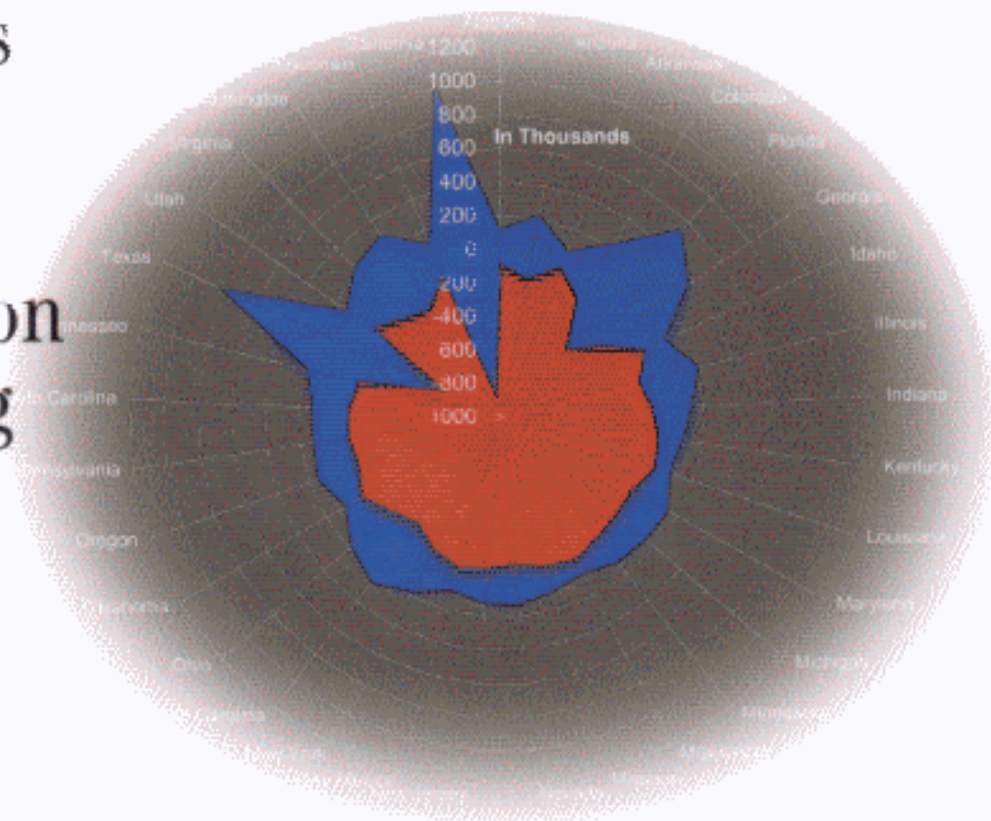




Visual Data Mining

Techniques
and Tools
for Data
Visualization
and Mining

Tom Soukup
Ian Davidson



| | |
|--------------------------|-------------|
| Acknowledgments | xv |
| About the Authors | xvii |
| Trademarks | xix |
| Introduction | xxi |

| | | |
|-----------------|--|----------|
| Part One | Introduction and Project Planning Phase | 1 |
|-----------------|--|----------|

| | | |
|------------------|---|-----------|
| Chapter 1 | Introduction to Data Visualization and Visual Data Mining | 3 |
| | Visualization Data Sets | 5 |
| | Visualization Data Types | 6 |
| | Visual versus Data Dimensions | 7 |
| | Data Visualization Tools | 8 |
| | Multidimensional Data Visualization Tools | 8 |
| | Column and Bar Graphs | 10 |
| | Distribution and Histogram Graphs | 10 |
| | Box Graphs | 12 |
| | Line Graphs | 14 |
| | Scatter Graphs | 16 |
| | Pie Graphs | 17 |
| | Hierarchical and Landscape Data Visualization Tools | 19 |
| | Tree Visualizations | 19 |
| | Map Visualizations | 20 |
| | Visual Data Mining Tools | 21 |
| | Summary | 23 |
| Chapter 2 | Step 1: Justifying and Planning the Data Visualization and Data Mining Project | 25 |
| | Classes of Projects | 26 |
| | Project Justifications | 27 |
| | Dayton Hudson Corp. Success Story | 29 |
| | Marketing Dynamics Success Story | 29 |

| | |
|---|-----------|
| Sprint Success Story | 30 |
| Lowestfare.com Success Story | 30 |
| Challenges to Visual Data Mining | 31 |
| Data Visualization, Analysis, and Statistics are Meaningless | 31 |
| Why Are the Predictions Not 100 Percent Accurate? | 31 |
| Our Data Can't Be Visualized or Mined | 32 |
| Closed-Loop Business Model | 32 |
| Using the Closed-Loop Business Model | 34 |
| Project Timeline | 36 |
| Project Resources and Roles | 36 |
| Data and Business Analyst Team | 38 |
| Domain Expert Team | 38 |
| Decision Maker Team | 40 |
| Operations Team | 41 |
| Data Warehousing Team | 42 |
| Project Justification and Plan for the Case Study | 44 |
| Summary | 48 |
| Chapter 3 Step 2: Identifying the Top Business Questions | 49 |
| Choosing the Top Business Questions | 49 |
| Problems Data Mining Does Not Address | 50 |
| Data Visualization Problem Definitions | 51 |
| Multidimensional or Comparative Visualization | |
| Problem Definitions | 51 |
| Geographic or Spatial Data Visualization Problem Definitions | 52 |
| Visual Data Mining Problem Definitions | 52 |
| Classification Data Mining Problem Definitions | 53 |
| Estimation Data Mining Problem Definitions | 54 |
| Association Grouping Data Mining Problem Definitions | 54 |
| Clustering and Segmentation Data Mining | |
| Problem Definitions | 54 |
| Prediction Data Mining Problem Definitions | 55 |
| Which Data Mining Techniques Can Address | |
| a Business Issue? | 55 |
| Mapping the ROI Targets | 57 |
| Determining the Visualization and Data Mining | |
| Analysis Goals and Success Criteria | 59 |
| Problem and Objective Definitions for the Case Study | 61 |
| Summary | 63 |

Part Two Data Preparation Phase 65

| | |
|--|-----------|
| Chapter 4 Step 3: Choosing the Business Data Set | 67 |
| Identifying the Operational Data | 68 |
| Exploratory Data Mart | 69 |
| Business Data Sets | 71 |
| Data Types | 74 |
| Experimental Unit | 74 |
| Surveying Discrete and Continuous Columns with Visualizations | 75 |
| Selecting Columns from the Operational Data Sources | 79 |
| Encoded Data Dimensions | 80 |
| Data Dimension Consistency | 82 |
| Business Rule Consistency | 82 |
| Unique Columns | 82 |
| Duplicate Columns | 83 |
| Correlated Columns | 84 |
| Insignificant Columns | 84 |
| Developing and Documenting the ECTL Procedures | 85 |
| Data Cleaning | 87 |
| Techniques for Handling Data Noise, NULLs, and Missing Values | 89 |
| Handling NULLs | 91 |
| Sampling the Operational Data Sources | 92 |
| Avoiding Biased Sampling | 94 |
| Available ECTL Tools | 96 |
| Documenting the ECTL Procedures | 97 |
| Choosing the Business Data Set for the Case Study | 98 |
| Identifying the Operational Data Sources | 100 |
| ECTL Processing of the Customer File | 102 |
| Documenting ECTL Procedure for the Customer File | 108 |
| ECTL Processing of the Contract File | 109 |
| Documenting ECTL Procedure for the Contact File | 113 |
| ECTL Processing of the Invoice File | 113 |
| Documenting ECTL Procedure for the Invoice File | 118 |
| ECTL Processing of the Demographic File | 118 |
| Documenting ECTL Procedure for the Demographic File | 122 |
| Creating the Production Business Data Set | 123 |
| Review of the ECTL Procedures for the Case Study | 126 |
| Summary | 127 |

| | | |
|------------------|--|------------|
| Chapter 5 | Step 4: Transforming the Business Data Set | 129 |
| | Types of Logical Transformations | 130 |
| | Table-Level Logical Transformations | 131 |
| | Transforming Weighted Data Sets | 132 |
| | Transforming Column Weights | 133 |
| | Transforming Record Weights | 135 |
| | Transforming Time Series Data Sets | 137 |
| | Aggregating the Data Sets | 140 |
| | Filtering Data Sets | 142 |
| | Column-Level Logical Transformations | 143 |
| | Simple Column Transformations | 144 |
| | Column Grouping Transformations | 146 |
| | Documenting the Logical Transformations | 151 |
| | Logically Transforming the Business Data Set for the Customer Retention VDM Case Study | 154 |
| | Logically Transforming the <i>customer_join</i> Business Data Set | 156 |
| | Documenting the Logical Transformations for the Business Data Set <i>customer_join</i> | 163 |
| | Logically Transforming the <i>customer_demographic</i> Business Data Set | 164 |
| | Documenting the Logical Transformations for the Business Data Set <i>customer_demographic</i> | 168 |
| | Review of the Logical Transformation Procedures for the Case Study | 168 |
| | Summary | 169 |
| Chapter 6 | Step 5: Verify the Business Data Set | 171 |
| | Verification Process | 172 |
| | Verifying the Integrity of the Data Preparation Operations | 173 |
| | Discrete Column Verification Techniques | 174 |
| | Continuous Column Verification Techniques | 178 |
| | Verifying Common ECTL Data Preparation Operations | 180 |
| | Verifying the Logic of the Data Preparation Operations | 181 |
| | Verifying Common Logical Transformation Operations | 181 |
| | Data Profiling Tools | 188 |
| | Verifying the Data Set for the Case Study | 189 |
| | Verifying the ECTL Procedures | 191 |
| | Verifying the ECTL Data Preparation Step for the Customer Table | 191 |
| | Verifying the ECTL Data Preparation Step for the Contract Table | 197 |
| | Verifying the ECTL Data Preparation Step for the Invoice Table | 197 |
| | Verifying the ECTL Data Preparation Step for the Demographic Table | 199 |
| | Verifying the Logical Transformations | 199 |
| | Summary | 201 |

Part Three Data Analysis Phase and Beyond**203**

| | | |
|------------------|--|------------|
| Chapter 7 | Step 6: Choosing the Visualization or Visual Mining Tool | 205 |
| | Choosing the Right Data Visualization Tool | 206 |
| | Multidimensional Visualizations | 208 |
| | Column and Bar Graphs | 208 |
| | Area, Line, High-Low-Close, and Radar Graphs | 216 |
| | Histogram, Distribution, Pie, and Doughnut Graphs | 219 |
| | Scatter Graphs | 220 |
| | Specialized Landscape and Hierarchical Visualizations | 221 |
| | Map Graphs | 222 |
| | Tree Graphs | 222 |
| | Choosing the Right Data Mining Tool | 225 |
| | Which Subset of the Available Tools Is Applicable? | 225 |
| | Business Questions to Address | 225 |
| | How Is the Model to Be Used? | 227 |
| | Supervised and Unsupervised Learning | 227 |
| | Supervised Learning Tools | 228 |
| | Decision Trees and Rule Set Models | 228 |
| | Neural Network Models for Classification | 230 |
| | Linear Regression Models | 231 |
| | Logistic Regression | 232 |
| | Unsupervised Learning Tools | 233 |
| | Association Rules | 233 |
| | K-Means and Clustering | 234 |
| | Kohonen Self-Organizing Maps | 235 |
| | Tools to Solve Typical Problems | 236 |
| | Which of the Applicable Tools Are Best for My Situation? | 236 |
| | How the Different Techniques Handle Data Types | 240 |
| | Choosing the Visualization or Mining Tool for the Case Study | 242 |
| | Choosing the Data Visualization Tools | 243 |
| | Choosing the Data Mining Tools | 248 |
| | Tuning the Data Mining Tool Selection | 248 |
| | Summary | 250 |
| Chapter 8 | Step 7: Analyzing the Visualization or Mining Tool | 253 |
| | Analyzing the Data Visualizations | 254 |
| | Using Frequency Graphs to Discover and Evaluate Key Business Indicators | 254 |
| | Using Pareto Graphs to Discover and Evaluate the Importance of Key Business Indicators | 262 |
| | Using Radar Graphs to Spot Seasonal Trends and Problem Areas | 265 |
| | Using Line Graphs to Analyze Time Relationships | 268 |
| | Using Scatter Graphs to Evaluate Cause-and-Effect Relationships | 270 |
| | Analyzing the Data Mining Models | 276 |

| | |
|--|------------|
| Visualizations to Understand the Performance of the | |
| Core Data Mining Tasks | 276 |
| Classification | 276 |
| Estimation | 283 |
| Association Grouping | 284 |
| Clustering and Segmenting | 284 |
| Using Visualization to Understand and Evaluate | |
| Supervised Learning Models | 288 |
| Decision Trees | 288 |
| Neural Networks | 290 |
| Uses of Visualizations after Model Deployment | 290 |
| Analyzing the Visualization or Mining Tools | |
| for the Case Study | 291 |
| Using Frequency Graphs with Trend Lines to Analyze | |
| Time Relationships | 294 |
| Using Pareto Graphs to Discover and Evaluate the | |
| Importance of Key Business Indicators | 295 |
| Using Scatter Graphs to Evaluate Cause-and-Effect | |
| Relationships | 296 |
| Using Data Mining Tools to Gain an Insight into Churn | 299 |
| Profiling the Ones That Got Away | 299 |
| Trying to Predict the Defectors | 305 |
| Explaining Why People Leave | 309 |
| Predicting When People Will Leave | 312 |
| Summary | 315 |
| Chapter 9 Step 8: Verifying and Presenting the Visualizations | 317 |
| or Mining Models | |
| Verifying the Data Visualizations and Mining Models | 318 |
| Verifying Logical Transformations to the Business Data Set | 318 |
| Verifying Your Business Assumptions | 319 |
| Organizing and Creating the Business Presentation | 320 |
| Parts of the Business Presentation | 320 |
| Description of the VDM Project Goals | 321 |
| Highlights of the Discoveries and Data Mining Models | 321 |
| Call to Action | 324 |
| VDM Project Implementation Phase | 326 |
| Create Action Plan | 327 |
| Approve Action Plan | 327 |
| Implement Action Plan | 327 |
| Measure the Results | 328 |
| Verifying and Presenting the Analysis for the Case Study | 329 |
| Verifying Logical Transformations to the Business Data Set | 329 |
| Verifying the Business Assumptions | 330 |

| | |
|--|------------|
| The Business Presentation | 331 |
| Customer Retention Project Goals and Objectives | 331 |
| Highlights of the Discoveries | 332 |
| Call to Action | 334 |
| Summary | 337 |
| Chapter 10 The Future of Visual Data Mining | 339 |
| The Project Planning Phase | 339 |
| The Data Preparation Phase | 341 |
| The Data Analysis Phase | 347 |
| Trends in Commercial Visual Data Mining Software | 350 |
| More Chart Types and User-Defined Layouts | 351 |
| Dynamic Visualizations That Allow User Interaction | 353 |
| Size and Complexity of Data Structures Visualized | 354 |
| Standards That Allow Exchanges between Tools | 354 |
| Summary | 355 |
| Glossary | 357 |
| References | 363 |
| Index | 365 |