# THE SUPERVISED LEARNING WORKSHOP

## A NEW, INTERACTIVE APPROACH TO UNDERSTANDING SUPERVISED LEARNING ALGORITHMS

SECOND EDITION

BLAINE BATEMAN, ASHISH RANJAN JHA,
BENJAMIN JOHNSTON, AND ISHITA MATHUR

# The Supervised Learning Workshop

## Second Edition

A New, Interactive Approach to Understanding Supervised Learning Algorithms

Blaine Bateman, Ashish Ranjan Jha,
Benjamin Johnston, and Ishita Mathur

Packt

## The Supervised Learning Workshop

## Second Edition

Copyright © 2020 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

**Authors:** Blaine Bateman, Ashish Ranjan Jha, Benjamin Johnston, and Ishita Mathur

**Reviewers:** Tiffany Ford, Sukanya Mandal, Ashish Pratik Patil, and Ratan Singh

**Managing Editor:** Snehal Tambe

**Acquisitions Editor:** Anindya Sil

**Production Editor:** Samita Warang

**Editorial Board:** Shubhopriya Banerjee, Bharat Botle, Ewan Buckingham, Megan Carlisle, Mahesh Dhyani, Manasa Kumar, Alex Mazonowicz, Bridget Neale, Dominic Pereira, Shiny Poojary, Abhishek Rane, Brendan Rodrigues, Mugdha Sawarkar, Erol Staveley, Ankita Thakur, Nitesh Thakur, and Jonathan Wray

First published: April 2019

Second edition: February 2020

Production reference: 1280220

ISBN 978-1-80020-904-6

Published by Packt Publishing Ltd.

Livery Place, 35 Livery Street

Birmingham B3 2PB, UK

# Table of Contents

Preface i

Chapter 1: Fundamentals of Supervised Learning Algorithms 1

Introduction ..................... 2

When to Use Supervised Learning ..................... 3

Python Packages and Modules ..................... 4

Loading Data in Pandas ..................... 5

Exercise 1.01: Loading and Summarizing the Titanic Dataset ..................... 7

Exercise 1.02: Indexing and Selecting Data ..................... 10

Exercise 1.03: Advanced Indexing and Selection ..................... 14

Pandas Methods ..................... 17

Exercise 1.04: Splitting, Applying, and Combining Data Sources ..................... 21

Quantiles ..................... 23

Lambda Functions ..................... 24

Exercise 1.05: Creating Lambda Functions ..................... 25

Data Quality Considerations ..................... 28

Managing Missing Data ..................... 28

Class Imbalance ..................... 33

Low Sample Size ..................... 34

Activity 1.01: Implementing Pandas Functions ..................... 35

Summary ..................... 37

## Chapter 2: Exploratory Data Analysis and Visualization 39

Introduction ........................................................................................ 40

Exploratory Data Analysis (EDA) ................................................................ 40

Summary Statistics and Central Values ...................................................... 42

Exercise 2.01: Summarizing the Statistics of Our Dataset ............................ 43

Missing Values ...................................................................................... 48

Finding Missing Values ........................................................................... 49

Exercise 2.02: Visualizing Missing Values ................................................... 50

Imputation Strategies for Missing Values ................................................... 54

Exercise 2.03: Performing Imputation Using Pandas ..................................... 55

Exercise 2.04: Performing Imputation Using Scikit-Learn ............................. 56

Exercise 2.05: Performing Imputation Using Inferred Values ........................ 58

Activity 2.01: Summary Statistics and Missing Values .................................. 61

Distribution of Values ............................................................................ 65

Target Variable .................................................................................... 65

Exercise 2.06: Plotting a Bar Chart ........................................................... 65

Categorical Data .................................................................................. 67

Exercise 2.07: Identifying Data Types for Categorical Variables .................. 68

Exercise 2.08: Calculating Category Value Counts ....................................... 70

Exercise 2.09: Plotting a Pie Chart ........................................................... 71

Continuous Data .................................................................................. 73

Skewness............................................................................................. 75

Kurtosis .............................................................................................. 75

Exercise 2.10: Plotting a Histogram ......................................................... 75

Exercise 2.11: Computing Skew and Kurtosis .............................................. 77

Activity 2.02: Visually Representing the Distribution of Values ................... 79

Relationships within the Data ........................................................ 84
Relationship between Two Continuous Variables ................................... 84
Pearson's Coefficient of Correlation ................................................. 85
Exercise 2.12: Plotting a Scatter Plot ................................................ 85
Exercise 2.13: Plotting a Correlation Heatmap ..................................... 87
Using Pairplots ........................................................................... 90
Exercise 2.14: Implementing a Pairplot ............................................... 90
Relationship between a Continuous and a Categorical Variable ............... 92
Exercise 2.15: Plotting a Bar Chart .................................................... 92
Exercise 2.16: Visualizing a Box Plot ................................................. 95
Relationship Between Two Categorical Variables .................................. 97
Exercise 2.17: Plotting a Stacked Bar Chart ......................................... 97
Activity 2.03: Relationships within the Data ........................................ 99
Summary ....................................................................................... 105

## Chapter 3: Linear Regression 107

Introduction .................................................................................. 108
Regression and Classification Problems ............................................ 108
The Machine Learning Workflow ....................................................... 109
Business Understanding ................................................................. 110
Data Understanding ...................................................................... 110
Data Preparation .......................................................................... 111
Modeling .................................................................................... 111
Evaluation .................................................................................. 112
Deployment ................................................................................ 112
Exercise 3.01: Plotting Data with a Moving Average ............................. 112
Activity 3.01: Plotting Data with a Moving Average .............................. 123

Linear Regression ........ 125
Least Squares Method ........ 126
The Scikit-Learn Model API ........ 126
Exercise 3.02: Fitting a Linear Model Using the Least Squares Method .. 127
Activity 3.02: Linear Regression Using the Least Squares Method ........ 132
Linear Regression with Categorical Variables ........ 137
Exercise 3.03: Introducing Dummy Variables ........ 139
Activity 3.03: Dummy Variables ........ 150
Polynomial Models with Linear Regression ........ 152
Exercise 3.04: Polynomial Models with Linear Regression ........ 154
Activity 3.04: Feature Engineering with Linear Regression ........ 159
Generic Model Training ........ 163
Gradient Descent ........ 165
Exercise 3.05: Linear Regression with Gradient Descent ........ 168
Exercise 3.06: Optimizing Gradient Descent ........ 176
Activity 3.05: Gradient Descent ........ 181
Multiple Linear Regression ........ 183
Exercise 3.07: Multiple Linear Regression ........ 185
Summary ........ 193

## Chapter 4: Autoregression 195

Introduction ........ 196
Autoregression Models ........ 196
Exercise 4.01: Creating an Autoregression Model ........ 197
Activity 4.01: Autoregression Model Based on Periodic Data ........ 214
Summary ........ 221

Chapter 5: Classification Techniques 223

Introduction ..... 224

Ordinary Least Squares as a Classifier ..... 224

Exercise 5.01: Ordinary Least Squares as a Classifier ..... 226

Logistic Regression ..... 232

Exercise 5.02: Logistic Regression as a Classifier – Binary Classifier ..... 236

Exercise 5.03: Logistic Regression – Multiclass Classifier ..... 242

Activity 5.01: Ordinary Least Squares Classifier – Binary Classifier ..... 248

Select K Best Feature Selection ..... 249

Exercise 5.04: Breast Cancer Diagnosis Classification Using Logistic Regression ..... 250

Classification Using K-Nearest Neighbors ..... 254

Exercise 5.05: KNN Classification ..... 257

Exercise 5.06: Visualizing KNN Boundaries ..... 259

Activity 5.02: KNN Multiclass Classifier ..... 266

Classification Using Decision Trees ..... 267

Exercise 5.07: ID3 Classification ..... 269

Classification and Regression Tree ..... 280

Exercise 5.08: Breast Cancer Diagnosis Classification Using a CART Decision Tree ..... 281

Activity 5.03: Binary Classification Using a CART Decision Tree ..... 284

Artificial Neural Networks ..... 286

Exercise 5.09: Neural Networks – Multiclass Classifier ..... 288

Activity 5.04: Breast Cancer Diagnosis Classification Using Artificial Neural Networks ..... 292

Summary ..... 294

## Chapter 6: Ensemble Modeling 297

Introduction ........ 298

One-Hot Encoding ........ 299

Exercise 6.01: Importing Modules and Preparing the Dataset ........ 300

Overfitting and Underfitting ........ 302

Underfitting ........ 304

Overfitting ........ 305

Overcoming the Problem of Underfitting and Overfitting ........ 306

Bagging ........ 307

Bootstrapping ........ 308

Exercise 6.02: Using the Bagging Classifier ........ 310

Random Forest ........ 312

Exercise 6.03: Building the Ensemble Model Using Random Forest ........ 313

Boosting ........ 314

Adaptive Boosting ........ 316

Exercise 6.04: Implementing Adaptive Boosting ........ 316

Gradient Boosting ........ 319

Exercise 6.05: Implementing GradientBoostingClassifier to Build an Ensemble Model ........ 320

Stacking ........ 321

Exercise 6.06: Building a Stacked Model ........ 324

Activity 6.01: Stacking with Standalone and Ensemble Algorithms ........ 328

Summary ........ 331

## Chapter 7: Model Evaluation 333

Introduction ........ 334

Importing the Modules and Preparing Our Dataset ........ 336

Evaluation Metrics ........................................ 338
Regression Metrics ........................................ 338
Exercise 7.01: Calculating Regression Metrics ........................................ 341
Classification Metrics ........................................ 342
Numerical Metrics ........................................ 342
Curve Plots ........................................ 346
Exercise 7.02: Calculating Classification Metrics ........................................ 348
Splitting a Dataset ........................................ 350
Hold-Out Data ........................................ 350
K-Fold Cross-Validation ........................................ 352
Sampling ........................................ 353
Exercise 7.03: Performing K-Fold Cross-Validation with Stratified Sampling ........................................ 354
Performance Improvement Tactics ........................................ 355
Variation in Train and Test Errors ........................................ 356
Learning Curve ........................................ 356
Validation Curve ........................................ 357
Hyperparameter Tuning ........................................ 358
Exercise 7.04: Hyperparameter Tuning with Random Search ........................................ 360
Feature Importance ........................................ 364
Exercise 7.05: Feature Importance Using Random Forest ........................................ 364
Activity 7.01: Final Test Project ........................................ 366
Summary ........................................ 369
Appendix 371
Index 467