Table of Contents

Preface 1

Chapter 1: A Process for Success 9

- **The process** 10
- **Business understanding** 11
  - Identifying the business objective 12
  - Assessing the situation 13
  - Determining the analytical goals 13
  - Producing a project plan 13
- **Data understanding** 14
- **Data preparation** 14
- **Modeling** 15
- **Evaluation** 16
- **Deployment** 16
- **Algorithm flowchart** 17
- **Summary** 22

Chapter 2: Linear Regression - The Blocking and Tackling of Machine Learning 23

- **Univariate linear regression** 24
  - Business understanding 27
- **Multivariate linear regression** 33
  - Business understanding 33
  - Data understanding and preparation 34
  - Modeling and evaluation 37
- **Other linear model considerations** 51
  - Qualitative features 51
  - Interaction terms 53
- **Summary** 55

Chapter 3: Logistic Regression and Discriminant Analysis 57

- **Classification methods and linear regression** 58
- **Logistic regression** 58
  - Business understanding 59
  - Data understanding and preparation 60
  - Modeling and evaluation 65

The logistic regression model 66
Logistic regression with cross-validation 69
Discriminant analysis overview 72
Discriminant analysis application 75
Multivariate Adaptive Regression Splines (MARS) 78
Model selection 84
Summary 87
Chapter 4: Advanced Feature Selection in Linear Models 89
Regularization in a nutshell 90
Ridge regression 91
LASSO 91
Elastic net 92
Business case 92
Business understanding 92
Data understanding and preparation 93
Modeling and evaluation 99
Best subsets 99
Ridge regression 103
LASSO 108
Elastic net 111
Cross-validation with glmnet 114
Model selection 116
Regularization and classification 117
Logistic regression example 117
Summary 120
Chapter 5: More Classification Techniques - K-Nearest Neighbors and Support Vector Machines 121
K-nearest neighbors 122
Support vector machines 123
Business case 127
Business understanding 127
Data understanding and preparation 128
Modeling and evaluation 134
KNN modeling 134
SVM modeling 139
Model selection 142
Feature selection for SVMs 145
Summary 147
Chapter 6: Classification and Regression Trees 149

An overview of the techniques 150
Understanding the regression trees 150
Classification trees 151
Random forest 152
Gradient boosting 153
Business case 154
Modeling and evaluation 154
Regression tree 154
Classification tree 158
Random forest regression 160
Random forest classification 163
Extreme gradient boosting - classification 167
Model selection 172
Feature Selection with random forests 172
Summary 175
Chapter 7: Neural Networks and Deep Learning 177
Introduction to neural networks 177
Deep learning, a not-so-deep overview 182
Deep learning resources and advanced methods 184
Business understanding 186
Data understanding and preparation 187
Modeling and evaluation 192
An example of deep learning 197
H2O background 198
Data upload to H2O 198
Create train and test datasets 200
Modeling 201
Summary 205
Chapter 8: Cluster Analysis 207
Hierarchical clustering 208
Distance calculations 209
K-means clustering 210
Gower and partitioning around medoids 211
Gower 212
PAM 213
Random forest 213
Business understanding 214
Data understanding and preparation 215
Modeling and evaluation 217

Hierarchical clustering 217
K-means clustering 227
Gower and PAM 231
Random Forest and PAM 233
Summary 235
Chapter 9: Principal Components Analysis 237
An overview of the principal components 238
Rotation 241
Business understanding 243
Data understanding and preparation 244
Modeling and evaluation 246
Component extraction 246
Orthogonal rotation and interpretation 247
Creating factor scores from the components 249
Regression analysis 250
Summary 256
Chapter 10: Market Basket Analysis, Recommendation Engines, and Sequential Analysis 257
An overview of a market basket analysis 258
Business understanding 259
Data understanding and preparation 260
Modeling and evaluation 262
An overview of a recommendation engine 266
User-based collaborative filtering 268
Item-based collaborative filtering 268
Singular value decomposition and principal components analysis 269
Business understanding and recommendations 273
Data understanding, preparation, and recommendations 273
Modeling, evaluation, and recommendations 276
Sequential data analysis 286
Sequential analysis applied 287
Summary 294
Chapter 11: Creating Ensembles and Multiclass Classification 295
Ensembles 296
Business and data understanding 297
Modeling evaluation and selection 298
Multiclass classification 301
Business and data understanding 302

Model evaluation and selection 306
Random forest 307
Ridge regression 309
MLR's ensemble 310
Summary 312
Chapter 12: Time Series and Causality 313
Univariate time series analysis 314
Understanding Granger causality 320
Business understanding 321
Data understanding and preparation 323
Modeling and evaluation 327
Univariate time series forecasting 327
Examining the causality 331
Linear regression 331
Vector autoregression 334
Summary 339
Chapter 13: Text Mining 341
Text mining framework and methods 342
Topic models 344
Other quantitative analyses 345
Business understanding 347
Data understanding and preparation 347
Modeling and evaluation 350
Word frequency and topic models 350
Additional quantitative analysis 355
Summary 364
Chapter 14: R on the Cloud 365
Creating an Amazon Web Services account 366
Launch a virtual machine 368
Start RStudio 372
Summary 374
Appendix A: R Fundamentals 375
Getting R up-and-running 375
Using R 381
Data frames and matrices 386
Creating summary statistics 388
Installing and loading R packages 392
Data manipulation with dplyr 393

Summary 396
Appendix B: Sources 397
Index 399