

Contents

Preface • xi

Author Bios • xv

1 Introduction • 1

- 1.1 A Simple Example • 4
- 1.2 Important Concepts • 7
- 1.3 A More Complex Example • 15
- 1.4 Feature Selection • 17
- 1.5 An Outline of the Book • 18
- 1.6 Computing • 20

2 Illustrative Example: Predicting Risk of Ischemic Stroke • 21

- 2.1 Splitting • 23
- 2.2 Preprocessing • 23
- 2.3 Exploration • 26
- 2.4 Predictive Modeling across Sets • 30
- 2.5 Other Considerations • 34
- 2.6 Computing • 34

3 A Review of the Predictive Modeling Process • 35

- 3.1 Illustrative Example: OkCupid Profile Data • 35
- 3.2 Measuring Performance • 36
- 3.3 Data Splitting • 46
- 3.4 Resampling • 47
- 3.5 Tuning Parameters and Overfitting • 56
- 3.6 Model Optimization and Tuning • 57
- 3.7 Comparing Models Using the Training Set • 61
- 3.8 Feature Engineering without Overfitting • 62
- 3.9 Summary • 64
- 3.10 Computing • 64

4 Exploratory Visualizations • 65

- 4.1 Introduction to the Chicago Train Ridership Data • 66

- 4.2 Visualizations for Numeric Data: Exploring Train Ridership Data • 69
 - 4.3 Visualizations for Categorical Data: Exploring the OkCupid Data • 83
 - 4.4 Postmodeling Exploratory Visualizations • 88
 - 4.5 Summary • 92
 - 4.6 Computing • 92
- 5 Encoding Categorical Predictors • 93**
- 5.1 Creating Dummy Variables for Unordered Categories • 94
 - 5.2 Encoding Predictors with Many Categories • 96
 - 5.3 Approaches for Novel Categories • 102
 - 5.4 Supervised Encoding Methods • 102
 - 5.5 Encodings for Ordered Data • 107
 - 5.6 Creating Features from Text Data • 109
 - 5.7 Factors versus Dummy Variables in Tree-Based Models • 114
 - 5.8 Summary • 119
 - 5.9 Computing • 120
- 6 Engineering Numeric Predictors • 121**
- 6.1 1:1 Transformations • 122
 - 6.2 1:Many Transformations • 126
 - 6.3 Many:Many Transformations • 133
 - 6.4 Summary • 154
 - 6.5 Computing • 155
- 7 Detecting Interaction Effects • 157**
- 7.1 Guiding Principles in the Search for Interactions • 161
 - 7.2 Practical Considerations • 164
 - 7.3 The Brute-Force Approach to Identifying Predictive Interactions • 165
 - 7.4 Approaches when Complete Enumeration Is Practically Impossible • 172
 - 7.5 Other Potentially Useful Tools • 184
 - 7.6 Summary • 185
 - 7.7 Computing • 186
- 8 Handling Missing Data • 187**
- 8.1 Understanding the Nature and Severity of Missing Information • 189
 - 8.2 Models that Are Resistant to Missing Values • 195
 - 8.3 Deletion of Data • 196
 - 8.4 Encoding Missingness • 197
 - 8.5 Imputation Methods • 198
 - 8.6 Special Cases • 203
 - 8.7 Summary • 203
 - 8.8 Computing • 204
- 9 Working with Profile Data • 205**
- 9.1 Illustrative Data: Pharmaceutical Manufacturing Monitoring • 209
 - 9.2 What Are the Experimental Unit and the Unit of Prediction? • 210
 - 9.3 Reducing Background • 214
 - 9.4 Reducing Other Noise • 215
 - 9.5 Exploiting Correlation • 217

- 9.6 Impacts of Data Processing on Modeling • 219
- 9.7 Summary • 224
- 9.8 Computing • 225

10 Feature Selection Overview • 227

- 10.1 Goals of Feature Selection • 227
- 10.2 Classes of Feature Selection Methodologies • 228
- 10.3 Effect of Irrelevant Features • 232
- 10.4 Overfitting to Predictors and External Validation • 235
- 10.5 A Case Study • 238
- 10.6 Next Steps • 240
- 10.7 Computing • 240

11 Greedy Search Methods • 241

- 11.1 Illustrative Data: Predicting Parkinson's Disease • 241
- 11.2 Simple Filters • 241
- 11.3 Recursive Feature Elimination • 248
- 11.4 Stepwise Selection • 252
- 11.5 Summary • 254
- 11.6 Computing • 255

12 Global Search Methods • 257

- 12.1 Naive Bayes Models • 257
- 12.2 Simulated Annealing • 260
- 12.3 Genetic Algorithms • 270
- 12.4 Test Set Results • 280
- 12.5 Summary • 281
- 12.6 Computing • 282

Bibliography • 283

Index • 295