessons Learned   | 162        |
| <b>Chapter 6</b> | <b>Decision Trees</b>   | <b>165</b> |
|                  | What Is a Decision Tree?  | 166        |
|                  | Classification  | 166        |
|                  | Scoring   | 169        |
|                  | Estimation  | 170        |
|                  | Trees Grow in Many Forms  | 170        |

|  |            |
|--|------------|
| How a Decision Tree Is Grown                           | 171        |
| Finding the Splits                                     | 172        |
| Splitting on a Numeric Input Variable                  | 173        |
| Splitting on a Categorical Input Variable              | 174        |
| Splitting in the Presence of Missing Values            | 174        |
| Growing the Full Tree                                  | 175        |
| Measuring the Effectiveness Decision Tree              | 176        |
| Tests for Choosing the Best Split                      | 176        |
| Purity and Diversity                                   | 177        |
| Gini or Population Diversity                           | 178        |
| Entropy Reduction or Information Gain                  | 179        |
| Information Gain Ratio                                 | 180        |
| Chi-Square Test  | 180        |
| Reduction in Variance                                  | 183        |
| F Test   | 183        |
| Pruning  | 184        |
| The CART Pruning Algorithm                             | 185        |
| Creating the Candidate Subtrees                        | 185        |
| Picking the Best Subtree                               | 189        |
| Using the Test Set to Evaluate the Final Tree          | 189        |
| The C5 Pruning Algorithm                               | 190        |
| Pessimistic Pruning                                    | 191        |
| Stability-Based Pruning                                | 191        |
| Extracting Rules from Trees                            | 193        |
| Taking Cost into Account                               | 195        |
| Further Refinements to the Decision Tree Method        | 195        |
| Using More Than One Field at a Time                    | 195        |
| Tilting the Hyperplane                                 | 197        |
| Neural Trees   | 199        |
| Piecewise Regression Using Trees                       | 199        |
| Alternate Representations for Decision Trees           | 199        |
| Box Diagrams   | 199        |
| Tree Ring Diagrams                                     | 201        |
| Decision Trees in Practice                             | 203        |
| Decision Trees as a Data Exploration Tool              | 203        |
| Applying Decision-Tree Methods to Sequential Events    | 205        |
| Simulating the Future                                  | 206        |
| Case Study: Process Control in a Coffee-Roasting Plant | 206        |
| Lessons Learned  | 209        |
| <b>Chapter 7 Artificial Neural Networks</b>            | <b>211</b> |
| A Bit of History                                       | 212        |
| Real Estate Appraisal                                  | 213        |
| Neural Networks for Directed Data Mining               | 219        |
| What Is a Neural Net?                                  | 220        |
| What Is the Unit of a Neural Network?                  | 222        |
| Feed-Forward Neural Networks                           | 226        |

|  |     |
|--|-----|
| How Does a Neural Network Learn Using Back Propagation?      | 228 |
| Heuristics for Using Feed-Forward, Back Propagation Networks | 231 |
| Choosing the Training Set                                    | 232 |
| Coverage of Values for All Features                          | 232 |
| Number of Features   | 233 |
| Size of Training Set   | 234 |
| Number of Outputs  | 234 |
| Preparing the Data   | 235 |
| Features with Continuous Values                              | 235 |
| Features with Ordered, Discrete (Integer) Values             | 238 |
| Features with Categorical Values                             | 239 |
| Other Types of Features                                      | 241 |
| Interpreting the Results                                     | 241 |
| Neural Networks for Time Series                              | 244 |
| How to Know What Is Going on Inside a Neural Network         | 247 |
| Self-Organizing Maps   | 249 |
| What Is a Self-Organizing Map?                               | 249 |
| Example: Finding Clusters                                    | 252 |
| Lessons Learned  | 254 |

|                  |  |            |
|------------------|--|------------|
| <b>Chapter 8</b> | <b>Nearest Neighbor Approaches: Memory-Based Reasoning and Collaborative Filtering</b> | <b>257</b> |
|                  | Memory Based Reasoning   | 258        |
|                  | Example: Using MBR to Estimate Rents in Tuxedo, New York                               | 259        |
|                  | Challenges of MBR  | 262        |
|                  | Choosing a Balanced Set of Historical Records  | 262        |
|                  | Representing the Training Data   | 263        |
|                  | Determining the Distance Function, Combination Function, and Number of Neighbors       | 265        |
|                  | Case Study: Classifying News Stories   | 265        |
|                  | What Are the Codes?  | 266        |
|                  | Applying MBR   | 267        |
|                  | Choosing the Training Set  | 267        |
|                  | Choosing the Distance Function   | 267        |
|                  | Choosing the Combination Function  | 267        |
|                  | Choosing the Number of Neighbors   | 270        |
|                  | The Results  | 270        |
|                  | Measuring Distance   | 271        |
|                  | What Is a Distance Function?   | 271        |
|                  | Building a Distance Function One Field at a Time                                       | 274        |
|                  | Distance Functions for Other Data Types  | 277        |
|                  | When a Distance Metric Already Exists  | 278        |
|                  | The Combination Function: Asking the Neighbors for the Answer                          | 279        |
|                  | The Basic Approach: Democracy  | 279        |
|                  | Weighted Voting  | 281        |

|  |            |
|--|------------|
| Collaborative Filtering: A Nearest Neighbor Approach to Making Recommendations | 282        |
| Building Profiles  | 283        |
| Comparing Profiles   | 284        |
| Making Predictions   | 284        |
| Lessons Learned  | 285        |
| <b>Chapter 9 Market Basket Analysis and Association Rules</b>                  | <b>287</b> |
| Defining Market Basket Analysis  | 289        |
| Three Levels of Market Basket Data   | 289        |
| Order Characteristics  | 292        |
| Item Popularity  | 293        |
| Tracking Marketing Interventions   | 293        |
| Clustering Products by Usage   | 294        |
| Association Rules  | 296        |
| Actionable Rules   | 296        |
| Trivial Rules  | 297        |
| Inexplicable Rules   | 297        |
| How Good Is an Association Rule?   | 299        |
| Building Association Rules   | 302        |
| Choosing the Right Set of Items  | 303        |
| Product Hierarchies Help to Generalize Items                                   | 305        |
| Virtual Items Go beyond the Product Hierarchy                                  | 307        |
| Data Quality   | 308        |
| Anonymous versus Identified  | 308        |
| Generating Rules from All This Data  | 308        |
| Calculating Confidence   | 309        |
| Calculating Lift   | 310        |
| The Negative Rule  | 311        |
| Overcoming Practical Limits  | 311        |
| The Problem of Big Data  | 313        |
| Extending the Ideas  | 315        |
| Using Association Rules to Compare Stores                                      | 315        |
| Dissociation Rules   | 317        |
| Sequential Analysis Using Association Rules                                    | 318        |
| Lessons Learned  | 319        |
| <b>Chapter 10 Link Analysis</b>  | <b>321</b> |
| Basic Graph Theory   | 322        |
| Seven Bridges of Königsberg  | 325        |
| Traveling Salesman Problem   | 327        |
| Directed Graphs  | 330        |
| Detecting Cycles in a Graph  | 330        |
| A Familiar Application of Link Analysis  | 331        |
| The Kleinberg Algorithm  | 332        |
| The Details: Finding Hubs and Authorities                                      | 333        |
| Creating the Root Set  | 333        |
| Identifying the Candidates   | 334        |
| Ranking Hubs and Authorities   | 334        |
| Hubs and Authorities in Practice   | 336        |



|   |     |
|---|-----|
| Case Study: Who Is Using Fax Machines from Home?    | 336 |
| Why Finding Fax Machines Is Useful                  | 336 |
| The Data as a Graph                                 | 337 |
| The Approach  | 338 |
| Some Results  | 340 |
| Case Study: Segmenting Cellular Telephone Customers | 343 |
| The Data  | 343 |
| Analyses without Graph Theory                       | 343 |
| A Comparison of Two Customers                       | 344 |
| The Power of Link Analysis                          | 345 |
| Lessons Learned                                     | 346 |

## **Chapter 11    Automatic Cluster Detection    349**

|   |     |
|---|-----|
| Searching for Islands of Simplicity               | 350 |
| Star Light, Star Bright                           | 351 |
| Fitting the Troops                                | 352 |
| K-Means Clustering                                | 354 |
| Three Steps of the K-Means Algorithm              | 354 |
| What K Means                                      | 356 |
| Similarity and Distance                           | 358 |
| Similarity Measures and Variable Type             | 359 |
| Formal Measures of Similarity                     | 360 |
| Geometric Distance between Two Points             | 360 |
| Angle between Two Vectors                         | 361 |
| Manhattan Distance                                | 363 |
| Number of Features in Common                      | 363 |
| Data Preparation for Clustering                   | 363 |
| Scaling for Consistency                           | 363 |
| Use Weights to Encode Outside Information         | 365 |
| Other Approaches to Cluster Detection             | 365 |
| Gaussian Mixture Models                           | 365 |
| Agglomerative Clustering                          | 368 |
| An Agglomerative Clustering Algorithm             | 368 |
| Distance between Clusters                         | 368 |
| Clusters and Trees                                | 370 |
| Clustering People by Age: An Example of           |     |
| Agglomerative Clustering                          | 370 |
| Divisive Clustering                               | 371 |
| Self-Organizing Maps                              | 372 |
| Evaluating Clusters                               | 372 |
| Inside the Cluster                                | 373 |
| Outside the Cluster                               | 373 |
| Case Study: Clustering Towns                      | 374 |
| Creating Town Signatures                          | 374 |
| The Data  | 375 |
| Creating Clusters                                 | 377 |
| Determining the Right Number of Clusters          | 377 |
| Using Thematic Clusters to Adjust Zone Boundaries | 380 |
| Lessons Learned                                   | 381 |

|                   |   |            |
|-------------------|---|------------|
| <b>Chapter 12</b> | <b>Knowing When to Worry: Hazard Functions and Survival Analysis in Marketing</b> | <b>383</b> |
|                   | Customer Retention  | 385        |
|                   | Calculating Retention   | 385        |
|                   | What a Retention Curve Reveals  | 386        |
|                   | Finding the Average Tenure from a Retention Curve                                 | 387        |
|                   | Looking at Retention as Decay   | 389        |
|                   | Hazards   | 394        |
|                   | The Basic Idea  | 394        |
|                   | Examples of Hazard Functions  | 397        |
|                   | Constant Hazard   | 397        |
|                   | Bathtub Hazard  | 397        |
|                   | A Real-World Example  | 398        |
|                   | Censoring   | 399        |
|                   | Other Types of Censoring  | 402        |
|                   | From Hazards to Survival  | 404        |
|                   | Retention   | 404        |
|                   | Survival  | 405        |
|                   | Proportional Hazards  | 408        |
|                   | Examples of Proportional Hazards  | 409        |
|                   | Stratification: Measuring Initial Effects on Survival                             | 410        |
|                   | Cox Proportional Hazards  | 410        |
|                   | Limitations of Proportional Hazards   | 411        |
|                   | Survival Analysis in Practice   | 412        |
|                   | Handling Different Types of Attrition   | 412        |
|                   | When Will a Customer Come Back?   | 413        |
|                   | Forecasting   | 415        |
|                   | Hazards Changing over Time  | 416        |
|                   | Lessons Learned   | 418        |
| <b>Chapter 13</b> | <b>Genetic Algorithms</b>   | <b>421</b> |
|                   | How They Work   | 423        |
|                   | Genetics on Computers   | 424        |
|                   | Selection   | 429        |
|                   | Crossover   | 430        |
|                   | Mutation  | 431        |
|                   | Representing Data   | 432        |
|                   | Case Study: Using Genetic Algorithms for  |            |
|                   | Resource Optimization   | 433        |
|                   | Schemata: Why Genetic Algorithms Work   | 435        |
|                   | More Applications of Genetic Algorithms   | 438        |
|                   | Application to Neural Networks  | 439        |
|                   | Case Study: Evolving a Solution for Response Modeling                             | 440        |
|                   | Business Context  | 440        |
|                   | Data  | 441        |
|                   | The Data Mining Task: Evolving a Solution   | 442        |
|                   | Beyond the Simple Algorithm   | 444        |
|                   | Lessons Learned   | 446        |

|  |            |
|--|------------|
| <b>Chapter 14 Data Mining throughout the Customer Life Cycle</b> | <b>447</b> |
| Levels of the Customer Relationship                              | 448        |
| Deep Intimacy  | 449        |
| Mass Intimacy  | 451        |
| In-between Relationships   | 453        |
| Indirect Relationships   | 453        |
| Customer Life Cycle  | 454        |
| The Customer's Life Cycle: Life Stages                           | 455        |
| Customer Life Cycle  | 456        |
| Subscription Relationships versus Event-Based Relationships      | 458        |
| Event-Based Relationships  | 458        |
| Subscription-Based Relationships                                 | 459        |
| Business Processes Are Organized around the Customer Life Cycle  | 461        |
| Customer Acquisition   | 461        |
| Who Are the Prospects?   | 462        |
| When Is a Customer Acquired?                                     | 462        |
| What Is the Role of Data Mining?                                 | 464        |
| Customer Activation  | 464        |
| Relationship Management  | 466        |
| Retention  | 467        |
| Winback  | 470        |
| Lessons Learned  | 470        |
| <b>Chapter 15 Data Warehousing, OLAP, and Data Mining</b>        | <b>473</b> |
| The Architecture of Data   | 475        |
| Transaction Data, the Base Level                                 | 476        |
| Operational Summary Data   | 477        |
| Decision-Support Summary Data                                    | 477        |
| Database Schema  | 478        |
| Metadata   | 483        |
| Business Rules   | 484        |
| A General Architecture for Data Warehousing                      | 484        |
| Source Systems   | 486        |
| Extraction, Transformation, and Load                             | 487        |
| Central Repository   | 488        |
| Metadata Repository  | 491        |
| Data Marts   | 491        |
| Operational Feedback   | 492        |
| End Users and Desktop Tools                                      | 492        |
| Analysts   | 492        |
| Application Developers   | 493        |
| Business Users   | 494        |
| Where Does OLAP Fit In?  | 494        |
| What's in a Cube?  | 497        |
| Three Varieties of Cubes   | 498        |
| Facts  | 501        |
| Dimensions and Their Hierarchies                                 | 502        |
| Conformed Dimensions   | 504        |

|  |            |
|--|------------|
| Star Schema  | 505        |
| OLAP and Data Mining                                   | 507        |
| Where Data Mining Fits in with Data Warehousing        | 508        |
| Lots of Data   | 509        |
| Consistent, Clean Data                                 | 510        |
| Hypothesis Testing and Measurement                     | 510        |
| Scalable Hardware and RDBMS Support                    | 511        |
| Lessons Learned  | 511        |
| <b>Chapter 16 Building the Data Mining Environment</b> | <b>513</b> |
| A Customer-Centric Organization                        | 514        |
| An Ideal Data Mining Environment                       | 515        |
| The Power to Determine What Data Is Available          | 515        |
| The Skills to Turn Data into Actionable Information    | 516        |
| All the Necessary Tools                                | 516        |
| Back to Reality  | 516        |
| Building a Customer-Centric Organization               | 516        |
| Creating a Single Customer View                        | 517        |
| Defining Customer-Centric Metrics                      | 519        |
| Collecting the Right Data                              | 520        |
| From Customer Interactions to Learning Opportunities   | 520        |
| Mining Customer Data                                   | 521        |
| The Data Mining Group                                  | 521        |
| Outsourcing Data Mining                                | 522        |
| Outsourcing Occasional Modeling                        | 522        |
| Outsourcing Ongoing Data Mining                        | 523        |
| Insourcing Data Mining                                 | 524        |
| Building an Interdisciplinary Data Mining Group        | 524        |
| Building a Data Mining Group in IT                     | 524        |
| Building a Data Mining Group in the Business Units     | 525        |
| What to Look for in Data Mining Staff                  | 525        |
| Data Mining Infrastructure                             | 526        |
| The Mining Platform                                    | 527        |
| The Scoring Platform                                   | 527        |
| One Example of a Production Data Mining Architecture   | 528        |
| Architectural Overview                                 | 528        |
| Customer Interaction Module                            | 529        |
| Analysis Module  | 530        |
| Data Mining Software                                   | 532        |
| Range of Techniques                                    | 532        |
| Scalability  | 533        |
| Support for Scoring                                    | 534        |
| Multiple Levels of User Interfaces                     | 535        |
| Comprehensible Output                                  | 536        |
| Ability to Handle Diverse Data Types                   | 536        |
| Documentation and Ease of Use                          | 536        |



|   |            |
|---|------------|
| Availability of Training for Both Novice and<br>Advanced Users, Consulting, and Support | 537        |
| Vendor Credibility  | 537        |
| Lessons Learned   | 537        |
| <b>Chapter 17 Preparing Data for Mining</b>   | <b>539</b> |
| What Data Should Look Like  | 540        |
| The Customer Signature  | 540        |
| The Columns   | 542        |
| Columns with One Value  | 544        |
| Columns with Almost Only One Value  | 544        |
| Columns with Unique Values  | 546        |
| Columns Correlated with Target  | 547        |
| Model Roles in Modeling   | 547        |
| Variable Measures   | 549        |
| Numbers   | 550        |
| Dates and Times   | 552        |
| Fixed-Length Character Strings  | 552        |
| IDs and Keys  | 554        |
| Names   | 555        |
| Addresses   | 555        |
| Free Text   | 556        |
| Binary Data (Audio, Image, Etc.)  | 557        |
| Data for Data Mining  | 557        |
| Constructing the Customer Signature   | 558        |
| Cataloging the Data   | 559        |
| Identifying the Customer  | 560        |
| First Attempt   | 562        |
| Identifying the Time Frames   | 562        |
| Taking a Recent Snapshot  | 562        |
| Pivoting Columns  | 563        |
| Calculating the Target  | 563        |
| Making Progress   | 564        |
| Practical Issues  | 564        |
| Exploring Variables   | 565        |
| Distributions Are Histograms  | 565        |
| Changes over Time   | 566        |
| Crosstabulations  | 567        |
| Deriving Variables  | 568        |
| Extracting Features from a Single Value   | 569        |
| Combining Values within a Record  | 569        |
| Looking Up Auxiliary Information  | 569        |
| Pivoting Regular Time Series  | 572        |
| Summarizing Transactional Records   | 574        |
| Summarizing Fields across the Model Set   | 574        |

|  |            |
|--|------------|
| Examples of Behavior-Based Variables               | 575        |
| Frequency of Purchase                              | 575        |
| Declining Usage                                    | 577        |
| Revolvers, Transactors, and Convenience Users:     |            |
| Defining Customer Behavior                         | 580        |
| Data   | 581        |
| Segmenting by Estimating Revenue                   | 581        |
| Segmentation by Potential                          | 583        |
| Customer Behavior by Comparison to Ideals          | 585        |
| The Ideal Convenience User                         | 587        |
| The Dark Side of Data                              | 590        |
| Missing Values                                     | 590        |
| Dirty Data   | 592        |
| Inconsistent Values                                | 593        |
| Computational Issues                               | 594        |
| Source Systems                                     | 594        |
| Extraction Tools                                   | 595        |
| Special-Purpose Code                               | 595        |
| Data Mining Tools                                  | 595        |
| Lessons Learned                                    | 596        |
| <b>Chapter 18 Putting Data Mining to Work</b>      | <b>597</b> |
| Getting Started                                    | 598        |
| What to Expect from a Proof-of-Concept Project     | 599        |
| Identifying a Proof-of-Concept Project             | 599        |
| Implementing the Proof-of-Concept Project          | 601        |
| Act on Your Findings                               | 602        |
| Measure the Results of the Actions                 | 603        |
| Choosing a Data Mining Technique                   | 605        |
| Formulate the Business Goal as a Data Mining Task  | 605        |
| Determine the Relevant Characteristics of the Data | 606        |
| Data Type  | 606        |
| Number of Input Fields                             | 607        |
| Free-Form Text                                     | 607        |
| Consider Hybrid Approaches                         | 608        |
| How One Company Began Data Mining                  | 608        |
| A Controlled Experiment in Retention               | 609        |
| The Data   | 611        |
| The Findings                                       | 613        |
| The Proof of the Pudding                           | 614        |
| Lessons Learned                                    | 614        |
| <b>Index</b>                                       | <b>615</b> |