

Contents

Preface	xv
0.1 What Is Veridical Data Science?	xix
0.2 The Structure of This Book	xxi
0.2.1 What This Book Does Not Contain	xxii
0.2.2 Use of Mathematical Equations	xxiv
0.2.3 Use of Code	xxiv
0.2.4 Terminology and Important Concepts	xxv
0.2.5 Exercises	xxv
Acknowledgments	xxvii
I AN INTRODUCTION TO VERIDICAL DATA SCIENCE	
1 An Introduction to Veridical Data Science	3
1.1 The Role of Data and Algorithms in Real-World Decision Making	3
1.2 Evaluating and Building Trustworthiness Using Critical Thinking	6
1.2.1 Case Study: Estimating Animal Deaths Caused by the 2019 Australian Bushfires	8
1.3 Evaluating and Building Trustworthiness Using the PCS Framework	11
1.3.1 Predictability	12
1.3.2 Stability	17
1.3.3 Computability	20
1.3.4 PCS as a Unification of Cultures	20
Exercises	21
2 The Data Science Life Cycle	23
2.1 Data Terminology	24
2.2 DSLC Stage 1: Problem Formulation and Data Collection	26

2.2.1	Formulating a Question	26
2.2.2	Data Collection	28
2.2.3	Making a Plan for Evaluating Predictability	29
2.3	DSLCL Stage 2: Data Cleaning and Exploratory Data Analysis	30
2.3.1	Data Cleaning	30
2.3.2	Preprocessing	30
2.3.3	Exploratory Data Analysis	31
2.3.4	Data Snooping and PCS	33
2.4	DSLCL Stage 3: Uncovering Intrinsic Data Structures	33
2.5	DSLCL Stage 4: Predictive and/or Inferential Analysis	34
2.5.1	Prediction	34
2.5.2	Data-Driven Inference	36
2.6	DSLCL Stage 5: Evaluation of Results	36
2.7	DSLCL Stage 6: Communication of Results and Updating Domain Knowledge	37
	Exercises	38
3	Setting Up Your Data Science Project	41
3.1	Programming Languages and IDEs	41
3.1.1	Code Guidelines	44
3.1.2	Writing Functions for Data Cleaning and Preprocessing	45
3.2	A Consistent Project Structure	45
3.2.1	Best Practices for Working with Data Files	47
3.3	Reproducibility	51
3.3.1	Defining Reproducibility	53
3.3.2	Tips and Tricks for Reproducibility	55
3.4	Tools for Collaboration	56
3.4.1	Git and GitHub	56
	Exercises	58
II	PREPARING, EXPLORING, AND DESCRIBING DATA	
4	Data Preparation	65
4.1	The Organ Donation Data	70
4.1.1	A Plan for Evaluating Predictability	70
4.2	A Generalizable Data Cleaning Procedure	71
4.3	Step 1: Learn About the Data Collection Process and the Problem Domain	74
4.3.1	Reviewing the Background and Domain Information of the Organ Donation Data	75
4.4	Step 2: Load the Data	77
4.4.1	Loading the Organ Donation Data	78
4.5	Step 3: Examine the Data and Create Action Items	78
4.5.1	Messy Data Trait 1: Invalid or Inconsistent Values	79

4.5.2	Messy Data Trait 2: Improperly Formatted Missing Values	83
4.5.3	Messy Data Trait 3: Nonstandard Data Format	89
4.5.4	Messy Data Trait 4: Messy Column Names	92
4.5.5	Messy Data Trait 5: Improper Variable Types	93
4.5.6	Messy Data Trait 6: Incomplete Data	95
4.5.7	Addressing Any Remaining Questions and Assumptions	98
4.6	Step 4: Clean the Data	98
4.6.1	Cleaning the Organ Donation Data	100
4.7	Additional Common Preprocessing Steps	101
	Exercises	101
5	Exploratory Data Analysis	107
5.1	A Question-and-Answer-Based Exploratory Data Analysis Workflow	109
5.1.1	Choosing a Data Visualization Technique	109
5.1.2	Exploratory and Explanatory Data Analysis	113
5.2	Common Explorations	119
5.2.1	A Typical Value of a Single Variable: Mean and Median	119
5.2.2	The “Spread” of a Variable: Variance and Standard Deviation	123
5.2.3	Linear Relationships Between Two Variables: Covariance and Correlation	125
5.3	Comparability	131
5.4	PCS Scrutinization of Exploratory Results	133
5.4.1	Predictability	134
5.4.2	Stability	134
	Exercises	139
6	Principal Component Analysis	147
6.1	The Nutrition Project	149
6.1.1	Data Source	149
6.1.2	A Predictability Evaluation Plan: The Legacy Dataset	152
6.1.3	Data Cleaning	152
6.1.4	Exploratory Data Analysis	153
6.1.5	The Analysis Plan: Summarizing Nutrient Groups Separately	153
6.2	Generating Summary Variables: Principal Component Analysis	156
6.2.1	The First Principal Component	158
6.3	Preprocessing: Standardization for Comparability	162
6.4	Singular Value Decomposition	164
6.4.1	The Proportion of Variability Explained	169
6.5	Preprocessing: Gaussianity and Transformations	173

6.6	Principal Component Analysis Step-by-Step Summary	178
6.7	PCS Evaluation of Principal Component Analysis	179
6.7.1	Predictability	179
6.7.2	Stability	181
6.8	Applying Principal Component Analysis to Each Nutrient Group	186
6.9	Alternatives to Principal Component Analysis Exercises	189
7	Clustering	195
7.1	Understanding Clusters	197
7.1.1	Evaluating Clusters	198
7.1.2	Defining Similarity	199
7.2	Hierarchical Clustering	204
7.2.1	Defining Intercluster Similarity (Linkage Methods)	205
7.2.2	A Demonstration of Single-Linkage Clustering	206
7.3	K-Means Clustering	211
7.3.1	The K-Means Objective Function	213
7.3.2	Preprocessing: Transformation	215
7.4	Visualizing Clusters in High Dimensions	215
7.4.1	Sampling Data Points from Each Cluster	216
7.4.2	Comparing Variable Distributions across Each Cluster	218
7.4.3	Projecting Clusters onto a Two-Dimensional Scatterplot	218
7.5	Quantitative Measures of Cluster Quality	220
7.5.1	Within-Cluster Sum of Squares	220
7.5.2	The Silhouette Score	224
7.6	The Rand Index for Comparing Cluster Similarity	228
7.6.1	The Adjusted Rand Index	231
7.7	Choosing the Number of Clusters	232
7.7.1	Using Cross-Validation to Choose K	232
7.8	PCS Scrutinization of Cluster Results	238
7.8.1	Predictability	238
7.8.2	Stability	239
7.9	The Final Clusters Exercises	243 246
III	PREDICTION	
8	An Introduction to Prediction Problems	253
8.1	Connecting the Past, Present, and Future for Prediction Problems	255
8.2	Setting up a Prediction Problem	258
8.2.1	Defining a Response Variable	258

8.2.2	Defining Predictor Variables	259
8.2.3	Quantity versus Quality	260
8.3	PCS and Evaluating Prediction Algorithms	261
8.3.1	Predictability	261
8.3.2	Stability	263
8.4	The Ames House Price Prediction Project	263
8.4.1	Data Source	263
8.4.2	The Ames Housing Market: Past versus Future	264
8.4.3	A Predictability Evaluation Plan	265
8.4.4	Cleaning and Preprocessing the Ames Housing Dataset	267
8.4.5	Exploring the Ames Housing Data	268
	Exercises	270
9	Continuous Responses and Least Squares	273
9.1	Visualizing Predictive Relationships	273
9.2	Using Fitted Lines to Generate Predictions	276
9.3	Computing Fitted Lines	277
9.3.1	Least Absolute Deviation	278
9.3.2	Least Squares	281
9.3.3	Least Squares versus Least Absolute Deviation	285
9.4	Quantitative Measures of Predictive Performance	289
9.4.1	Mean Squared Error, Mean Absolute Error, and Median Absolute Deviation	290
9.4.2	Correlation and R-Squared	293
9.5	PCS Scrutinization of Prediction Results	297
9.5.1	Predictability	297
9.5.2	Stability	298
	Exercises	302
10	Extending the Least Squares Algorithm	307
10.1	Linear Fits with Multiple Predictive Features	307
10.1.1	Interpreting the Coefficients of a Linear Fit	308
10.1.2	Comparing Coefficients	310
10.2	Pre-processing: One-Hot-Encoding	312
10.3	Pre-processing: Variable Transformations	316
10.3.1	Heteroskedasticity	317
10.3.2	Fitting Nonlinear Curves	317
10.4	Feature Selection	319
10.5	Regularization	321
10.5.1	Ridge	324
10.5.2	Lasso	326
10.5.3	Standardization for Lasso and Ridge	329
10.5.4	Cross-Validation for Choosing Regularization Hyperparameters	329
10.5.5	Lasso as a Feature Selection Technique	331

10.6	PCS Evaluations	333
10.6.1	Predictability	333
10.6.2	Stability	334
10.7	Appendix: Matrix Formulation of a Linear Fit	343
	Exercises	344
11	Binary Responses and Logistic Regression	349
11.1	The Online Shopping Purchase Prediction Project	349
11.1.1	Data Source	349
11.1.2	A Predictability Evaluation Plan	351
11.1.3	Cleaning and Preprocessing the Online Shopping Dataset	352
11.1.4	Exploring the Online Shopping Data	354
11.2	Least Squares for Binary Prediction	357
11.3	Logistic Regression	358
11.3.1	Log Odds and the Logistic Function	360
11.3.2	Generating Binary Response Predictions for Logistic Regression	362
11.3.3	Fitting Logistic Regression Coefficients	364
11.3.4	Logistic Regression with Multiple Predictive Features	365
11.3.5	Interpreting and Comparing the Coefficients	366
11.4	Quantitative Measures of Binary Predictive Performance	369
11.4.1	Prediction Accuracy and Prediction Error	369
11.4.2	The Confusion Matrix	371
11.4.3	True Positive Rate	374
11.4.4	True Negative Rate	375
11.4.5	The Sensitivity-Specificity Trade-Off: Choosing the Binary Threshold Value	376
11.4.6	ROC Curves	378
11.4.7	Class Imbalance	381
11.5	PCS Scrutinization of Binary Prediction Results	382
11.5.1	Predictability	383
11.5.2	Stability	385
	Exercises	392
12	Decision Trees and the Random Forest Algorithm	397
12.1	Decision Trees	397
12.2	The Classification and Regression Trees Algorithm	400
12.2.1	Choosing the Splits	400
12.2.2	Stopping Criteria and Hyperparameters	406
12.2.3	Generating Predictions	408
12.3	The Random Forest Algorithm	410
12.3.1	Tuning the RF Hyperparameters	412
12.4	Random Forest Feature Importance Measures	413
12.4.1	The Permutation Feature Importance Score	414
12.4.2	The Gini Impurity Feature Importance Score	416

12.5	PCS Evaluation of the CART and RF Algorithms	418
12.5.1	Predictability	418
12.5.2	Stability	421
	Exercises	425
13	Producing the Final Prediction Results	429
13.1	Approach 1: Choosing a Single Predictive Fit Using PCS	432
13.1.1	Choosing a Single Fit for the Ames House Price Prediction Project	435
13.1.2	Choosing a Single Fit for the Online Shopping Purchase Intent Project	438
13.2	Approach 2: PCS Ensemble	441
13.2.1	The Ensemble for the Ames House Price Project	444
13.2.2	The Ensemble for the Binary Online Shopping Project	445
13.3	Approach 3: Calibrated PCS Prediction Perturbation Intervals	447
13.3.1	Coverage and Calibration	450
13.3.2	Computing Prediction Perturbation Intervals	452
13.3.3	Perturbation Prediction Intervals for the Ames House Price Project	454
13.4	Choosing the Final Prediction Approach	456
13.5	Using Your Predictions in the Real World	457
	Exercises	457
14	Conclusion	463
14.1	Predictability	464
14.2	Stability and Uncertainty	465
14.3	Future PCS Directions: Inference	468
14.4	Farewell	469
	Answers to True or False Exercises	471
	Chapter 1	471
	Chapter 2	471
	Chapter 3	472
	Chapter 4	473
	Chapter 5	474
	Chapter 6	475
	Chapter 7	475
	Chapter 8	476
	Chapter 9	477
	Chapter 10	478
	Chapter 11	479
	Chapter 12	480
	Chapter 13	480
	References	483
	Index	489