

---

# Data Science for Business

*Foster Provost and Tom Fawcett*

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

 O'REILLY®

---

# Table of Contents

<b>Preface</b> .....	<b>xiii</b>
<b>1. Introduction: Data-Analytic Thinking</b> .....	<b>1</b>
The Ubiquity of Data Opportunities	1
Example: Hurricane Frances	3
Example: Predicting Customer Churn	4
Data Science, Engineering, and Data-Driven Decision Making	4
Data Processing and “Big Data”	7
From Big Data 1.0 to Big Data 2.0	8
Data and Data Science Capability as a Strategic Asset	9
Data-Analytic Thinking	12
This Book	14
Data Mining and Data Science, Revisited	14
Chemistry Is Not About Test Tubes: Data Science Versus the Work of the Data Scientist	15
Summary	16
<b>2. Business Problems and Data Science Solutions</b> .....	<b>19</b>
<i>Fundamental concepts: A set of canonical data mining tasks; The data mining process; Supervised versus unsupervised data mining.</i>	
From Business Problems to Data Mining Tasks	19
Supervised Versus Unsupervised Methods	24
Data Mining and Its Results	25
The Data Mining Process	26
Business Understanding	27
Data Understanding	28
Data Preparation	29
Modeling	31
Evaluation	31

Deployment	32
Implications for Managing the Data Science Team	34
Other Analytics Techniques and Technologies	35
Statistics	35
Database Querying	37
Data Warehousing	38
Regression Analysis	39
Machine Learning and Data Mining	39
Answering Business Questions with These Techniques	40
Summary	41
<b>3. Introduction to Predictive Modeling: From Correlation to Supervised Segmentation.</b>	<b>43</b>
<i>Fundamental concepts: Identifying informative attributes; Segmenting data by progressive attribute selection.</i>	
<i>Exemplary techniques: Finding correlations; Attribute/variable selection; Tree induction.</i>	
Models, Induction, and Prediction	44
Supervised Segmentation	48
Selecting Informative Attributes	49
Example: Attribute Selection with Information Gain	56
Supervised Segmentation with Tree-Structured Models	62
Visualizing Segmentations	67
Trees as Sets of Rules	71
Probability Estimation	71
Example: Addressing the Churn Problem with Tree Induction	73
Summary	78
<b>4. Fitting a Model to Data.</b>	<b>81</b>
<i>Fundamental concepts: Finding “optimal” model parameters based on data; Choosing the goal for data mining; Objective functions; Loss functions.</i>	
<i>Exemplary techniques: Linear regression; Logistic regression; Support-vector machines.</i>	
Classification via Mathematical Functions	83
Linear Discriminant Functions	85
Optimizing an Objective Function	88
An Example of Mining a Linear Discriminant from Data	89
Linear Discriminant Functions for Scoring and Ranking Instances	91
Support Vector Machines, Briefly	92
Regression via Mathematical Functions	95
Class Probability Estimation and Logistic “Regression”	97
* Logistic Regression: Some Technical Details	100
Example: Logistic Regression versus Tree Induction	103
Nonlinear Functions, Support Vector Machines, and Neural Networks	107

Summary	110
<b>5. Overfitting and Its Avoidance.....</b>	<b>111</b>
<i>Fundamental concepts: Generalization; Fitting and overfitting; Complexity control.</i>	
<i>Exemplary techniques: Cross-validation; Attribute selection; Tree pruning; Regularization.</i>	
Generalization	111
Overfitting	113
Overfitting Examined	113
Holdout Data and Fitting Graphs	113
Overfitting in Tree Induction	116
Overfitting in Mathematical Functions	118
Example: Overfitting Linear Functions	119
* Example: Why Is Overfitting Bad?	124
From Holdout Evaluation to Cross-Validation	126
The Churn Dataset Revisited	129
Learning Curves	130
Overfitting Avoidance and Complexity Control	133
Avoiding Overfitting with Tree Induction	133
A General Method for Avoiding Overfitting	134
* Avoiding Overfitting for Parameter Optimization	136
Summary	140
<b>6. Similarity, Neighbors, and Clusters.....</b>	<b>141</b>
<i>Fundamental concepts: Calculating similarity of objects described by data; Using similarity for prediction; Clustering as similarity-based segmentation.</i>	
<i>Exemplary techniques: Searching for similar entities; Nearest neighbor methods; Clustering methods; Distance metrics for calculating similarity.</i>	
Similarity and Distance	142
Nearest-Neighbor Reasoning	144
Example: Whiskey Analytics	145
Nearest Neighbors for Predictive Modeling	147
How Many Neighbors and How Much Influence?	149
Geometric Interpretation, Overfitting, and Complexity Control	151
Issues with Nearest-Neighbor Methods	155
Some Important Technical Details Relating to Similarities and Neighbors	157
Heterogeneous Attributes	157
* Other Distance Functions	158
* Combining Functions: Calculating Scores from Neighbors	162
Clustering	163
Example: Whiskey Analytics Revisited	164
Hierarchical Clustering	165

Nearest Neighbors Revisited: Clustering Around Centroids	170
Example: Clustering Business News Stories	175
Understanding the Results of Clustering	178
* Using Supervised Learning to Generate Cluster Descriptions	180
Stepping Back: Solving a Business Problem Versus Data Exploration	183
Summary	185
<b>7. Decision Analytic Thinking I: What Is a Good Model?.....</b>	<b>187</b>
<i>Fundamental concepts: Careful consideration of what is desired from data science results; Expected value as a key evaluation framework; Consideration of appropriate comparative baselines.</i>	
<i>Exemplary techniques: Various evaluation metrics; Estimating costs and benefits; Calculating expected profit; Creating baseline methods for comparison.</i>	
Evaluating Classifiers	188
Plain Accuracy and Its Problems	189
The Confusion Matrix	189
Problems with Unbalanced Classes	190
Problems with Unequal Costs and Benefits	193
Generalizing Beyond Classification	193
A Key Analytical Framework: Expected Value	194
Using Expected Value to Frame Classifier Use	195
Using Expected Value to Frame Classifier Evaluation	196
Evaluation, Baseline Performance, and Implications for Investments in Data	204
Summary	207
<b>8. Visualizing Model Performance.....</b>	<b>209</b>
<i>Fundamental concepts: Visualization of model performance under various kinds of uncertainty; Further consideration of what is desired from data mining results.</i>	
<i>Exemplary techniques: Profit curves; Cumulative response curves; Lift curves; ROC curves.</i>	
Ranking Instead of Classifying	209
Profit Curves	212
ROC Graphs and Curves	214
The Area Under the ROC Curve (AUC)	219
Cumulative Response and Lift Curves	219
Example: Performance Analytics for Churn Modeling	223
Summary	231
<b>9. Evidence and Probabilities.....</b>	<b>233</b>
<i>Fundamental concepts: Explicit evidence combination with Bayes' Rule; Probabilistic reasoning via assumptions of conditional independence.</i>	
<i>Exemplary techniques: Naive Bayes classification; Evidence lift.</i>	

Example: Targeting Online Consumers With Advertisements	233
Combining Evidence Probabilistically	235
Joint Probability and Independence	236
Bayes' Rule	237
Applying Bayes' Rule to Data Science	239
Conditional Independence and Naive Bayes	241
Advantages and Disadvantages of Naive Bayes	243
A Model of Evidence "Lift"	244
Example: Evidence Lifts from Facebook "Likes"	246
Evidence in Action: Targeting Consumers with Ads	248
Summary	248
<b>10. Representing and Mining Text. ....</b>	<b>251</b>
<i>Fundamental concepts: The importance of constructing mining-friendly data representations; Representation of text for data mining.</i>	
<i>Exemplary techniques: Bag of words representation; TFIDF calculation; N-grams; Stemming; Named entity extraction; Topic models.</i>	
Why Text Is Important	252
Why Text Is Difficult	252
Representation	253
Bag of Words	254
Term Frequency	254
Measuring Sparseness: Inverse Document Frequency	256
Combining Them: TFIDF	258
Example: Jazz Musicians	258
* The Relationship of IDF to Entropy	263
Beyond Bag of Words	265
N-gram Sequences	265
Named Entity Extraction	266
Topic Models	266
Example: Mining News Stories to Predict Stock Price Movement	268
The Task	268
The Data	270
Data Preprocessing	272
Results	273
Summary	277
<b>11. Decision Analytic Thinking II: Toward Analytical Engineering. ....</b>	<b>279</b>
<i>Fundamental concept: Solving business problems with data science starts with analytical engineering: designing an analytical solution, based on the data, tools, and techniques available.</i>	
<i>Exemplary technique: Expected value as a framework for data science solution design.</i>	

Targeting the Best Prospects for a Charity Mailing	280
The Expected Value Framework: Decomposing the Business Problem and Recomposing the Solution Pieces	280
A Brief Digression on Selection Bias	282
Our Churn Example Revisited with Even More Sophistication	283
The Expected Value Framework: Structuring a More Complicated Business Problem	283
Assessing the Influence of the Incentive	285
From an Expected Value Decomposition to a Data Science Solution	286
Summary	289
<b>12. Other Data Science Tasks and Techniques</b>	<b>291</b>
<i>Fundamental concepts: Our fundamental concepts as the basis of many common data science techniques; The importance of familiarity with the building blocks of data science.</i>	
<i>Exemplary techniques: Association and co-occurrences; Behavior profiling; Link prediction; Data reduction; Latent information mining; Movie recommendation; Bias-variance decomposition of error; Ensembles of models; Causal reasoning from data.</i>	
Co-occurrences and Associations: Finding Items That Go Together	292
Measuring Surprise: Lift and Leverage	293
Example: Beer and Lottery Tickets	294
Associations Among Facebook Likes	295
Profiling: Finding Typical Behavior	298
Link Prediction and Social Recommendation	303
Data Reduction, Latent Information, and Movie Recommendation	304
Bias, Variance, and Ensemble Methods	308
Data-Driven Causal Explanation and a Viral Marketing Example	311
Summary	312
<b>13. Data Science and Business Strategy</b>	<b>315</b>
<i>Fundamental concepts: Our principles as the basis of success for a data-driven business; Acquiring and sustaining competitive advantage via data science; The importance of careful curation of data science capability.</i>	
Thinking Data-Analytically, Redux	315
Achieving Competitive Advantage with Data Science	317
Sustaining Competitive Advantage with Data Science	318
Formidable Historical Advantage	319
Unique Intellectual Property	319
Unique Intangible Collateral Assets	320
Superior Data Scientists	320
Superior Data Science Management	322
Attracting and Nurturing Data Scientists and Their Teams	323

Examine Data Science Case Studies	325
Be Ready to Accept Creative Ideas from Any Source	326
Be Ready to Evaluate Proposals for Data Science Projects	326
Example Data Mining Proposal	327
Flaws in the Big Red Proposal	328
A Firm's Data Science Maturity	329
<b>14. Conclusion.....</b>	<b>333</b>
The Fundamental Concepts of Data Science	333
Applying Our Fundamental Concepts to a New Problem: Mining Mobile Device Data	336
Changing the Way We Think about Solutions to Business Problems	339
What Data Can't Do: Humans in the Loop, Revisited	340
Privacy, Ethics, and Mining Data About Individuals	343
Is There More to Data Science?	344
Final Example: From Crowd-Sourcing to Cloud-Sourcing	345
Final Words	346
<b>A. Proposal Review Guide.....</b>	<b>349</b>
<b>B. Another Sample Proposal.....</b>	<b>353</b>
<b>Glossary.....</b>	<b>357</b>
<b>Bibliography.....</b>	<b>361</b>
<b>Index.....</b>	<b>369</b>