Table of Contents

Preface 1

Chapter 1: Preparing and Understanding Data 9

Overview 10

Reading the data 11

Handling duplicate observations 13

Descriptive statistics 14

Exploring categorical variables 15

Handling missing values 17

Zero and near-zero variance features 19

Treating the data 21

Correlation and linearity 23

Summary 27

Chapter 2: Linear Regression 29

Univariate linear regression 30

Building a univariate model 33

Reviewing model assumptions 36

Multivariate linear regression 38

Loading and preparing the data 39

Modeling and evaluation – stepwise regression 46

Modeling and evaluation – MARS 52

Reverse transformation of natural log predictions 56

Summary 59

Chapter 3: Logistic Regression 61

Classification methods and linear regression 62

Logistic regression 62

Model training and evaluation 63

Training a logistic regression algorithm 64

Weight of evidence and information value 66

Feature selection 68

Cross-validation and logistic regression 71

Multivariate adaptive regression splines 77

Model comparison 81

Summary 83

Chapter 4: Advanced Feature Selection in Linear Models 85

Regularization overview 86

Ridge regression 87

LASSO 87

Elastic net 88
Data creation 88
Modeling and evaluation 91
Ridge regression 91
LASSO 96
Elastic net 99
Summary 104
Chapter 5: K-Nearest Neighbors and Support Vector Machines 105
K-nearest neighbors 106
Support vector machines 107
Manipulating data 111
Dataset creation 111
Data preparation 114
Modeling and evaluation 117
KNN modeling 117
Support vector machine 125
Summary 131
Chapter 6: Tree-Based Classification 133
An overview of the techniques 134
Understanding a regression tree 134
Classification trees 135
Random forest 136
Gradient boosting 137
Datasets and modeling 138
Classification tree 138
Random forest 144
Extreme gradient boosting – classification 152
Feature selection with random forests 157
Summary 160
Chapter 7: Neural Networks and Deep Learning 161
Introduction to neural networks 162
Deep learning – a not-so-deep overview 167
Deep learning resources and advanced methods 169
Creating a simple neural network 171
Data understanding and preparation 171
Modeling and evaluation 173
An example of deep learning 176
Keras and TensorFlow background 176
Loading the data 177
Creating the model function 178
Model training 180
Summary 182

Chapter 8: Creating Ensembles and Multiclass Methods 183
Ensembles 184
Data understanding 185
Modeling and evaluation 187
Random forest model 187
Creating an ensemble 190
Summary 191
Chapter 9: Cluster Analysis 193
Hierarchical clustering 195
Distance calculations 196
K-means clustering 197
Gower and PAM 197
Gower 198
PAM 199
Random forest 200
Dataset background 201
Data understanding and preparation 201
Modeling 204
Hierarchical clustering 204
K-means clustering 212
Gower and PAM 215
Random forest and PAM 216
Summary 218
Chapter 10: Principal Component Analysis 219
An overview of the principal components 220
Rotation 223
Data 225
Data loading and review 226
Training and testing datasets 229
PCA modeling 231
Component extraction 231
Orthogonal rotation and interpretation 234
Creating scores from the components 235
Regression with MARS 236
Test data evaluation 240
Summary 242
Chapter 11: Association Analysis 243
An overview of association analysis 243
Creating transactional data 245
Data understanding 246
Data preparation 247
Modeling and evaluation 249

Summary 254

Chapter 12: Time Series and Causality 255

Univariate time series analysis 256

Understanding Granger causality 264

Time series data 265

Data exploration 267

Modeling and evaluation 272

Univariate time series forecasting 272

Examining the causality 282

Linear regression 283

Vector autoregression 284

Summary 290

Chapter 13: Text Mining 291

Text mining framework and methods 292

Topic models 294

Other quantitative analysis 295

Data overview 297

Data frame creation 297

Word frequency 299

Word frequency in all addresses 299

Lincoln's word frequency 301

Sentiment analysis 305

N-grams 309

Topic models 311

Classifying text 316

Data preparation 316

LASSO model 318

Additional quantitative analysis 320

Summary 327

Chapter 14: Exploring the Machine Learning Landscape 329

ML versus software engineering 329

Types of ML methods 334

Supervised learning 334

Unsupervised learning 336

Semi-supervised learning 338

Reinforcement learning 339

Transfer learning 341

ML terminology – a quick review 344

Deep learning 344

Big data 345

Natural language processing 345

Computer vision 345

Cost function 346

Model accuracy 346
Confusion matrix 346
Predictor variables 347
Response variable 347
Dimensionality reduction 347
Class imbalance problem 348
Model bias and variance 349
Underfitting and overfitting 349
Data preprocessing 350
Holdout sample 351
Hyperparameter tuning 351
Performance metrics 352
Feature engineering 352
Model interpretability 353
ML project pipeline 354
Business understanding 355
Understanding and sourcing the data 355
Preparing the data 356
Model building and evaluation 358
Model deployment 358
Learning paradigm 360
Datasets 361
Summary 361
Chapter 15: Predicting Employee Attrition Using Ensemble Models 363
Philosophy behind ensembling 364
Getting started 366
Understanding the attrition problem and the dataset 366
K-nearest neighbors model for benchmarking the performance 381
Bagging 384
Bagged classification and regression trees (treeBag) implementation 386
Support vector machine bagging (SVMBag) implementation 387
Naive Bayes (nbBag) bagging implementation 388
Randomization with random forests 389
Implementing an attrition prediction model with random forests 390
Boosting 391
The GBM implementation 392
Building attrition prediction model with XGBoost 394
Stacking 395
Building attrition prediction model with stacking 396
Summary 399
Chapter 16: Implementing a Jokes Recommendation Engine 401
Fundamental aspects of recommendation engines 402
Recommendation engine categories 404

Content-based filtering 404
Collaborative filtering 404
Hybrid filtering 405
Getting started 405
Understanding the Jokes recommendation problem and the dataset 405
Converting the DataFrame 409
Dividing the DataFrame 410
Building a recommendation system with an item-based collaborative filtering technique 411
Building a recommendation system with a user-based collaborative filtering technique 415
Building a recommendation system based on an association-rule mining technique 419
The Apriori algorithm 420
Content-based recommendation engine 427
Differentiating between ITCF and content-based recommendations 429
Building a hybrid recommendation system for Jokes recommendations 430
Summary 433
References 434
Chapter 17: Sentiment Analysis of Amazon Reviews with NLP 435
The sentiment analysis problem 437
Getting started 438
Understanding the Amazon reviews dataset 438
Building a text sentiment classifier with the BoW approach 443
Pros and cons of the BoW approach 449
Understanding word embedding 450
Building a text sentiment classifier with pretrained word2vec word embedding based on Reuters news corpus 453
Building a text sentiment classifier with GloVe word embedding 458
Building a text sentiment classifier with fastText 462
Summary 469
Chapter 18: Customer Segmentation Using Wholesale Data 471
Understanding customer segmentation 472
Understanding the wholesale customer dataset and the segmentation problem 474
Categories of clustering algorithms 478
Identifying the customer segments in wholesale customer data using k-means clustering 479
Working mechanics of the k-means algorithm 485
Identifying the customer segments in the wholesale customer data using DIANA 490

Identifying the customer segments in the wholesale customers data using AGNES 496
Summary 501

Chapter 19: Image Recognition Using Deep Neural Networks 503
Technical requirements 504
Understanding computer vision 504
Achieving computer vision with deep learning 505
Convolutional Neural Networks 506
Layers of CNNs 507
Introduction to the MXNet framework 510
Understanding the MNIST dataset 511
Implementing a deep learning network for handwritten digit recognition 515
Implementing dropout to avoid overfitting 522
Implementing the LeNet architecture with the MXNet library 527
Implementing computer vision with pretrained models 532
Summary 536

Chapter 20: Credit Card Fraud Detection Using Autoencoders 537
Machine learning in credit card fraud detection 538
Autoencoders explained 540
Types of AEs based on hidden layers 542
Types of AEs based on restrictions 543
Applications of AEs 545
The credit card fraud dataset 546
Building AEs with the H2O library in R 547
Autoencoder code implementation for credit card fraud detection 548
Summary 572

Chapter 21: Automatic Prose Generation with Recurrent Neural Networks 573
Understanding language models 574
Exploring recurrent neural networks 577
Comparison of feedforward neural networks and RNNs 581
Backpropagation through time 584
Problems and solutions to gradients in RNN 585
Exploding gradients 585
Vanishing gradients 586
Building an automated prose generator with an RNN 587
Implementing the project 593
Summary 603

Chapter 22: Winning the Casino Slot Machines with Reinforcement Learning 605

Understanding RL 606
Comparison of RL with other ML algorithms 608
Terminology of RL 609
The multi-arm bandit problem 610
Strategies for solving MABP 614
The epsilon-greedy algorithm 615
Boltzmann or softmax exploration 616
Decayed epsilon greedy 616
The upper confidence bound algorithm 616
Thompson sampling 618
Multi-arm bandit – real-world use cases 618
Solving the MABP with UCB and Thompson sampling algorithms 619
Summary 628
Creating a Package 629
Other Books You May Enjoy 635
Index 639