Preprocessing and Cleaning the Data ........................................ 23
2.5 Predictive Power and Overfitting ........................................ 27
Creation and Use of Data Partitions ........................................ 28
Overfitting ............................................................................. 30
2.6 Building a Predictive Model with XLMiner ............................. 32
Predicting Home Values in the West Roxbury Neighborhood ....... 32
Modeling Process .................................................................... 34
2.7 Using Excel for Data Mining .................................................. 41
2.8 Automating Data Mining Solutions ....................................... 42
Data Mining Software Tools: the State of the Market (by Herb Edelstein) ................................................................. 43
Problems ................................................................................... 47

PART II DATA EXPLORATION AND DIMENSION REDUCTION

CHAPTER 3 Data Visualization .................................................... 52

3.1 Uses of Data Visualization .................................................... 52
3.2 Data Examples ................................................................. 54
Example 1: Boston Housing Data .............................................. 54
Example 2: Ridership on Amtrak Trains .................................... 55
3.3 Basic Charts: Bar Charts, Line Graphs, and Scatter Plots ....... 55
Distribution Plots: Boxplots and Histograms ............................ 57
Heatmaps: Visualizing Correlations and Missing Values .......... 60
3.4 Multidimensional Visualization ............................................ 62
Adding Variables: Color, Size, Shape, Multiple Panels, and Animation .................................................. 62
Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering .................................................. 64
Reference: Trend Line and Labels ............................................. 67
Scaling up to Large Datasets ..................................................... 67
Multivariate Plot: Parallel Coordinates Plot ............................. 69
Interactive Visualization ............................................................ 70
3.5 Specialized Visualizations .................................................... 73
Visualizing Networked Data ..................................................... 73
Visualizing Hierarchical Data: Treemaps ................................. 75
Visualizing Geographical Data: Map Charts ............................. 76
3.6 Summary: Major Visualizations and Operations, by Data Mining Goal .................................................. 78
Prediction .............................................................................. 78
Classification .......................................................................... 78
Time Series Forecasting ............................................................ 78
Unsupervised Learning .............................................................. 79
Problems ................................................................................... 80
CHAPTER 4  Dimension Reduction 82

4.1 Introduction .................................. 82
4.2 Curse of Dimensionality ...................... 83
4.3 Practical Considerations ..................... 83
    Example 1: House Prices in Boston ......... 84
4.4 Data Summaries .............................. 84
    Summary Statistics .......................... 86
    Pivot Tables ................................ 87
4.5 Correlation Analysis ......................... 88
4.6 Reducing the Number of Categories in
    Categorical Variables ........................ 89
4.7 Converting a Categorical Variable to a Numerical Variable 90
4.8 Principal Components Analysis ............... 90
    Example 2: Breakfast Cereals ................ 91
    Principal Components ....................... 95
    Normalizing the Data ....................... 97
    Using Principal Components for Classification and Prediction 100
4.9 Dimension Reduction Using Regression Models .... 100
4.10 Dimension Reduction Using Classification and
    Regression Trees ............................ 101
Problems ..................................... 102

PART III  PERFORMANCE EVALUATION

CHAPTER 5  Evaluating Predictive Performance 106

5.1 Introduction .................................. 107
5.2 Evaluating Predictive Performance .......... 107
    Benchmark: The Average .................... 108
    Prediction Accuracy Measures ............... 108
    Comparing Training and Validation Performance 109
    Lift Chart .................................. 110
5.3 Judging Classifier Performance ............... 112
    Benchmark: The Naive Rule .................. 112
    Class Separation ............................ 112
    The Classification Matrix .................... 113
    Using the Validation Data ................... 114
    Accuracy Measures .......................... 115
    Propensities and Cutoff for Classification 115
    Performance in Unequal Importance of Classes 119
    Asymmetric Misclassification Costs .......... 121
    Generalization to More Than Two Classes .... 124
5.4 Judging Ranking Performance ................. 124
    Lift Charts for Binary Data ................ 125
Using the Cutoff Probability Method .................................................. 172
Practical Difficulty with the Complete (Exact) Bayes Procedure .......... 172
Solution: Naive Bayes ................................................................. 173
Example 2: Predicting Fraudulent Financial Reports, Two Predictors .......... 175
Example 3: Predicting Delayed Flights ............................................. 176
8.3 Advantages and Shortcomings of the Naive Bayes Classifier .............. 180
Problems ....................................................................................... 184

CHAPTER 9 Classification and Regression Trees .................................. 186
9.1 Introduction ............................................................................. 187
9.2 Classification Trees .................................................................. 188
Recursive Partitioning ................................................................. 188
Example 1: Riding Mowers ............................................................ 189
Measures of Impurity .................................................................. 191
Tree Structure ............................................................................. 195
Classifying a New Observation ...................................................... 195
9.3 Evaluating the Performance of a Classification Tree ....................... 196
Example 2: Acceptance of Personal Loan ........................................ 196
9.4 Avoiding Overfitting ................................................................. 199
Stopping Tree Growth: CHAID ...................................................... 199
Pruning the Tree .......................................................................... 201
9.5 Classification Rules from Trees .................................................. 206
9.6 Classification Trees for More Than two Classes .............................. 207
9.7 Regression Trees ..................................................................... 207
Prediction .................................................................................... 207
Measuring Impurity ..................................................................... 208
Evaluating Performance ............................................................... 208
9.8 Advantages, Weaknesses and Extensions ...................................... 209
9.9 Improving Prediction: Multiple Trees ......................................... 210
Problems ....................................................................................... 214

CHAPTER 10 Logistic Regression .......................................................... 218
10.1 Introduction ............................................................................ 219
10.2 The Logistic Regression Model ................................................ 220
Example: Acceptance of Personal Loan ........................................... 222
Model with a Single Predictor ....................................................... 223
Estimating the Logistic Model from Data: Computing Parameter Estimates ................................................. 225
Interpreting Results in Terms of Odds (for a Profiling Goal) .............. 227
10.3 Evaluating Classification Performance ....................................... 229
Variable Selection ....................................................................... 231
**CONTENTS**

10.4 Example of Complete Analysis: Predicting
   Delayed Flights ........................................... 231
   Data Preprocessing ....................................... 234
   Model Fitting and Estimation ............................ 234
   Model Interpretation ..................................... 235
   Model Performance ....................................... 236
   Variable Selection ....................................... 237

10.5 Appendix: Logistic Regression for Profiling ......... 240
   Appendix A: Why Linear Regression Is Problematic for a
   Categorical Response ...................................... 240
   Appendix B: Evaluating Explanatory Power ............. 241
   Appendix C: Logistic Regression for More Than Two Classes
   Problems ................................................... 244

**CHAPTER 11 Neural Nets**

11.1 Introduction ............................................. 250
11.2 Concept and Structure of a Neural Network .......... 251
11.3 Fitting a Network to Data .............................. 252
   Example 1: Tiny Dataset .................................. 252
   Computing Output of Nodes .............................. 253
   Preprocessing the Data ................................... 256
   Training the Model ....................................... 257
   Example 2: Classifying Accident Severity .............. 261
   Avoiding Overfitting ..................................... 263
   Using the Output for Prediction and Classification ..... 264
11.4 Required User Input ..................................... 266
11.5 Exploring the Relationship Between Predictors
   and Response ............................................. 268
   Unsupervised Feature Extraction and Deep Learning ... 270
11.6 Advantages and Weaknesses of Neural Networks ....... 268
Problems ................................................... 271

**CHAPTER 12 Discriminant Analysis**

12.1 Introduction ............................................. 273
   Example 1: Riding Mowers ................................ 274
   Example 2: Personal Loan Acceptance .................. 275
12.2 Distance of an Observation from a Class .............. 275
12.3 Fisher’s Linear Classification Functions .............. 278
12.4 Classification Performance of Discriminant
   Analysis ................................................... 281
12.5 Prior Probabilities ..................................... 282
12.6 Unequal Misclassification Costs ....................... 283
12.7 Classifying More Than Two Classes .................... 284
### Example 3: Medical Dispatch to Accident Scenes

\[ 284 \]

### 12.8 Advantages and Weaknesses

\[ 286 \]

### Problems

\[ 289 \]

---

**CHAPTER 13**

**Combining Methods: Ensembles and Uplift Modeling**

\[ 292 \]

13.1 Ensembles

- Why Ensembles Can Improve Predictive Power
- Simple Averaging
- Bagging
- Boosting
- Advantages and Weaknesses of Ensembles

\[ 297 \]

13.2 Uplift (Persuasion) Modeling

- A-B Testing
- Uplift
- Gathering the Data
- A Simple Model
- Modeling Individual Uplift
- Using the Results of an Uplift Model

\[ 303 \]

13.3 Summary

\[ 303 \]

Problems

\[ 304 \]

---

**PART V**

**MINING RELATIONSHIPS AMONG RECORDS**

**CHAPTER 14**

**Association Rules and Collaborative Filtering**

\[ 308 \]

14.1 Association Rules

- Discovering Association Rules in Transaction Databases
- Example 1: Synthetic Data on Purchases of Phone Faceplates
- Generating Candidate Rules
- The Apriori Algorithm
- Selecting Strong Rules
- Data Format
- The Process of Rule Selection
- Interpreting the Results
- Rules and Chance
- Example 2: Rules for Similar Book Purchases

\[ 320 \]

14.2 Collaborative Filtering

- Data Type and Format
- Example 3: Netflix Prize Contest
- User-Based Collaborative Filtering: “People Like You”
- Item-Based Collaborative Filtering
- Advantages and Weaknesses of Collaborative Filtering
- Collaborative Filtering vs. Association Rules

\[ 328 \]

14.3 Summary

\[ 330 \]

Problems

\[ 332 \]
CHAPTER 15 Cluster Analysis

15.1 Introduction ........................................ 337
   Example: Public Utilities .......................... 338
15.2 Measuring Distance Between Two Observations ........ 340
   Euclidean Distance .............................. 340
   Normalizing Numerical Measurements .......... 341
   Other Distance Measures for Numerical Data .... 341
   Distance Measures for Categorical Data ....... 343
   Distance Measures for Mixed Data ............. 344
15.3 Measuring Distance Between Two Clusters ............. 345
   Minimum Distance .............................. 345
   Maximum Distance .............................. 345
   Average Distance .............................. 345
   Centroid Distance .............................. 345
15.4 Hierarchical (Agglomerative) Clustering .............. 347
   Single Linkage .................................. 348
   Complete Linkage .............................. 348
   Average Linkage (in XLMiner: “Group Average Linkage”) . 349
   Centroid Linkage .............................. 349
   Ward’s Method .................................. 349
   Dendrograms: Displaying Clustering Process and Results . 350
   Validating Clusters ............................. 352
   Limitations of Hierarchical Clustering ......... 353
15.5 Non-hierarchical Clustering: The k-Means Algorithm .... 354
   Initial Partition into k Clusters ............... 356
Problems ........................................... 360

PART VI FORECASTING TIME SERIES

CHAPTER 16 Handling Time Series .......................... 364

16.1 Introduction ........................................ 364
16.2 Descriptive vs. Predictive Modeling .................. 366
16.3 Popular Forecasting Methods in Business ............ 366
   Combining Methods ............................. 366
16.4 Time Series Components .......................... 367
   Example: Ridership on Amtrak Trains ............ 367
16.5 Data Partitioning and Performance Evaluation ........ 371
   Benchmark Performance: Naive Forecasts ........ 372
   Generating Future Forecasts .................... 372
Problems ........................................... 374
CHAPTER 17  Regression-Based Forecasting  
17.1 A Model with Trend  
  Linear Trend  
  Exponential Trend  
  Polynomial Trend  
17.2 A Model with Seasonality  
17.3 A Model with Trend and Seasonality  
17.4 Autocorrelation and ARIMA Models  
  Computing Autocorrelation  
  Improving Forecasts by Integrating Autocorrelation  
  Evaluating Predictability  
Problems  

CHAPTER 18  Smoothing Methods  
18.1 Introduction  
18.2 Moving Average  
  Centered Moving Average for Visualization  
  Trailing Moving Average for Forecasting  
  Choosing Window Width (w)  
18.3 Simple Exponential Smoothing  
  Choosing Smoothing Parameter \( \alpha \)  
  Relation between Moving Average and Simple Exponential Smoothing  
18.4 Advanced Exponential Smoothing  
  Series with a Trend  
  Series with a Trend and Seasonality  
  Series with Seasonality (No Trend)  
Problems  

PART VII  DATA ANALYTICS  
CHAPTER 19  Social Network Analytics  
19.1 Introduction  
19.2 Directed vs. Undirected Networks  
19.3 Visualizing and Analyzing Networks  
  Graph Layout  
  Adjacency List  
  Adjacency Matrix  
  Using Network Data in Classification and Prediction  
19.4 Social Data Metrics and Taxonomy  
  Node-Level Centrality Metrics  
  Egocentric Network  

### Network Metrics

- 19.5 Using Network Metrics in Prediction and Classification
  - Link Prediction
  - Entity Resolution
  - Collaborative Filtering

#### Advantages and Disadvantages

Problems

### CHAPTER 20  Text Mining

20.1 Introduction

20.2 The Spreadsheet Representation of Text: “Bag-of-Words”

20.3 Bag-of-Words vs. Meaning Extraction at Document Level

20.4 Preprocessing the Text
  - Tokenization
  - Text Reduction
  - Presence/Absence vs. Frequency
  - Term Frequency—Inverse Document Frequency (TF-IDF)
  - From Terms to Concepts: Latent Semantic Indexing
  - Extracting Meaning

20.5 Implementing Data Mining Methods

20.6 Example: Online Discussions on Autos and Electronics
  - Importing and Labeling the Records
  - Tokenization
  - Text Processing and Reduction
  - Producing a Concept Matrix
  - Labeling the Documents
  - Fitting a Model
  - Prediction

20.7 Summary

Problems

### PART VIII  CASES

### CHAPTER 21  Cases

21.1 Charles Book Club
  - The Book Industry
  - Database Marketing at Charles
  - Data Mining Techniques
  - Assignment

21.2 German Credit
  - Background
  - Data
  - Assignment

21.3 Tayko Software Cataloger