## **Data Science for Business**

Foster Provost and Tom Fawcett



## **Data Science for Business**

by Foster Provost and Tom Fawcett

Copyright © 2013 Foster Provost and Tom Fawcett. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/ institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Mike Loukides and Meghan Blanchette Production Editor: Christopher Hearse Proofreader: Kiel Van Horn Indexer: WordCo Indexing Services, Inc. **Cover Designer:** Mark Paglietti **Interior Designer:** David Futato **Illustrator:** Rebecca Demarest

July 2013: First Edition

## **Revision History for the First Edition:**

2013-07-25: First release

See http://oreilly.com/catalog/errata.csp?isbn=9781449361327 for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps. *Data Science for Business* is a trademark of Foster Provost and Tom Fawcett.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-36132-7

[LSI]

## **Table of Contents**

Pre	face	xi
1.	Introduction: Data-Analytic Thinking	. 1
	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
2.	Business Problems and Data Science Solutions	19
	Fundamental concepts: A set of canonical data mining tasks; The data mining proce. Supervised versus unsupervised data mining.	ss;
	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 Implications for Managing the Data Science Team Other Analytics Techniques and Technologies Statistics Database Querying Data Warehousing Regression Analysis Machine Learning and Data Mining Answering Business Questions with These Techniques Summary	32 34 35 35 37 38 39 39 40 41
3.	<b>Introduction to Predictive Modeling: From Correlation to Supervised Segmentation.</b> Fundamental concepts: Identifying informative attributes; Segmenting data by progressive attribute selection. Exemplary techniques: Finding correlations; Attribute/variable selection; Tree induction	43
	Models Induction and Production	4.4
	Supervised Segmentation	44
	Selecting Informative Attributes	40
	Example: Attribute Selection with Information Cain	49
	Example: Attribute Selection with Tree Structured Models	50
	Visualizing Segmentations	67
	Trace of Sets of Dules	71
	Drahakilita Datimation	/1
	Probability Estimation	/1
	Example: Addressing the Churn Problem with Tree Induction	73
	Summary	/8
4.	<b>Fitting a Model to Data</b> . Fundamental concepts: Finding "optimal" model parameters based on data; Choos the goal for data mining; Objective functions; Loss functions.	<b>. 81</b> ing
	Exemplary techniques: Linear regression; Logistic regression; Support-vector machine	nes.
	Classification via Mathematical Functions	83
	Linear Discriminant Functions	85
	Optimizing an Objective Function	87
	An Example of Mining a Linear Discriminant from Data	88
	Linear Discriminant Functions for Scoring and Ranking Instances	90
	Support Vector Machines, Briefly	91
	Regression via Mathematical Functions	94
	Class Probability Estimation and Logistic "Regression"	96
	* Logistic Regression: Some Technical Details	99
	Example: Logistic Regression versus Tree Induction	102
	Nonlinear Functions, Support Vector Machines, and Neural Networks	105

Summary

5.	<b>Overfitting and Its Avoidance</b> Fundamental concepts: Generalization: Fittina and overfittina: Complexity control	111
	Exemplary techniques: Cross-validation: Attribute selection: Tree pruning:	
	Regularization.	
	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
6.	Similarity, Neighbors, and Clusters	141
	Fundamental concepts: Calculating similarity of objects described by data; Using	
	similarity for prediction; Clustering as similarity-based segmentation.	
	Exemplary techniques: Searching for similar entities; Nearest neighbor methods; Clustering methods; Distance metrics for calculating similarity.	
	Similarity and Distance	142
	Nearest-Neighbor Reasoning	144
	Example: Whiskey Analytics	144
	Nearest Neighbors for Predictive Modeling	146
	How Many Neighbors and How Much Influence?	149
	Geometric Interpretation, Overfitting, and Complexity Control	151
	Issues with Nearest-Neighbor Methods	154
	Some Important Technical Details Relating to Similarities and Neighbors	157
	Heterogeneous Attributes	157
	* Other Distance Functions	158
	* Combining Functions: Calculating Scores from Neighbors	161
	Clustering	163
	Example: Whiskey Analytics Revisited	163
	Hierarchical Clustering	164

	Nearest Neighbors Revisited: Clustering Around Centroids	169
	Example: Clustering Business News Stories	174
	Understanding the Results of Clustering	177
	* Using Supervised Learning to Generate Cluster Descriptions	179
	Stepping Back: Solving a Business Problem Versus Data Exploration	182
	Summary	184
7.	<b>Decision Analytic Thinking I: What Is a Good Model?</b> <i>Fundamental concepts: Careful consideration of what is desired from data science results; Expected value as a key evaluation framework; Consideration of appropriat comparative baselines.</i> <i>Exemplary techniques: Various evaluation metrics; Estimating costs and benefits;</i>	<b>187</b> e
	Calculating expected profit; Creating baseline methods for comparison.	
	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
8.	<b>Visualizing Model Performance</b> . Fundamental concepts: Visualization of model performance under various kinds of uncertainty; Further consideration of what is desired from data mining results.	209
	curves	
	Ranking Instead of Classifying	209
	Profit Curves	207
	ROC Graphs and Curves	212
	The Area Under the ROC Curve (AUC)	219
	Cumulative Response and Lift Curves	219
	Example: Performance Analytics for Churn Modeling	223
	Summary	231
9.	<b>Evidence and Probabilities.</b> Fundamental concepts: Explicit evidence combination with Bayes' Rule; Probabilisti reasoning via assumptions of conditional independence. Exemplary techniques: Naive Bayes classification; Evidence lift.	<b>233</b> ic

	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	240
	Advantages and Disadvantages of Naive Bayes	242
	A Model of Evidence "Lift"	244
	Example: Evidence Lifts from Facebook "Likes"	245
	Evidence in Action: Targeting Consumers with Ads	247
	Summary	247
10.	<b>Representing and Mining Text.</b> Fundamental concepts: The importance of constructing mining-friendly data representations; Representation of text for data mining.	249
	Exemplary techniques: Bag of words representation; TFIDF calculation; N-grams; Stemming; Named entity extraction; Topic models.	
	Why Text Is Important	250
	Why Text Is Difficult	250
	Representation	251
	Bag of Words	252
	Term Frequency	252
	Measuring Sparseness: Inverse Document Frequency	254
	Combining Them: TFIDF	256
	Example: Jazz Musicians	256
	* The Relationship of IDF to Entropy	261
	Beyond Bag of Words	263
	N-gram Sequences	263
	Named Entity Extraction	264
	Topic Models	264
	Example: Mining News Stories to Predict Stock Price Movement	266
	The Task	266
	The Data	268
	Data Preprocessing	270
	Results	271
	Summary	275
11.	<b>Decision Analytic Thinking II: Toward Analytical Engineering.</b> Fundamental concept: Solving business problems with data science starts with analytical engineering: designing an analytical solution, based on the data, tools, a techniques available. Exemplary technique: Expected value as a framework for data science solution desi	<b>277</b> Ind an.

	Targeting the Best Prospects for a Charity Mailing The Expected Value Framework: Decomposing the Business Problem and	278
	Recomposing the Solution Pieces	278
	A Brief Digression on Selection Bias	280
	Our Churn Example Revisited with Even More Sonhistication	200
	The Expected Value Framework: Structuring a More Complicated Business	201
	Droblam	201
	Producting the Influence of the Incentive	201
	Assessing the influence of the incentive	203
	Summary	284 287
	Summery	207
12.	Other Data Science Tasks and Techniques	289
	Fundamental concepts: Our fundamental concepts as the basis of many common or science techniques; The importance of familiarity with the building blocks of data science.	lata
	Exemplary techniques: Association and co-occurrences; Behavior profiling; Link prediction; Data reduction; Latent information mining; Movie recommendation; Bio variance decomposition of error: Ensembles of models: Causal reasoning from data	as-
	Co-occurrences and Associations: Finding items That Go Together	290
	Measuring Surprise: Lift and Leverage	291
	Example: Beer and Lottery Tickets	292
	Associations Among Facebook Likes	293
	Profiling: Finding Typical Behavior	296
	Link Prediction and Social Recommendation	301
	Data Reduction, Latent Information, and Movie Recommendation	302
	Bias, Variance, and Ensemble Methods	306
	Data-Driven Causal Explanation and a Viral Marketing Example	309
	Summary	310
13.	Data Science and Business Strategy.	313
	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.	
	Thinking Data-Analytically, Redux	313
	Achieving Competitive Advantage with Data Science	315
	Sustaining Competitive Advantage with Data Science	316
	Formidable Historical Advantage	317
	Unique Intellectual Property	317
	Unique Intangible Collateral Assets	318
	Superior Data Scientists	318
	Superior Data Science Management	320
	Attracting and Nurturing Data Scientists and Their Teams	321

	Examine Data Science Case Studies Be Ready to Accept Creative Ideas from Any Source Be Ready to Evaluate Proposals for Data Science Projects Example Data Mining Proposal Flaws in the Big Red Proposal A Firm's Data Science Maturity	<ul> <li>323</li> <li>324</li> <li>324</li> <li>325</li> <li>326</li> <li>327</li> </ul>
14.	<b>Conclusion.</b> The Fundamental Concepts of Data Science Applying Our Fundamental Concepts to a New Problem: Mining Mobile Device Data Changing the Way We Think about Solutions to Business Problems What Data Can't Do: Humans in the Loop, Revisited Privacy, Ethics, and Mining Data About Individuals Is There More to Data Science? Final Example: From Crowd Sourcing to Cloud Sourcing	<ul> <li>331</li> <li>331</li> <li>334</li> <li>337</li> <li>338</li> <li>341</li> <li>342</li> <li>343</li> </ul>
	Final Words	344 344
A.	Proposal Review Guide	347
B.	Another Sample Proposal	351
Glo	Glossary	
Bibliography		359
Index		