Table of Contents

FOREWORD by William Hersh xi

PREFACE xiii

CHAPTER 1: INTRODUCTION TO DATA SCIENCE 1

What Is Data Science? 1
Background 2
Artificial Intelligence and Machine Learning. 2
Artificial Intelligence and Data Science 5
Foundations and Frontiers: Integrating DMBoK Practices with Data Science. . . . 5
Statistical Foundations and Predictive Frontiers: A Data Voyage 6
The 8-Step Data Science Discover Process 6
The 3 Levels of AI for Healthcare 7
What Do Data Scientists Do?. 9
Why Learn Data Science? 10
Data Science Educational Challenges 10
Why Did We Create This Textbook? 11
Who Is The Target Audience? 12
Why Use Open-Source Software? 12
Orange Data Mining 13
Orange General Information. 13
JASP Stats Package 16
Introduction 16
JASP Unique Features 16
JASP Organization. 17
JASP limitations 18
Orange vs JASP 18
Putting Knowledge Into Practice 18
Exercise 18
Maturity Levels 19
Resources 19
References 20

CHAPTER 2: DATA PREPARATION AND WRANGLING 23

Understanding Data And Data Types. 24
Introduction 24
Data Science DISCOVER Process Step 2 . . 24
Data Types 27
Step 3 In The Data Science Process 29
Data Cleaning 29
What is Feature Engineering? 29
Handling Missing Data with Imputation or Deletion 30
Handling Outliers 33
Binning / Discretize 37
Log Transform 41
Feature Creation 42
One Hot Encoding Versus N-1 Dummy Encoding 42
Group Calculation Operations 45
Feature Splitting 45
Feature Scaling. 45
Univariate, Bivariate, And Multivariate Analysis 49
Important Data Challenges 50
Curse of Dimensionality 50
Multicollinearity 50
Data Leakage 62
Imbalanced Data 63
Synthetic Data 63
Introduction 63
Synthea. 63
AI-Generated Synthetic Tabular Data 64
Data Science Experiments (DSEs) 66
Experimental Thinking in Data Science. . . 66
Nine Steps for Creating Effective Data Science Experiments 67
Data Science Experiment Complexity Level 68

Algorithm Preprocessing Defaults 67
Algorithm Compatibility with Other Widgets 69
Data Science Experiment Complexity Level Examples 70
Orange Data And Transform Menu Widgets . . 77
JASP Data Features 97
Putting Knowledge Into Practice 97
Exercise 97
Maturity Level 98
Resources 99
References 99

CHAPTER 3: DATA VISUALIZATION 105

Introduction 106
Data Visualization Tables And Plots 109
Statistical Tables and Displays 109
Basic Data Displays 110
Distributions 113
Relationships Between Variables 117
Time Series 122
Geographical Displays 122
Hierarchical Displays 124
Multivariate Displays 124
Statistical Models and Displays 129
Text Displays 135
Variable Ranks - Non-Regularized Versus Regularized 136
Dashboards And Infographics 136
Dashboards 136
Infographics 138
Putting Knowledge Into Practice 140
Exercise 140
Maturity Levels 141
Resources 141
References 141

CHAPTER 4: MACHINE LEARNING MODELS 143

Introduction 143
How Do Models Work? 146
More Modeling Concepts 147
Bias-Variance Tradeoff 147
Model Interpretability and Explainability . . 149
Gradient Descent 150
Local and Global Minimum 150
Types of Gradient Descent 151
Orange Algorithms 151
JASP Models 159
Linear Regression 160
Logistic Regression 160
JASP Machine Learning Algorithms 160
MLOps 161
Tackling Model Challenges 162
Model Bias and Fairness 162
Navigating Imbalanced Datasets 162
Hyperparameter Optimization 173
Avoiding Model Overfitting 173
Putting Knowledge Into Practice 173
Exercise 173
Maturity Levels 175
Resources 175
References 176

CHAPTER 5: MODEL EVALUATION 179

Model Validation 180
Internal Validation (IV) 180
External Validation (EV) 181
Optimal Validation (OV) 182
Further Validation Guidance 182
Model Discrimination And Calibration 182
Probability, Odds, Odds Ratios And Log Odds Ratios 185
Probability vs Odds 185
Odds Ratios and Log Odds Ratios 185
Diagnostic Tests and Likelihood Ratios . . . 187
Orange Model Evaluation Widgets 188
Test and Score Widget 189
Predictions Widget 191
Confusion Matrix 192
ROC Curve Widget 193
Precision-Recall (PR) Curve 196
Other Valuable Orange Curves 196
Lift Curve 196
Calibration Plot 196
Residual Plots 198
Explain Module 201
Rank Widget 201
Feature Importance Widget 203
Explain Prediction Widget 204
Explain Model Widget 206

JASP Model Evaluation 206
Hypothesis Testing 208
Hypothesis Testing Example 208
How Hypothesis Tests Make Decisions . . . 208
Type 1 and Type 2 Hypothesis Testing Errors 208
Hypothesis Tests Available in Orange, JASP, and BlueSky Statistics 209
Identifying And Addressing Concept Drift And Model Drift 212
Concept Drift 212
Model Drift 212
When to Retrain a Model 213
Putting Knowledge Into Practice 213
Exercise 213
Maturity Levels 214
Resources 215
References 215

CHAPTER 6: SUPERVISED LEARNING 217

Classification Modeling 218
Introduction 218
Regression Modeling 219
Linear regression 219
Simple Linear Regression (SLR) 219
Multiple Linear Regression (MLR) 221
The Dummy Variable Trap 222
Logistic Regression 223
Polynomial Regression 223
Linear Regression Model Regularization . . 224
Regression Performance Metrics 224
Putting Knowledge Into Practice 225
Exercises 225
Maturity Levels 228
Resources 229
References 229

CHAPTER 7: UNSUPERVISED LEARNING . . 231

Unsupervised Learning 232
K-Means Clustering 232
K-Means Clustering Exercises 233
Hierarchical Clustering 237
Principle Component Analysis (PCA) 237
Introduction 237
Routine Steps to Achieve PCA 238
PCA Exercises 239
Association Rules 241
Association Rules Exercise 242
Putting Knowledge Into Practice 243
Exercise 243
Maturity Levels 246
Resources 247
References 247

CHAPTER 8: TIME SERIES FORECASTING AND SURVIVAL ANALYSIS 249

Time Series Forecasting With Arima 250
Arima Time Series Forecasting in Orange . 250
Pros and Cons of ARIMA Models 252
Survival Analysis 254
The General Steps to Perform Kaplan-Meier Survival Analysis 254
Orange Example of Kaplan-Meier Survival Analysis 255
Putting Knowledge Into Practice 259
Exercise 259
Maturity Levels 260
Resources 261
References 261

CHAPTER 9: GEOLOCATION 262

Introduction 264
Geolocation Exercises 265
Geolocation with Synthetic Data 275
Multi-Industry Application Examples Of Geo-Mapping And Geo-Analysis 267
Putting Knowledge Into Practice 269
Exercise 269
Maturity Levels 270
Resources 271
References 271

CHAPTER 10: IMAGE ANALYTICS 273

Introduction 273
Medical Field Applications 273
Non- Medical Field Applications 274
AI Imaging: Industry Application Examples . 274
How Do CNNs Work? 275
The Normal Steps Of Image Analytics With a CNN 275

CNN Layers . . . 276
Putting Knowledge Into Practice . . . 277
Exercises . . . 277
Maturity Levels . . . 280
Resources . . . 281
References . . . 281

CHAPTER 11: TEXT MINING . . . 283

Introduction . . . 284
Neural Networks And NLP . . . 285
Recurrent Neural Networks . . . 285
Transformer and Attention Models . . . 286
NLP Concepts . . . 286
NLP Preprocessing Steps . . . 287
Topic Analysis . . . 288
Large Language Models (LLMs) . . . 288
Graphs From Text . . . 289
Code Assistants . . . 291
Small Language Models (SLM) . . . 291
Orange NLP Widgets . . . 291
Text Document Widgets . . . 292
Preprocess Text Widget . . . 292
Document Embedding Widget . . . 292
Bag of Words Widget . . . 292
Topic Modeling Widget . . . 292
Putting Knowledge Into Practice . . . 292
Exercises . . . 292
Maturity Levels . . . 298
Resources . . . 299
References . . . 299

CHAPTER 12: INTEGRATING CONTINUOUS IMPROVEMENT AND DATA SCIENCE INTO INDUSTRY 4.0 . . . 301

Introduction . . . 302
Industry And Sector 4.0 Initiatives . . . 303
Accelerating Industry 4.0: The Pivotal Role Of Continuous Improvement (CI) Programs . . . 305
Eleven Powerful Continuous Improvement Techniques . . . 307
Hoshin Kanri with Catchball Strategic Planning . . . 307
Six Sigma . . . 308
Lean / Kaizen . . . 311
Lean Six Sigma . . . 314
Predictive Modeling . . . 315
High Reliability Organizations (HROs) . . . 316
Agile Versus Waterfall Thinking . . . 318
Cascading Risk Management (CRM) . . . 319
AI Systems . . . 321
AI Chatbots . . . 322
Innovation-On-Demand Techniques . . . 324
The Redwood Approach: Building Business Resilience Through Cooperative Strength . . . 327
Maturity Levels . . . 327
Resources . . . 328
References . . . 329

APPENDIX A: DATA VISUALIZATION . . . 335

APPENDIX B: MACHINE LEARNING ALGORITHMS . . . 337

APPENDIX C: GLOSSARY . . . 341

INDEX . . . 347