Business Analytics

Methods & Cases for Data-Driven Decisions

Richard Huntsinger
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

Contents

About the Author xv
Preface xvi
Acknowledgements xxi

Executive Overview **1**

1 Data and Decisions **13**
Learning Objectives 14
1.1 Introduction 15
1.1.1 Let's Make a Deal 15
1.1.2 Let's Make Another Deal 17
1.1.3 Data Landscape 17
1.2 Data-to-Decision Process Model 20
1.2.1 Introduction 20
1.2.2 From Data to Decision 20
1.2.3 Process Diagram 21
1.3 Decision Models 23
1.3.1 Introduction 23
1.3.2 About Decision Models 23
1.3.3 Decision Model 28
1.4 Sensitivity Analysis 32
1.4.1 Introduction 32
1.4.2 About Sensitivity Analysis 33
1.4.3 Sensitivity to Decision Method Performance 33
1.4.4 Sensitivity to Business Parameter Values 34

2 Data Preparation **37**
Learning Objectives 38
2.1 Data Objects 39
2.1.1 Introduction 39
2.1.2 About Data Objects 39
2.1.3 Anatomy of a Dataset 40
2.1.4 Respecting Value Types 42

2.2 Selection 45
2.2.1 Introduction 45
2.2.2 About Selection 45
2.2.3 Index-Based Selection | Rows 45
2.2.4 Index-Based Selection | Columns 47
2.2.5 Index-Based Selection | Rows & Columns 47
2.2.6 Name-Based Selection | One Column 49
2.2.7 Name-Based Selection | Columns 50
2.2.8 Selection with Reorder 51
2.2.9 Selection with Random Reorder | Rows 52
2.2.10 Criterion-Based Selection | Rows 52
2.2.11 Criterion-Based Selection | Columns 54
2.3 Amalgamation 56
2.3.1 Introduction 56
2.3.2 About Amalgamation 56
2.3.3 Row-wise Concatenation 56
2.3.4 Column-wise Concatenation 58
2.3.5 Join by One Variable 59
2.3.6 Join by Many Variables 61
2.4 Synthetic Variables 63
2.4.1 Introduction 63
2.4.2 About Synthetic Variables 63
2.4.3 Synthetic Variables by Unit Conversion 64
2.4.4 Synthetic Variables by Linear Recombination 64
2.4.5 Synthetic Variables by Non-Linear Recombination 64
2.4.6 Synthetic Variables by Descriptive Statistic 65
2.4.7 Synthetic Variables by Lag 65
2.5 Normalization 67
2.5.1 Introduction 67
2.5.2 About Normalization 67
2.5.3 Normalization 68
2.6 Dummy Variables 71
2.6.1 Introduction 71
2.6.2 About Dummy Variables 71

2.6.3 Categorical to Dummy in a Dataset . . . 71
2.6.4 Categorical to Dummy in a New Observation . . . 73
2.7 CASE | High-Tech Stocks . . . 74
2.7.1 Business Situation . . . 74
2.7.2 Data . . . 74
2.7.3 Index-Based Selection | Rows . . . 75
2.7.4 Index-Based Selection | Rows & Columns . . . 76
2.7.5 Name-Based Selection | Rows & One Column . . . 78
2.7.6 Name-Based Selection | Rows & Columns . . . 79
2.7.7 Selection with Random Reorder | Rows . . . 79
2.7.8 Criterion-Based Selection | Rows . . . 80
2.7.9 Row-wise Concatenation . . . 82
2.7.10 Column-wise Concatenation . . . 84
2.7.11 Join . . . 85
2.7.12 Synthetic Variables by Descriptive Statistic . . . 86
2.7.13 Synthetic Variables by Lag . . . 86

3 Data Exploration 91

Learning Objectives . . . 92
3.1 Descriptive Statistics . . . 94
3.1.1 Introduction . . . 94
3.1.2 About Descriptive Statistics . . . 94
3.1.3 Data . . . 94
3.1.4 Descriptive Statistics of One Variable . . . 95
3.1.5 Descriptive Statistics of Two Variables . . . 97
3.2 Similarity . . . 99
3.2.1 Introduction . . . 99
3.2.2 About Similarity . . . 99
3.2.3 Euclidean Distance . . . 100
3.3 Cross-Tabulation . . . 103
3.3.1 Introduction . . . 103
3.3.2 About Cross-Tabulation . . . 103
3.3.3 Data . . . 104
3.3.4 Aggregate Table . . . 105

3.3.5 Long Table . . . 106
3.3.6 Cross-Table . . . 106
3.4 Data Visualization . . . 109
3.4.1 Introduction . . . 109
3.4.2 About Data Visualization . . . 111
3.4.3 Data . . . 111
3.4.4 One-Axis Scatterplot . . . 112
3.4.5 Two-Axis Scatterplot . . . 112
3.4.6 Three-Axis Scatterplot Projection . . . 114
3.4.7 Lineplot . . . 114
3.4.8 Stepplot . . . 116
3.4.9 Pathplot . . . 117
3.4.10 Bar Chart . . . 118
3.4.11 Histogram . . . 119
3.4.12 Pie Chart . . . 120
3.4.13 Violinplot . . . 120
3.4.14 Boxplot . . . 120
3.4.15 Heat Map & Conditional Format . . . 121
3.5 Kernel Density Estimation . . . 123
3.5.1 Introduction . . . 123
3.5.2 About Kernel Density Estimation . . . 123
3.5.3 Guess the Underlying Process . . . 124
3.5.4 Histogram Density Estimation | One Variable . . . 128
3.5.5 Kernel Density Estimation | One Variable . . . 130
3.5.6 Kernel Density Estimation | Two Variables . . . 135
3.5.7 Probability from Kernel Density Estimate . . . 137
3.6 CASE | Fundraising Strategy . . . 139
3.6.1 Business Situation . . . 139
3.6.2 Data . . . 139
3.6.3 Donations . . . 143
3.6.4 Donor Occupations . . . 149
3.6.5 Donor Cities . . . 151
3.6.6 Timing . . . 153
3.6.7 Comparative Analysis . . . 155

3.7 CASE | Iowa Liquor Sales . . . 156
3.7.1 Business Situation . . . 156
3.7.2 Data . . . 156
3.7.3 High-Level Trend Analysis . . . 159
3.7.4 Low-Level Trend Analysis . . . 159
3.7.5 Implications for Bundling & Pricing . . . 162

4 Data Transformation 165

Learning Objectives . . . 166
4.1 Balance . . . 167
4.1.1 Introduction . . . 167
4.1.2 About Balance . . . 167
4.1.3 Balance by Downsample . . . 168
4.1.4 Balance by Bootstrap . . . 169
4.1.5 Balance by Downsample & Bootstrap . . . 170
4.2 Imputation . . . 172
4.2.1 Introduction . . . 172
4.2.2 About Imputation . . . 172
4.2.3 Data . . . 173
4.2.4 Remove Observations with Missing Values . . . 173
4.2.5 Remove Variables with Missing Values . . . 173
4.2.6 Impute by Variable Mean . . . 174
4.2.7 Impute by Neighbor Mean . . . 174
4.2.8 Impute by Linear Interpolation . . . 175
4.2.9 Compare Imputation Methods . . . 175
4.3 Alignment . . . 176
4.3.1 Introduction . . . 176
4.3.2 About Alignment . . . 176
4.3.3 Data . . . 177
4.3.4 Alignment by Contraction . . . 177
4.3.5 Alignment by Expansion . . . 179
4.4 Principal Component Analysis . . . 181
4.4.1 Introduction . . . 181
4.4.2 About Principal Component Analysis . . . 182

4.4.3 Principal Component Analysis | Two Variables . . . 184
4.4.4 Principal Component Analysis | Two Normalized Variables . . . 187
4.4.5 Principal Component Analysis | Three Normalized Variables . . . 192
4.4.6 Transform a New Observation to Principal Component Representation . . . 195
4.4.7 An Analogy for Principal Component Analysis . . . 198
4.5 CASE | Loan Portfolio . . . 199
4.5.1 Business Situation . . . 199
4.5.2 Data . . . 199
4.5.3 Data Exploration . . . 203
4.5.4 Try to Distinguish Observations . . . 203
4.5.5 Principal Component Analysis . . . 205
4.5.6 Try Again to Distinguish Observations . . . 209
4.5.7 Decision about a New Loan . . . 211

5 Classification I **213**

Learning Objectives . . . 214
5.1 Classification Methodology . . . 216
5.1.1 Introduction . . . 216
5.1.2 About Classification Methodology . . . 217
5.1.3 Construction . . . 220
5.1.4 Prediction . . . 221
5.1.5 Evaluation . . . 222
5.2 Classifier Evaluation . . . 225
5.2.1 Introduction . . . 225
5.2.2 Confusion Matrix . . . 225
5.2.3 Performance Metrics . . . 227
5.2.4 Evaluation by In-Sample Performance . . . 228
5.2.5 Evaluation by Out-of-Sample Performance . . . 230
5.2.6 Evaluation by Cross-Validation Performance . . . 233
5.3 k-Nearest Neighbors . . . 239
5.3.1 Introduction . . . 239
5.3.2 About k-Nearest Neighbors . . . 239
5.3.3 k-Nearest Neighbors . . . 240
5.4 Logistic Regression . . . 244

5.4.1 Introduction . . . 244
5.4.2 About Logistic Regression . . . 244
5.4.3 Logistic Regression with One Predictor Variable . . . 245
5.4.4 Logistic Regression with Many Predictor Variables . . . 247
5.5 Decision Tree . . . 250
5.5.1 Introduction . . . 250
5.5.2 Prediction Based on Probabilities . . . 250
5.5.3 About Finding Best Splits . . . 253
5.5.4 Finding Best Splits . . . 255
5.5.5 Pruning . . . 257
5.6 CASE | Loan Portfolio Revisited . . . 259
5.6.1 Business Situation . . . 259
5.6.2 Decision Model . . . 259
5.6.3 Data . . . 260
5.6.4 Baseline . . . 264
5.6.5 Classifier Construction . . . 265
5.6.6 Classifier Evaluation by In-Sample Performance . . . 265
5.6.7 Classifier Evaluation by Cross-Validation Performance . . . 266
5.6.8 Classifier Tuning . . . 266

6 Classification II **269**

Learning Objectives . . . 270
6.1 Naive Bayes . . . 271
6.1.1 Introduction . . . 271
6.1.2 Naive Bayes | One Numerical Predictor Variable . . . 272
6.1.3 Naive Bayes | One Categorical Predictor Variable . . . 276
6.1.4 Naive Bayes | Two Predictor Variables . . . 280
6.1.5 Naive Bayes | Many Predictor Variables . . . 284
6.1.6 Naive Bayes | Gaussian Density Estimation . . . 285
6.1.7 LaPlace Smoothing . . . 287
6.1.8 Statistical Formulation of Naive Bayes Method . . . 290
6.2 Support Vector Machine . . . 293
6.2.1 Introduction . . . 293
6.2.2 Linearly Separable Data . . . 293

6.2.3 Support Vector Machine | One Predictor Variable & Linearly Separable Data . . . 294
6.2.4 Support Vector Machine | One Predictor Variable & Penalties . . . 297
6.2.5 Support Vector Machine | One Predictor Variable & Kernel Trick . . . 300
6.2.6 Support Vector Machine | More of One Predictor Variable & Kernel Trick 304
6.2.7 Support Vector Machine | Two Predictor Variables & Kernel Trick . . . 307
6.2.8 Support Vector Machine | Many Predictor Variables . . . 309
6.3 Neural Network . . . 310
6.3.1 Introduction . . . 310
6.3.2 About Neural Network . . . 310
6.3.3 Perceptron | Linearly Separable Data . . . 316
6.3.4 Perceptron | Non-Linearly Separable Data . . . 322
6.3.5 Two-Layer Neural Network . . . 323
6.3.6 Multi-Layer Neural Network . . . 327
6.3.7 Details of Back-propagation Algorithm . . . 329
6.4 CASE | Telecom Customer Churn . . . 332
6.4.1 Business Situation . . . 332
6.4.2 Decision Model . . . 332
6.4.3 Data . . . 333
6.4.4 Analysis . . . 337
6.4.5 Sensitivity Analysis . . . 345
6.5 CASE | Truck Fleet Maintenance . . . 347
6.5.1 Business Situation . . . 347
6.5.2 Decision Model . . . 348
6.5.3 Data . . . 348
6.5.4 Analysis I . . . 354
6.5.5 Custom Data for Test . . . 357
6.5.6 Analysis II . . . 361

7 Classification III **363**

Learning Objectives . . . 364
7.1 Multinomial Classification . . . 365
7.1.1 Introduction . . . 365
7.1.2 About Multinomial Classification . . . 366
7.1.3 Multinomial k-Nearest Neighbors . . . 367

7.1.4 Multinomial Decision Tree . . . 368
7.1.5 Multinomial Naive Bayes . . . 369
7.1.6 Multinomial Neural Network . . . 370
7.1.7 One vs. Rest . . . 371
7.1.8 One vs. One . . . 374
7.2 CASE | Facial Recognition . . . 377
7.2.1 Business Situation . . . 377
7.2.2 Decision Model . . . 377
7.2.3 Data . . . 378
7.2.4 Construct Models . . . 381
7.2.5 Evaluate Best Model . . . 382
7.3 CASE | Credit Card Fraud . . . 385
7.3.1 Business Situation . . . 385
7.3.2 Decision Model . . . 385
7.3.3 Data . . . 388
7.3.4 Baseline . . . 389
7.3.5 Predict Fraud by k-Nearest Neighbors . . . 390
7.3.6 Predict Fraud by Logistic Regression . . . 392
7.3.7 Predict Fraud by Decision Tree . . . 393
7.3.8 Predict Fraud by Naive Bayes . . . 394
7.3.9 Predict Fraud by Support Vector Machine . . . 395
7.3.10 Predict Fraud by Neural Network . . . 397
7.3.11 Predict Fraud by Stacking . . . 398
7.3.12 Comparative Analysis . . . 399

8 Regression **401**

Learning Objectives . . . 402
8.1 Regression Methodology . . . 404
8.1.1 Introduction . . . 404
8.1.2 About Regression Methodology . . . 405
8.1.3 Construction . . . 407
8.1.4 Prediction . . . 408
8.1.5 Evaluation . . . 409
8.2 Regressor Evaluation . . . 412

8.2.1 Introduction . . . 412
8.2.2 Error Table . . . 412
8.2.3 Performance Metrics . . . 413
8.2.4 Evaluation by In-Sample Performance . . . 414
8.2.5 Evaluation by Out-of-Sample Performance . . . 417
8.2.6 Evaluation by Cross-Validation Performance . . . 420
8.3 Linear Regression . . . 425
8.3.1 Introduction . . . 425
8.3.2 About Linear Regression . . . 425
8.3.3 Simple Linear Regression . . . 428
8.3.4 Multiple Linear Regression . . . 429
8.3.5 Linear Regression with Log Outcome Preparation . . . 431
8.3.6 Linear Regression with Log Predictor Preparation . . . 434
8.3.7 Linear Regression with Polynomial Predictor Preparation . . . 436
8.3.8 Linear Regression with Interaction Predictor Preparation . . . 438
8.3.9 Linear Regression with Categorical Predictor Variables . . . 441
8.3.10 Complex Linear Regression . . . 444
8.4 Regression Versions . . . 447
8.4.1 Introduction . . . 447
8.4.2 Data . . . 447
8.4.3 Regression by k-Nearest Neighbors . . . 448
8.4.4 Regression by Decision Tree . . . 449
8.4.5 Regression by Support Vector Machine . . . 450
8.4.6 Regression by Neural Network . . . 450
8.5 CASE | Call Center Scheduling . . . 452
8.5.1 Business Situation . . . 452
8.5.2 Decision Models . . . 452
8.5.3 Data . . . 456
8.5.4 Model A: Staffing Level is Set to Fixed High . . . 456
8.5.5 Model B: Staffing Level is Set to Fixed Low . . . 461
8.5.6 Model C: Staffing Level Is Set by Linear Regression . . . 463
8.5.7 Model D: Staffing Level Is Set by Linear Regression with Buffer . . . 465
8.5.8 Model E: Staffing Level Is Set by Piece-wise Linear Regression . . . 467
8.5.9 Comparative Analysis . . . 470

9 Ensemble Assembly **473**

Learning Objectives . . . 474

9.1 Bagging . . . 475

9.1.1 Introduction . . . 475

9.1.2 About Bagging . . . 476

9.1.3 Bagging . . . 476

9.2 Boosting . . . 479

9.2.1 Introduction . . . 479

9.2.2 About Boosting . . . 480

9.2.3 Boosting . . . 480

9.3 Stacking . . . 486

9.3.1 Introduction . . . 486

9.3.2 About Stacking . . . 487

9.3.3 Stacking . . . 487

10 Cluster Analysis **491**

Learning Objectives . . . 492

10.1 Cluster Analysis Methodology . . . 494

10.1.1 Introduction . . . 494

10.1.2 About Cluster Analysis Methodology . . . 495

10.1.3 Cluster Analysis Methodology . . . 497

10.2 Cluster Model Evaluation . . . 505

10.2.1 Introduction . . . 505

10.2.2 About Cluster Model Evaluation . . . 505

10.2.3 Cluster Model Evaluation by Dispersion Ratio . . . 506

10.3 k-Means . . . 513

10.3.1 Introduction . . . 513

10.3.2 About k-Means . . . 513

10.3.3 k-Means . . . 514

10.4 Hierarchical Agglomeration . . . 520

10.4.1 Introduction . . . 520

10.4.2 About Hierarchical Agglomeration . . . 521

10.4.3 Hierarchical Agglomeration . . . 522

10.4.4 More Dendrogram Analysis . . . 529

10.5 Gaussian Mixture . . . 531
10.5.1 Introduction . . . 531
10.5.2 About Gaussian Mixture . . . 532
10.5.3 Gaussian Mixture | One Variable, Two Classes . . . 535
10.5.4 Gaussian Mixture | One Variable, Three Classes . . . 540
10.5.5 More About Gaussian Mixture . . . 546
10.5.6 Gaussian Mixture | Two Variables, Two Classes . . . 549
10.5.7 Gaussian Mixture | Two Variables, Three Classes . . . 555
10.5.8 Gaussian Mixture | Many Variables, Many Classes . . . 561
10.6 CASE | Fortune 500 Diversity . . . 563
10.6.1 Business Situation . . . 563
10.6.2 Data . . . 563
10.6.3 Analysis I . . . 565
10.6.4 Analysis II . . . 569
10.7 CASE | Music Market Segmentation . . . 571
10.7.1 Business Situation . . . 571
10.7.2 Decision Model . . . 572
10.7.3 Data . . . 572
10.7.4 Market Segmentation Models . . . 574
10.7.5 Evaluation . . . 576
10.7.6 Business Results . . . 578
10.7.7 Sensitivity Analysis . . . 579
10.7.8 Reveal . . . 580

11 Special Data Types **585**

Learning Objectives . . . 586
11.1 Text Data . . . 588
11.1.1 Introduction . . . 588
11.1.2 About Text Data . . . 588
11.1.3 Data . . . 590
11.1.4 Transform . . . 590
11.1.5 Model . . . 593
11.1.6 Predict . . . 594
11.2 Time Series Data . . . 596

11.2.1 Introduction . . . 596
11.2.2 About Time Series Data . . . 596
11.2.3 Data . . . 599
11.2.4 Prepare Data . . . 599
11.2.5 Construct Models . . . 601
11.2.6 Forecast . . . 602
11.2.7 Evaluate by In-Sample Performance . . . 605
11.2.8 Evaluate by Time Series Out-of-Sample Performance . . . 607
11.2.9 Evaluate by Time Series Cross-Validation Performance . . . 610
11.3 Network Data . . . 615
11.3.1 Introduction . . . 615
11.3.2 About Network Data . . . 615
11.3.3 Data . . . 618
11.3.4 Node-Level Statistics . . . 619
11.3.5 Network-Level Statistics . . . 621
11.4 PageRank for Network Data . . . 622
11.4.1 Introduction . . . 622
11.4.2 About PageRank . . . 622
11.4.3 Data . . . 624
11.4.4 Random Walk . . . 624
11.4.5 PageRank . . . 625
11.4.6 PageRank with Dampening . . . 628
11.5 Collaborative Filtering for Network Data . . . 632
11.5.1 Introduction . . . 632
11.5.2 About Collaborative Filtering . . . 632
11.5.3 Data . . . 634
11.5.4 Similarity Matrix . . . 636
11.5.5 Simple Collaborative Filtering . . . 636
11.5.6 Collaborative Filtering with Weighting . . . 637
11.5.7 Collaborative Filtering with Calibrated Weighting . . . 638
11.5.8 Collaborative Filtering with Calibrated Weighting, Item-Based . . . 640
11.5.9 Recommendation . . . 641
11.6 CASE | Deceptive Hotel Reviews . . . 642
11.6.1 Business Situation . . . 642

11.6.2 Data . . . 643
11.6.3 Transform Reference Dataset . . . 644
11.6.4 Model . . . 645
11.6.5 Predict . . . 646
11.6.6 Business Result . . . 649
11.7 CASE | Targeted Marketing . . . 650
11.7.1 Business Situation . . . 650
11.7.2 Decision Model . . . 650
11.7.3 Data . . . 652
11.7.4 Descriptive Statistics . . . 654
11.7.5 Important Managers . . . 655
11.7.6 Decision Method #1: No Marketing . . . 656
11.7.7 Decision Method #2: Non-Targeted Marketing . . . 659
11.7.8 Decision Method #3: Targeted Marketing to Random Prospects . . . 660
11.7.9 Decision Method #4: Targeted Marketing to Important Prospects . . . 661
11.7.10 Comparative Analysis . . . 663
11.7.11 Reveal . . . 663

Epilog . . . 666
Photo Credits . . . 667
Index . . . 669