Contents

Preface xvii

Introduction 1

Chapter 1
Data Exploration as a Process 9

1.1 **The Data Exploration Process 10**

1.1.1 Stage 1: Exploring the Problem Space 12
1.1.2 Stage 2: Exploring the Solution Space 19
1.1.3 Stage 3: Specifying the Implementation Method 22
1.1.4 Stage 4: Mining the Data 22
1.1.5 Exploration: Mining and Modeling 28

1.2 **Data Mining, Modeling, and Modeling Tools 28**

1.2.1 Ten Golden Rules 29
1.2.2 Introducing Modeling Tools 30
1.2.3 Types of Models 32
1.2.4 Active and Passive Models 33
1.2.5 Explanatory and Predictive Models 33
1.2.6 Static and Continuously Learning Models 35

1.3 **Summary 37**

Supplemental Material 39

A Continuously Learning Model Application 39
How the Continuously Learning Model Worked 40

Chapter 2
The Nature of the World and Its Impact on Data Preparation 45

2.1 **Measuring the World 46**

2.1.1 Objects 46
2.1.2 Capturing Measurements 47

2.1.3 Errors of Measurement 48
2.1.4 Tying Measurements to the Real World 53

2.2 Types of Measurements 53

2.2.1 Scalar Measurements 54
2.2.2 Nonscalar Measurements 60

2.3 Continua of Attributes of Variables 60

2.3.1 The Qualitative-Quantitative Continuum 61
2.3.2 The Discrete-Continuous Continuum 61

2.4 Scale Measurement Example 66

2.5 Transformations and Difficulties—Variables, Data, and Information 66

2.6 Building Mineable Data Representations 67

2.6.1 Data Representation 68
2.6.2 Building Data—Dealing with Variables 69
2.6.3 Building Mineable Data Sets 77

2.7 Summary 86

Supplemental Material 87

Combinations 87

Chapter 3

Data Preparation as a Process 89

3.1 Data Preparation: Inputs, Outputs, Models, and Decisions 90

3.1.1 Step 1: Prepare the Data 92
3.1.2 Step 2: Survey the Data 97
3.1.3 Step 3: Model the Data 98
3.1.4 Use the Model 98

3.2 Modeling Tools and Data Preparation 100

3.2.1 How Modeling Tools Drive Data Preparation 102
3.2.2 Decision Trees 104
3.2.3 Decision Lists 104
3.2.4 Neural Networks 107
3.2.5 Evolution Programs 107
3.2.6 Modeling Data with the Tools 107
3.2.7 Predictions and Rules 109
3.2.8 Choosing Techniques 111
3.2.9 Missing Data and Modeling Tools 111

3.3 Stages of Data Preparation 112

3.3.1 Stage 1: Accessing the Data 112

3.3.2 Stage 2: Auditing the Data 113

3.3.3 Stage 3: Enhancing and Enriching the Data 114

3.3.4 Stage 4: Looking for Sampling Bias 114

3.3.5 Stage 5: Determining Data Structure (Super-, Macro-, and Micro-) 115

3.3.6 Stage 6: Building the PIE 116

3.3.7 Stage 7: Surveying the Data 121

3.3.8 Stage 8: Modeling the Data 122

3.4 And the Result Is . . . ? 122

Chapter 4

Getting the Data: Basic Preparation 125

4.1 Data Discovery 127

4.1.1 Data Access Issues 127

4.2 Data Characterization 129

4.2.1 Detail/Aggregation Level (Granularity) 129

4.2.2 Consistency 131

4.2.3 Pollution 132

4.2.4 Objects 133

4.2.5 Relationship 133

4.2.6 Domain 133

4.2.7 Defaults 134

4.2.8 Integrity 134

4.2.9 Concurrency 135

4.2.10 Duplicate or Redundant Variables 135

4.3 Data Set Assembly 135

4.3.1 Reverse Pivoting 136

4.3.2 Feature Extraction 137

4.3.3 Physical or Behavioral Data Sets 138

4.3.4 Explanatory Structure 138

4.3.5 Data Enhancement or Enrichment 139

4.3.6 Sampling Bias 140

4.4 Example 1: CREDIT 141

4.4.1 Looking at the Variables 141

4.4.2 Relationships between Variables 146

4.5 Example 2: SHOE 149

4.5.1 Looking at the Variables 149

4.5.2 Relationships between Variables 150

4.6 The Data Assay 151

Chapter 5
Sampling, Variability, and Confidence 155

5.1 **Sampling, or First Catch Your Hare! 155**
- 5.1.1 How Much Data? 155
- 5.1.2 Variability 156
- 5.1.3 Converging on a Representative Sample 159
- 5.1.4 Measuring Variability 162
- 5.1.5 Variability and Deviation 162

5.2 **Confidence 166**

5.3 **Variability of Numeric Variables 167**
- 5.3.1 Variability and Sampling 168
- 5.3.2 Variability and Convergence 168

5.4 **Variability and Confidence in Alpha Variables 170**
- 5.4.1 Ordering and Rate of Discovery 171

5.5 **Measuring Confidence 172**
- 5.5.1 Modeling and Confidence with the Whole Population 172
- 5.5.2 Testing for Confidence 173
- 5.5.3 Confidence Tests and Variability 176

5.6 **Confidence in Capturing Variability 178**
- 5.6.1 A Brief Introduction to the Normal Distribution 178
- 5.6.2 Normally Distributed Probabilities 180
- 5.6.3 Capturing Normally Distributed Probabilities: An Example 181
- 5.6.4 Capturing Confidence, Capturing Variance 182

5.7 **Problems and Shortcomings of Taking Samples Using Variability 184**
- 5.7.1 Missing Values 184
- 5.7.2 Constants (Variables with Only One Value) 185
- 5.7.3 Problems with Sampling 185
- 5.7.4 Monotonic Variable Detection 186
- 5.7.5 Interstitial Linearity 187
- 5.7.6 Rate of Discovery 187

5.8 **Confidence and Instance Count 188**

5.9 **Summary 188**

Supplemental Material 189

Confidence Samples 189

Chapter 6
Handling Nonnumerical Variables 191

6.1 Representing Alphas and Remapping 192
6.1.1 One-of-*n* Remapping 193
6.1.2 *m*-of-*n* Remapping 194
6.1.3 Remapping to Eliminate Ordering 195
6.1.4 Remapping One-to-Many Patterns, or Ill-Formed Problems 196
6.1.5 Remapping Circular Discontinuity 200

6.2 State Space 202
6.2.1 Unit State Space 202
6.2.2 Pythagoras in State Space 204
6.2.3 Position in State Space 204
6.2.4 Neighbors and Associates 205
6.2.5 Density and Sparsity 206
6.2.6 Nearby and Distant Nearest Neighbors 211
6.2.7 Normalizing Measured Point Separation 211
6.2.8 Contours, Peaks, and Valleys 213
6.2.9 Mapping State Space 213
6.2.10 Objects in State Space 213
6.2.11 Phase Space 214
6.2.12 Mapping Alpha Values 215
6.2.13 Location, Location, Location! 216
6.2.14 Numerics, Alphas, and the Montreal Canadiens 216

6.3 Joint Distribution Tables 222
6.3.1 Two-Way Tables 223
6.3.2 More Values, More Variables, and Meaning of the Numeration 228
6.3.3 Dealing with Low-Frequency Alpha Labels and Other Problems 229

6.4 Dimensionality 230
6.4.1 Multidimensional Scaling 230
6.4.2 Squashing a Triangle 231
6.4.3 Projecting Alpha Values 234
6.4.4 Scree Plots 234

6.5 Practical Consideration—Implementing Alpha Numeration in the Demonstration Code 235
6.5.1 Implementing Neighborhoods 235
6.5.2 Implementing Numeration in All Alpha Data Sets 237
6.5.3 Implementing Dimensionality Reduction for Variables 237

6.6 Summary 238

Chapter 7

Normalizing and Redistributing Variables 239

7.1 Normalizing a Variable's Range 240

7.1.1 Review of Data Preparation and Modeling (Training, Testing, and Execution) 241

7.1.2 The Nature and Scope of the Out-of-Range Values Problem 242

7.1.3 Discovering the Range of Values When Building the PIE 243

7.1.4 Out-of-Range Values When Training 247

7.1.5 Out-of-Range Values When Testing 249

7.1.6 Out-of-Range Values When Executing 250

7.1.7 Scaling Transformations 251

7.1.8 Softmax Scaling 257

7.1.9 Normalizing Ranges 258

7.2 Redistributing Variable Values 259

7.2.1 The Nature of Distributions 259

7.2.2 Distributive Difficulties 260

7.2.3 Adjusting Distributions 261

7.2.4 Modified Distributions 266

7.3 Summary 269

Supplemental Material 271

The Logistic Function 271

Modifying the Linear Part of the Logistic Function Range 274

Chapter 8

Replacing Missing and Empty Values 275

8.1 Retaining Information about Missing Values 275

8.1.1 Missing-Value Patterns 276

8.1.2 Capturing Patterns 277

8.2 Replacing Missing Values 278

8.2.1 Unbiased Estimators 279

8.2.2 Variability Relationships 279

8.2.3 Relationships between Variables 282

8.2.4 Preserving Between-Variable Relationships 284

8.3 Summary 285

Supplemental Material 286

Using Regression to Find Least Information-Damaging Missing Values 286

Alternative Methods of Missing-Value Replacement 294

Chapter 9

Series Variables 299

9.1 **Here There Be Dragons! 300**

9.2 **Types of Series 300**

→9.3 **Describing Series Data 301**

9.3.1 Constructing a Series 302

9.3.2 Features of a Series 302

9.3.3 Describing a Series—Fourier 303

9.3.4 Describing a Series—Spectrum 307

9.3.5 Describing a Series—Trend, Seasonality, Cycles, Noise 314

9.3.6 Describing a Series—Autocorrelation 316

9.4 **Modeling Series Data 320**

→9.5 **Repairing Series Data Problems 320**

9.5.1 Missing Values 320

9.5.2 Outliers 322

9.5.3 Nonuniform Displacement 322

9.5.4 Trend 323

→9.6 **Tools 325**

9.6.1 Filtering 325

9.6.2 Moving Averages 326

9.6.3 Smoothing 1—PVM Smoothing 333

9.6.4 Smoothing 2—Median Smoothing, Resmoothing, and Hanning 333

9.6.5 Extraction 335

9.6.6 Differencing 336

9.7 **Other Problems 339**

9.7.1 Numerating Alpha Values 341

9.7.2 Distribution 341

9.7.3 Normalization 344

9.8 **Preparing Series Data 344**

9.8.1 Looking at the Data 346

9.8.2 Signposts on the Rocky Road 346

9.9 **Implementation Notes 348**

Chapter 10

Preparing the Data Set 351

10.1 **Using Sparsely Populated Variables 351**

10.1.1 Increasing Information Density Using Sparsely Populated Variables 352

10.1.2 Binning Sparse Numerical Values 353
10.1.3 Present-Value Patterns (PVPs) 353

10.2 Problems with High-Dimensionality Data Sets 355

10.2.1 Information Representation 357
10.2.2 Representing High-Dimensionality Data in Fewer Dimensions 358

10.3 Introducing the Neural Network 360

10.3.1 Training a Neural Network 361
10.3.2 Neurons 362
10.3.3 Reshaping the Logistic Curve 363
10.3.4 Single-Input Neurons 363
10.3.5 Multiple-Input Neurons 366
10.3.6 Networking Neurons to Estimate a Function 368
10.3.7 Network Learning 368
10.3.8 Network Prediction—Hidden Layer 371
10.3.9 Network Prediction—Output Layer 371
10.3.10 Stochastic Network Performance 372
10.3.11 Network Architecture 1— The Autoassociative Network 373
10.3.12 Network Architecture 2— The Sparsely Connected Network 375

10.4 Compressing Variables 376

10.4.1 Using Compressed Dimensionality Data 376

10.5 Removing Variables 378

10.5.1 Estimating Variable Importance 1: What Doesn't Work 379
10.5.2 Estimating Variable Importance 2: Clues 379
10.5.3 Estimating Variable Importance 3: Configuring and Training the Network 380

10.6 How Much Data Is Enough? 383

10.6.1 Joint Distribution 384
10.6.2 Capturing Joint Variability 390
10.6.3 Degrees of Freedom 391

10.7 Beyond Joint Distribution 392

10.7.1 Enhancing the Data Set 393
10.7.2 Data Sets in Perspective 396

10.8 Implementation Notes 396

10.8.1 Collapsing Extremely Sparsely Populated Variables 397
10.8.2 Reducing Excessive Dimensionality 397
10.8.3 Measuring Variable Importance 398
10.8.4 Feature Enhancement 398

10.9 Where Next? 399

Chapter 11
The Data Survey 401

11.1 Introduction to the Data Survey 402

11.2 Information and Communication 403

11.2.1 Measuring Information: Signals and Dictionaries 405
11.2.2 Measuring Information: Signals 406
11.2.3 Measuring Information: Bits of Information 407
11.2.4 Measuring Information: Surprise 410
11.2.5 Measuring Information: Entropy 411
11.2.6 Measuring Information: Dictionaries 412

11.3 Mapping Using Entropy 414

11.3.1 Whole Data Set Entropy 416
11.3.2 Conditional Entropy between Inputs and Outputs 417
11.3.3 Mutual Information 420
11.3.4 Other Survey Uses for Entropy and Information 420
11.3.5 Looking for Information 421

11.4 Identifying Problems with a Data Survey 423

11.4.1 Confidence and Sufficient Data 424
11.4.2 Detecting Sparsity 426
11.4.3 Manifold Definition 427

11.5 Clusters 435

11.6 Sampling Bias 436

11.7 Making the Data Survey 439

11.8 Novelty Detection 442

11.9 Other Directions 443

Supplemental Material 446

Entropic Analysis—Example 446
Surveying Data Sets 451

Chapter 12
Using Prepared Data 483

12.1 Modeling Data 485

12.1.1 Assumptions 485
12.1.2 Models 485
12.1.3 Data Mining vs. Exploratory Data Analysis 486

12.2 Characterizing Data 489

12.2.1 Decision Trees 490
12.2.2 Clusters 491
12.2.3 Nearest Neighbor 492
12.2.4 Neural Networks and Regression 493

12.3 Prepared Data and Modeling Algorithms 494

12.3.1 Neural Networks and the CREDIT Data Set 494
12.3.2 Decision Trees and the CREDIT Data Set 499

12.4 Practical Use of Data Preparation and Prepared Data 500

12.5 Looking at Present Modeling Tools and Future Directions 501

12.5.1 Near Future 503
12.5.2 Farther Out 504

Appendix
Using the Demonstration Code on the CD-ROM 505

Further Reading 509

Index 513

About the Author 537

About the CD-ROM 539