Data Science Solutions with Python

Fast and Scalable Models Using Keras, PySpark MLlib, H2O, XGBoost, and Scikit-Learn

Tshepo Chris Nokeri

Apress®

Table of Contents

About the Author ix

About the Technical Reviewer xi

Acknowledgments xiii

Introduction xv

Chapter 1: Exploring Machine Learning 1

Exploring Supervised Methods 1

Exploring Nonlinear Models 2

Exploring Ensemble Methods 3

Exploring Unsupervised Methods 3

Exploring Cluster Methods 3

Exploring Dimension Reduction 4

Exploring Deep Learning 4

Conclusion 5

Chapter 2: Big Data, Machine Learning, and Deep Learning Frameworks 7

Big Data 7

Big Data Features 8

Impact of Big Data on Business and People 8

Better Customer Relationships 8

Refined Product Development 9

Improved Decision-Making 9

Big Data Warehousing 9

Big Data ETL 9

Big Data Frameworks 10

Apache Spark 10

ML Frameworks 13
Scikit-Learn 13
H2O 13
XGBoost 14
DL Frameworks 14
Keras 14

Chapter 3: Linear Modeling with Scikit-Learn, PySpark, and H2O 15

Exploring the Ordinary Least-Squares Method 15
Scikit-Learn in Action 17
PySpark in Action 20
H2O in Action 22
Conclusion 28

Chapter 4: Survival Analysis with PySpark and Lifelines 29

Exploring Survival Analysis 29
Exploring Cox Proportional Hazards Method 29
Lifeline in Action 30
Exploring the Accelerated Failure Time Method 34
PySpark in Action 34
Conclusion 37

Chapter 5: Nonlinear Modeling With Scikit-Learn, PySpark, and H2O 39

Exploring the Logistic Regression Method 39
Scikit-Learn in Action 41
PySpark in Action 48
H2O in Action 52
Conclusion 57

Chapter 6: Tree Modeling and Gradient Boosting with Scikit-Learn, XGBoost, PySpark, and H2O 59

Decision Trees 59

Preprocessing Features 60

Scikit-Learn in Action 61

Gradient Boosting 66

XGBoost in Action 66

PySpark in Action 69

H2O in Action 71

Conclusion 74

Chapter 7: Neural Networks with Scikit-Learn, Keras, and H2O 75

Exploring Deep Learning 75

Multilayer Perceptron Neural Network 75

Preprocessing Features 76

Scikit-Learn in Action 77

Keras in Action 82

Deep Belief Networks 87

H2O in Action 87

Conclusion 88

Chapter 8: Cluster Analysis with Scikit-Learn, PySpark, and H2O 89

Exploring the K-Means Method 89

Scikit-Learn in Action 91

PySpark in Action 93

H2O in Action 97

Conclusion 99

Chapter 9: Principal Component Analysis with Scikit-Learn, PySpark, and H2O 101

Exploring the Principal Component Method 101

Scikit-Learn in Action 102

PySpark in Action 105

H2O in Action 109

Conclusion 110

Chapter 10: Automating the Machine Learning Process with H2O 111

Exploring Automated Machine Learning 111

Preprocessing Features 112

H2O AutoML in Action 112

Conclusion 116

Index 117