

# Contents

|          |                                              |          |
|----------|----------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                          | <b>1</b> |
| 1.1      | Prediction Versus Interpretation             | 4        |
| 1.2      | Key Ingredients of Predictive Models         | 5        |
| 1.3      | Terminology                                  | 6        |
| 1.4      | Example Data Sets and Typical Data Scenarios | 7        |
| 1.5      | Overview                                     | 14       |
| 1.6      | Notation                                     | 15       |

## Part I General Strategies

|          |                                                         |           |
|----------|---------------------------------------------------------|-----------|
| <b>2</b> | <b>A Short Tour of the Predictive Modeling Process</b>  | <b>19</b> |
| 2.1      | Case Study: Predicting Fuel Economy                     | 19        |
| 2.2      | Themes                                                  | 24        |
| 2.3      | Summary                                                 | 26        |
| <b>3</b> | <b>Data Pre-processing</b>                              | <b>27</b> |
| 3.1      | Case Study: Cell Segmentation in High-Content Screening | 28        |
| 3.2      | Data Transformations for Individual Predictors          | 30        |
| 3.3      | Data Transformations for Multiple Predictors            | 33        |
| 3.4      | Dealing with Missing Values                             | 41        |
| 3.5      | Removing Predictors                                     | 43        |
| 3.6      | Adding Predictors                                       | 47        |
| 3.7      | Binning Predictors                                      | 49        |
| 3.8      | Computing                                               | 51        |
|          | Exercises                                               | 58        |
| <b>4</b> | <b>Over-Fitting and Model Tuning</b>                    | <b>61</b> |
| 4.1      | The Problem of Over-Fitting                             | 62        |
| 4.2      | Model Tuning                                            | 64        |
| 4.3      | Data Splitting                                          | 67        |
| 4.4      | Resampling Techniques                                   | 69        |

|     |                                        |    |
|-----|----------------------------------------|----|
| 4.5 | Case Study: Credit Scoring .....       | 73 |
| 4.6 | Choosing Final Tuning Parameters ..... | 74 |
| 4.7 | Data Splitting Recommendations .....   | 77 |
| 4.8 | Choosing Between Models .....          | 78 |
| 4.9 | Computing .....                        | 80 |
|     | Exercises .....                        | 89 |

## Part II Regression Models

|     |                                                                         |     |
|-----|-------------------------------------------------------------------------|-----|
| 5   | Measuring Performance in Regression Models .....                        | 95  |
| 5.1 | Quantitative Measures of Performance .....                              | 95  |
| 5.2 | The Variance-Bias Trade-off .....                                       | 97  |
| 5.3 | Computing .....                                                         | 98  |
| 6   | Linear Regression and Its Cousins .....                                 | 101 |
| 6.1 | Case Study: Quantitative Structure-Activity Relationship Modeling ..... | 102 |
| 6.2 | Linear Regression .....                                                 | 105 |
| 6.3 | Partial Least Squares .....                                             | 112 |
| 6.4 | Penalized Models .....                                                  | 122 |
| 6.5 | Computing .....                                                         | 128 |
|     | Exercises .....                                                         | 137 |
| 7   | Nonlinear Regression Models .....                                       | 141 |
| 7.1 | Neural Networks .....                                                   | 141 |
| 7.2 | Multivariate Adaptive Regression Splines .....                          | 145 |
| 7.3 | Support Vector Machines .....                                           | 151 |
| 7.4 | K-Nearest Neighbors .....                                               | 159 |
| 7.5 | Computing .....                                                         | 161 |
|     | Exercises .....                                                         | 168 |
| 8   | Regression Trees and Rule-Based Models .....                            | 173 |
| 8.1 | Basic Regression Trees .....                                            | 175 |
| 8.2 | Regression Model Trees .....                                            | 184 |
| 8.3 | Rule-Based Models .....                                                 | 190 |
| 8.4 | Bagged Trees .....                                                      | 192 |
| 8.5 | Random Forests .....                                                    | 198 |
| 8.6 | Boosting .....                                                          | 203 |
| 8.7 | Cubist .....                                                            | 208 |
| 8.8 | Computing .....                                                         | 212 |
|     | Exercises .....                                                         | 218 |

|      |                                                                    |     |
|------|--------------------------------------------------------------------|-----|
| 9    | A Summary of Solubility Models .....                               | 221 |
| 10   | Case Study: Compressive Strength of Concrete Mixtures .....        | 225 |
| 10.1 | Model Building Strategy .....                                      | 229 |
| 10.2 | Model Performance .....                                            | 230 |
| 10.3 | Optimizing Compressive Strength .....                              | 233 |
| 10.4 | Computing .....                                                    | 236 |
|      | Part III Classification Models                                     |     |
| 11   | Measuring Performance in Classification Models .....               | 247 |
| 11.1 | Class Predictions .....                                            | 247 |
| 11.2 | Evaluating Predicted Classes .....                                 | 254 |
| 11.3 | Evaluating Class Probabilities .....                               | 262 |
| 11.4 | Computing .....                                                    | 266 |
| 12   | Discriminant Analysis and Other Linear Classification Models ..... | 275 |
| 12.1 | Case Study: Predicting Successful Grant Applications .....         | 275 |
| 12.2 | Logistic Regression .....                                          | 282 |
| 12.3 | Linear Discriminant Analysis .....                                 | 287 |
| 12.4 | Partial Least Squares Discriminant Analysis .....                  | 297 |
| 12.5 | Penalized Models .....                                             | 302 |
| 12.6 | Nearest Shrunken Centroids .....                                   | 306 |
| 12.7 | Computing .....                                                    | 308 |
|      | Exercises .....                                                    | 326 |
| 13   | Nonlinear Classification Models .....                              | 329 |
| 13.1 | Nonlinear Discriminant Analysis .....                              | 329 |
| 13.2 | Neural Networks .....                                              | 333 |
| 13.3 | Flexible Discriminant Analysis .....                               | 338 |
| 13.4 | Support Vector Machines .....                                      | 343 |
| 13.5 | K-Nearest Neighbors .....                                          | 350 |
| 13.6 | Naïve Bayes .....                                                  | 353 |
| 13.7 | Computing .....                                                    | 358 |
|      | Exercises .....                                                    | 366 |
| 14   | Classification Trees and Rule-Based Models .....                   | 369 |
| 14.1 | Basic Classification Trees .....                                   | 370 |
| 14.2 | Rule-Based Models .....                                            | 383 |
| 14.3 | Bagged Trees .....                                                 | 385 |
| 14.4 | Random Forests .....                                               | 386 |
| 14.5 | Boosting .....                                                     | 389 |
| 14.6 | C5.0 .....                                                         | 392 |

|                                                             |            |
|-------------------------------------------------------------|------------|
| 4.7 Comparing Two Encodings of Categorical Predictors ..... | 400        |
| 4.8 Computing .....                                         | 400        |
| Exercises .....                                             | 411        |
| <b>A Summary of Grant Application Models .....</b>          | <b>415</b> |
| <b>Remedies for Severe Class Imbalance .....</b>            | <b>419</b> |
| 6.1 Case Study: Predicting Caravan Policy Ownership .....   | 419        |
| 6.2 The Effect of Class Imbalance .....                     | 420        |
| 6.3 Model Tuning .....                                      | 423        |
| 6.4 Alternate Cutoffs .....                                 | 423        |
| 6.5 Adjusting Prior Probabilities .....                     | 426        |
| 6.6 Unequal Case Weights .....                              | 426        |
| 6.7 Sampling Methods .....                                  | 427        |
| 6.8 Cost-Sensitive Training .....                           | 429        |
| 6.9 Computing .....                                         | 435        |
| Exercises .....                                             | 442        |
| <b>Case Study: Job Scheduling .....</b>                     | <b>445</b> |
| 17.1 Data Splitting and Model Strategy .....                | 450        |
| 17.2 Results .....                                          | 454        |
| 17.3 Computing .....                                        | 457        |
| <b>IV Other Considerations</b>                              |            |
| <b>Measuring Predictor Importance .....</b>                 | <b>463</b> |
| 18.1 Numeric Outcomes .....                                 | 464        |
| 18.2 Categorical Outcomes .....                             | 468        |
| 18.3 Other Approaches .....                                 | 472        |
| 18.4 Computing .....                                        | 478        |
| Exercises .....                                             | 484        |
| <b>An Introduction to Feature Selection .....</b>           | <b>487</b> |
| 19.1 Consequences of Using Non-informative Predictors ..... | 488        |
| 19.2 Approaches for Reducing the Number of Predictors ..... | 490        |
| 19.3 Wrapper Methods .....                                  | 491        |
| 19.4 Filter Methods .....                                   | 499        |
| 19.5 Selection Bias .....                                   | 500        |
| 19.6 Case Study: Predicting Cognitive Impairment .....      | 502        |
| 19.7 Computing .....                                        | 511        |
| Exercises .....                                             | 518        |

|                                                           |            |
|-----------------------------------------------------------|------------|
| <b>20 Factors That Can Affect Model Performance .....</b> | <b>521</b> |
| 20.1 Type III Errors .....                                | 522        |
| 20.2 Measurement Error in the Outcome .....               | 524        |
| 20.3 Measurement Error in the Predictors .....            | 527        |
| 20.4 Discretizing Continuous Outcomes .....               | 531        |
| 20.5 When Should You Trust Your Model's Prediction? ..... | 534        |
| 20.6 The Impact of a Large Sample .....                   | 538        |
| 20.7 Computing .....                                      | 541        |
| Exercises .....                                           | 542        |

## Appendix

|                                                 |            |
|-------------------------------------------------|------------|
| <b>A A Summary of Various Models .....</b>      | <b>549</b> |
| <b>B An Introduction to R .....</b>             | <b>551</b> |
| B.1 Start-Up and Getting Help .....             | 551        |
| B.2 Packages .....                              | 552        |
| B.3 Creating Objects .....                      | 553        |
| B.4 Data Types and Basic Structures .....       | 554        |
| B.5 Working with Rectangular Data Sets .....    | 558        |
| B.6 Objects and Classes .....                   | 560        |
| B.7 R Functions .....                           | 561        |
| B.8 The Three Faces of = .....                  | 562        |
| B.9 The AppliedPredictiveModeling Package ..... | 562        |
| B.10 The caret Package .....                    | 563        |
| B.11 Software Used in this Text .....           | 565        |
| <b>C Interesting Web Sites .....</b>            | <b>567</b> |
| <b>References .....</b>                         | <b>569</b> |
| <b>Indices</b>                                  |            |
| <b>Computing .....</b>                          | <b>591</b> |
| <b>General .....</b>                            | <b>595</b> |