

Table of Contents

About the Authors	xvii
About the Technical Reviewer	xix
Acknowledgments	xxi
Introduction	xxiii
Chapter 1: Introduction to Machine Learning and R	1
1.1 Understanding the Evolution	2
1.1.1 Statistical Learning.....	2
1.1.2 Machine Learning (ML).....	3
1.1.3 Artificial Intelligence (AI).....	4
1.1.4 Data Mining	5
1.1.5 Data Science	5
1.2 Probability and Statistics	7
1.2.1 Counting and Probability Definition	8
1.2.2 Events and Relationships	10
1.2.3 Randomness, Probability, and Distributions	13
1.2.4 Confidence Interval and Hypothesis Testing.....	15
1.3 Getting Started with R.....	21
1.3.1 Basic Building Blocks	21
1.3.2 Data Structures in R	22
1.3.3 Subsetting	24
1.3.4 Functions and the Apply Family.....	27
1.4 Machine Learning Process Flow	30
1.4.1 Plan.....	31
1.4.2 Explore.....	31

TABLE OF CONTENTS

1.4.3 Build	31
1.4.4 Evaluate	32
1.5 Other Technologies.....	33
1.6 Summary.....	33
Chapter 2: Data Preparation and Exploration	35
2.1 Planning the Gathering of Data	36
2.1.1 Variables Types	36
2.1.2 Data Formats	38
2.1.3 Types of Data Sources	46
2.2 Initial Data Analysis (IDA)	47
2.2.1 Discerning a First Look.....	48
2.2.2 Organizing Multiple Sources of Data into One.....	50
2.2.3 Cleaning the Data	54
2.2.4 Supplementing with More Information.....	58
2.2.5 Reshaping.....	58
2.3 Exploratory Data Analysis	60
2.3.1 Summary Statistics	61
2.3.2 Moment	65
2.4 Case Study: Credit Card Fraud	72
2.4.1 Data Import.....	72
2.4.2 Data Transformation	73
2.4.3 Data Exploration	74
2.5 Summary.....	77
Chapter 3: Sampling and Resampling Techniques	79
3.1 Introduction to Sampling.....	80
3.2 Sampling Terminology.....	81
3.2.1 Sample	82
3.2.2 Sampling Distribution	82
3.2.3 Population Mean and Variance	82
3.2.4 Sample Mean and Variance	83

TABLE OF CONTENTS

3.2.5 Pooled Mean and Variance	83
3.2.6 Sample Point	84
3.2.7 Sampling Error.....	84
3.2.8 Sampling Fraction	84
3.2.9 Sampling Bias.....	85
3.2.10 Sampling Without Replacement (SWOR)	85
3.2.11 Sampling with Replacement (SWR).....	85
3.3 Credit Card Fraud: Population Statistics	86
3.4 Data Description	86
3.5 Population Mean	88
3.6 Population Variance	88
3.7 Pooled Mean and Variance.....	88
3.8 Business Implications of Sampling.....	93
3.8.1 Shortcomings of Sampling	94
3.9 Probability and Non-Probability Sampling	94
3.9.1 Types of Non-Probability Sampling.....	95
3.10 Statistical Theory on Sampling Distributions	96
3.10.1 Law of Large Numbers: LLN	97
3.10.2 Central Limit Theorem	100
3.11 Probability Sampling Techniques	106
3.11.1 Population Statistics.....	106
3.11.2 Simple Random Sampling	110
3.11.3 Systematic Random Sampling	118
3.11.4 Stratified Random Sampling.....	123
3.11.5 Cluster Sampling	131
3.11.6 Bootstrap Sampling	139
3.12 Monte Carlo Method: Acceptance-Rejection Method.....	147
3.13 Summary.....	150

TABLE OF CONTENTS

Chapter 4: Data Visualization in R	151
4.1 Introduction to the ggplot2 Package	152
4.2 World Development Indicators	153
4.3 Line Chart.....	153
4.4 Stacked Column Charts.....	160
4.5 Scatterplots.....	167
4.6 Boxplots	168
4.7 Histograms and Density Plots	172
4.8 Pie Charts.....	177
4.9 Correlation Plots.....	180
4.10 Heatmaps	182
4.11 Bubble Charts	184
4.12 Waterfall Charts	189
4.13 Dendogram	192
4.14 Wordclouds	195
4.15 Sankey Plots	197
4.16 Time Series Graphs	198
4.17 Cohort Diagrams	201
4.18 Spatial Maps	203
4.19 Summary.....	208
Chapter 5: Feature Engineering	211
5.1 Introduction to Feature Engineering	212
5.2 Understanding the Data	213
5.2.1 Data Summary.....	215
5.2.2 Properties of Dependent Variable	215
5.2.3 Features Availability: Continuous or Categorical	219
5.2.4 Setting Up Data Assumptions	221
5.3 Feature Ranking.....	221

5.4 Variable Subset Selection	226
5.4.1 Filter Method	227
5.4.2 Wrapper Methods	231
5.4.3 Embedded Methods.....	240
5.5 Principal Component Analysis.....	245
5.6 Summary.....	251
Chapter 6: Machine Learning Theory and Practice	253
6.1 Machine Learning Types	256
6.1.1 Supervised Learning.....	257
6.1.2 Unsupervised Learning.....	257
6.1.3 Semi-Supervised Learning	257
6.1.4 Reinforcement Learning	258
6.2 Groups of Machine Learning Algorithms.....	258
6.3 Real-World Datasets	264
6.3.1 House Sale Prices.....	264
6.3.2 Purchase Preference	265
6.3.3 Twitter Feeds and Article	266
6.3.4 Breast Cancer	266
6.3.5 Market Basket	267
6.3.6 Amazon Food Reviews.....	268
6.4 Regression Analysis	268
6.5 Correlation Analysis	271
6.5.1 Linear Regression.....	273
6.5.2 Simple Linear Regression.....	275
6.5.3 Multiple Linear Regression.....	279
6.5.4 Model Diagnostics: Linear Regression	283
6.5.5 Polynomial Regression	297
6.5.6 Logistic Regression	302
6.5.7 Logit Transformation.....	303
6.5.8 Odds Ratio	304

TABLE OF CONTENTS

6.5.9 Model Diagnostics: Logistic Regression	313
6.5.10 Multinomial Logistic Regression	326
6.5.11 Generalized Linear Models	330
6.5.12 Conclusion	332
6.6 Support Vector Machine SVM.....	332
6.6.1 Linear SVM	334
6.6.2 Binary SVM Classifier	335
6.6.3 Multi-Class SVM	338
6.6.4 Conclusion	340
6.7 Decision Trees	340
6.7.1 Types of Decision Trees	342
6.7.2 Decision Measures	344
6.7.3 Decision Tree Learning Methods	346
6.7.4 Ensemble Trees	367
6.7.5 Conclusion	376
6.8 The Naive Bayes Method.....	376
6.8.1 Conditional Probability.....	376
6.8.2 Bayes Theorem	377
6.8.3 Prior Probability	377
6.8.4 Posterior Probability	378
6.8.5 Likelihood and Marginal Likelihood.....	378
6.8.6 Naïve Bayes Methods	378
6.8.7 Conclusion	385
6.9 Cluster Analysis.....	385
6.9.1 Introduction to Clustering	386
6.9.2 Clustering Algorithms	387
6.9.3 Internal Evaluation.....	401
6.9.4 External Evaluation.....	403
6.9.5 Conclusion	405

6.10 Association Rule Mining.....	405
6.10.1 Introduction to Association Concepts	406
6.10.2 Rule-Mining Algorithms.....	408
6.10.3 Recommendation Algorithms	417
6.10.4 Conclusion.....	426
6.11 Artificial Neural Networks	426
6.11.1 Human Cognitive Learning	427
6.11.2 Perceptron.....	429
6.11.3 Sigmoid Neuron.....	432
6.11.4 Neural Network Architecture	433
6.11.5 Supervised versus Unsupervised Neural Nets.....	435
6.11.6 Neural Network Learning Algorithms	436
6.11.7 Feed-Forward Back-Propagation	439
6.11.8 Conclusion.....	447
6.12 Text-Mining Approaches.....	448
6.12.1 Introduction to Text Mining	449
6.12.2 Text Summarization	451
6.12.3 TF-IDF	453
6.12.4 Part-of-Speech (POS) Tagging	455
6.12.5 Word Cloud	460
6.12.6 Text Analysis: Microsoft Cognitive Services.....	462
6.12.7 Conclusion	473
6.13 Online Machine Learning Algorithms	473
6.13.1 Fuzzy C-Means Clustering.....	475
6.13.2 Conclusion	479
6.14 Model Building Checklist	479
6.15 Summary.....	481

TABLE OF CONTENTS

Chapter 7: Machine Learning Model Evaluation	483
7.1 Dataset.....	484
7.1.1 House Sale Prices.....	484
7.1.2 Purchase Preference	487
7.2 Introduction to Model Performance and Evaluation.....	489
7.3 Objectives of Model Performance Evaluation	491
7.4 Population Stability Index	492
7.5 Model Evaluation for Continuous Output.....	498
7.5.1 Mean Absolute Error	500
7.5.2 Root Mean Square Error	503
7.5.3 R-Square	504
7.6 Model Evaluation for Discrete Output	508
7.6.1 Classification Matrix.....	509
7.6.2 Sensitivity and Specificity	515
7.6.3 Area Under ROC Curve.....	517
7.7 Probabilistic Techniques	520
7.7.1 K-Fold Cross-Validation	521
7.7.2 Bootstrap Sampling	524
7.8 The Kappa Error Metric	525
7.9 Summary.....	529
Chapter 8: Model Performance Improvement.....	533
8.1 Overview of the Caret Package.....	535
8.2 Introduction to Hyper-Parameters.....	537
8.3 Hyper-Parameter Optimization.....	541
8.3.1 Manual Search.....	543
8.3.2 Manual Grid Search	545
8.3.3 Automatic Grid Search.....	548
8.3.4 Optimal Search	550
8.3.5 Random Search	553
8.3.6 Custom Searching	555

TABLE OF CONTENTS

8.4 The Bias and Variance Tradeoff.....	559
8.5 Introduction to Ensemble Learning.....	564
8.5.1 Voting Ensembles.....	565
8.5.2 Advanced Methods in Ensemble Learning.....	567
8.6 Ensemble Techniques Illustration in R.....	570
8.6.1 Bagging Trees.....	571
8.6.2 Gradient Boosting with a Decision Tree.....	573
8.6.3 Blending KNN and Rpart.....	578
8.6.4 Stacking Using caretEnsemble.....	580
8.7 Advanced Topic: Bayesian Optimization of Machine Learning Models.....	586
8.8 Summary.....	592
Chapter 9: Time Series Modeling.....	595
9.1 Components of Time Series.....	596
9.2 Test of Stationarity.....	600
9.3 ACF and AR Model.....	604
9.4 PACF and MA Model.....	608
9.5 ARIMA Model.....	612
9.5.1 Box-Jenkins Approach.....	613
9.6 Linear Regression with AR Errors.....	621
9.7 Summary.....	626
Chapter 10: Scalable Machine Learning and Related Technologies.....	629
10.1 Distributed Processing and Storage.....	630
10.1.1 Google File System (GFS).....	631
10.1.2 MapReduce.....	632
10.1.3 Parallel Execution in R.....	633
10.2 The Hadoop Ecosystem.....	638
10.2.1 MapReduce.....	639
10.2.2 Hive.....	644
10.2.3 Apache Pig.....	648

TABLE OF CONTENTS

10.2.4 HBase	652
10.2.5 Spark	654
10.3 Machine Learning in R with Spark	655
10.3.1 Setting the Environment Variable	656
10.3.2 Initializing the Spark Session	656
10.3.3 Loading Data and the Running Preprocess	657
10.3.4 Creating SparkDataFrame	658
10.3.5 Building the ML Model.....	659
10.3.6 Predicting the Test Data.....	660
10.3.7 Stopping the SparkR Session	661
10.4 Machine Learning in R with H2O.....	661
10.4.1 Installation of Packages	663
10.4.2 Initialization of H2O Clusters	664
10.5 Summary.....	665
Chapter 11: Deep Learning Using Keras and TensorFlow	667
11.1 Introduction to Deep Learning	668
11.2 Deep Learning Architectures.....	669
11.2.1 Convolutional Neural Networks (CNN)	669
11.2.2 Recurrent Neural Networks (RNN).....	670
11.2.3 Generative Adversarial Network (GAN)	672
11.3 Deep Learning Toolset.....	674
11.3.1 High-Level Library	674
11.3.2 Backend Engine or Frameworks.....	674
11.3.3 Hardware Capability	675
11.3.4 Programming Language Choice	675
11.3.5 Cloud Infrastructure.....	675
11.4 Use Case: Identify Duplicate Questions in Quora	676
11.4.1 Environment Setup	676
11.4.2 Data Preprocessing	676
11.4.3 Benchmark Model	678

TABLE OF CONTENTS

11.4.4 Siamese Recurrent Architectures.....	680
11.4.5 The Keras Model.....	683
11.4.6 The Model Summary.....	683
11.4.7 The Validation Sample	684
11.4.8 Train the Model.....	684
11.4.9 Save the Model.....	685
11.4.10 Model Performance	686
11.4.11 Make Predictions.....	687
11.4.12 Example Predictions.....	687
11.5 Summary.....	688
Index.....	689