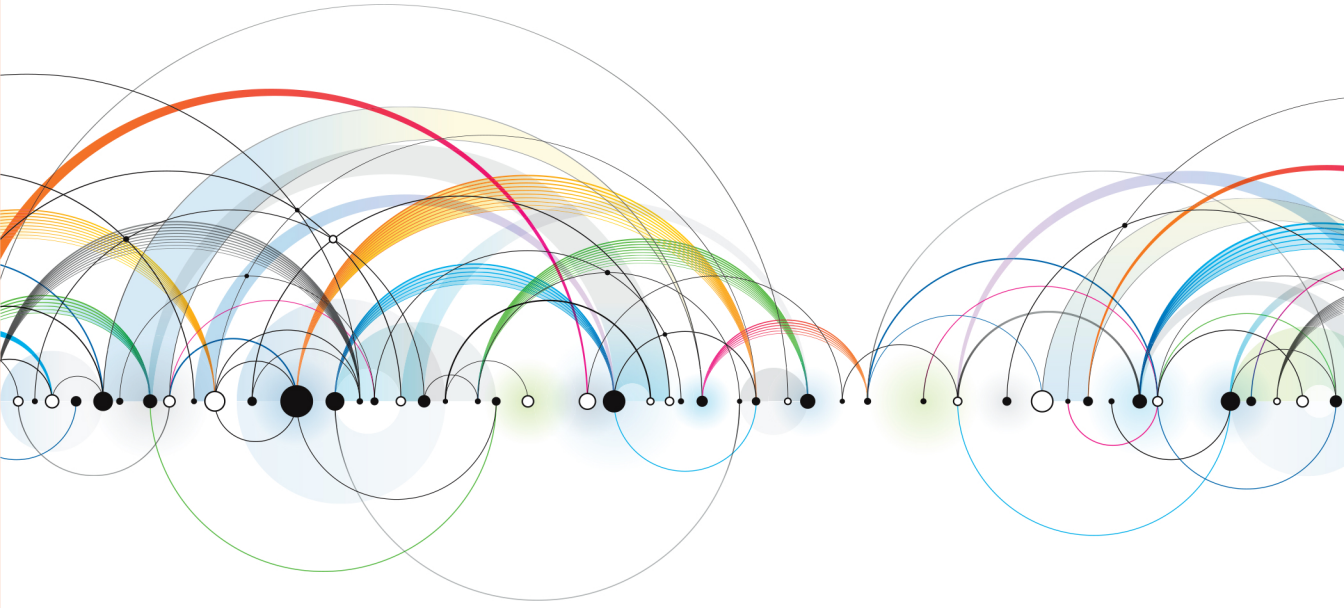
"A must-read resource for anyone who is serious about embracing the opportunity of big data."

—**Craig Vaughan, Global Vice President, SAP**

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking



Foster Provost & Tom Fawcett

www.it-ebooks.info

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking

This broad, deep, but not-too-technical guide introduces you to the fundamental principles of data science and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. By learning data science principles, you will understand the many data mining techniques in use today. More importantly, these principles underpin the processes and strategies necessary to solve business problems through data mining techniques.

"This book goes beyond data analytics 101. It's the essential guide for those of us (all of us?) whose businesses are built on the ubiquity of data opportunities and the new mandate for data-driven decision-making."

—Tom Phillips, CEO Media6Degrees; former Head of Google Search and Analytics

"The authors, both renowned experts in data science before it had a name, have taken a complex topic and made it accessible to all levels. This is the first book of its kind, with a focus on data science concepts as applied to practical business problems. It is liberally sprinkled with compelling real-world examples outlining familiar, accessible problems in the business world: customer churn, targeted marketing, even whiskey analytics!

The book is unique in that it does not give a cookbook of algorithms, rather it helps the reader understand the underlying concepts behind data science, and most importantly how to approach and be successful at problem solving. Whether you are looking for a good comprehensive overview of data science or are a budding data scientist in need of the basics, this is a must-read."

—Chris Volinsky, Director, Statistics Research, AT&T Labs
Winner of the $1 Million Netflix Challenge

"Data is the foundation of new waves of productivity growth, innovation, and richer customer insight. Only recently viewed broadly as a source of competitive advantage, dealing well with data is rapidly becoming table stakes to stay in the game. The authors' deep applied experience makes this a must read—a window into your competitor's strategy."

—Alan Murray, Serial Entrepreneur; Partner Coriolis Ventures

US $34.99 CAN $36.99
ISBN: 978-1-449-36132-7

Twitter: @oreillymedia
facebook.com/oreilly

O'REILLY®
oreilly.com

www.it-ebooks.info

Data Science for Business

Foster Provost and Tom Fawcett

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

www.it-ebooks.info

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]

www.it-ebooks.info

Table of Contents

Preface 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 process; Supervised versus unsupervised data mining.
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

www.it-ebooks.info

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

3. Introduction to Predictive Modeling: From Correlation to Supervised Segmentation. 43
Fundamental concepts: Identifying informative attributes; Segmenting data by progressive attribute selection.
Exemplary techniques: Finding correlations; Attribute/variable selection; Tree induction.
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

4. Fitting a Model to Data. 81
Fundamental concepts: Finding "optimal" model parameters based on data; Choosing the goal for data mining; Objective functions; Loss functions.
Exemplary techniques: Linear regression; Logistic regression; Support-vector machines.
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

www.it-ebooks.info

Summary 108

5. Overfitting and Its Avoidance. 111
Fundamental concepts: Generalization; Fitting and overfitting; Complexity control.
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

www.it-ebooks.info

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. Decision Analytic Thinking I: What Is a Good Model? 187
Fundamental concepts: Careful consideration of what is desired from data science results; Expected value as a key evaluation framework; Consideration of appropriate comparative baselines.
Exemplary techniques: Various evaluation metrics; Estimating costs and benefits; 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. Visualizing Model Performance 209
Fundamental concepts: Visualization of model performance under various kinds of uncertainty; Further consideration of what is desired from data mining results.
Exemplary techniques: Profit curves; Cumulative response curves; Lift curves; ROC curves.
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

9. Evidence and Probabilities 233
Fundamental concepts: Explicit evidence combination with Bayes' Rule; Probabilistic reasoning via assumptions of conditional independence.
Exemplary techniques: Naive Bayes classification; Evidence lift.

www.it-ebooks.info

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. Representing and Mining Text. 249
Fundamental concepts: The importance of constructing mining-friendly data representations; Representation of text for data mining.
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. Decision Analytic Thinking II: Toward Analytical Engineering. 277
Fundamental concept: Solving business problems with data science starts with analytical engineering: designing an analytical solution, based on the data, tools, and techniques available.
Exemplary technique: Expected value as a framework for data science solution design.

www.it-ebooks.info

Targeting the Best Prospects for a Charity Mailing 278
The Expected Value Framework: Decomposing the Business Problem and Recomposing the Solution Pieces 278
A Brief Digression on Selection Bias 280
Our Churn Example Revisited with Even More Sophistication 281
The Expected Value Framework: Structuring a More Complicated Business Problem 281
Assessing the Influence of the Incentive 283
From an Expected Value Decomposition to a Data Science Solution 284
Summary 287

12. Other Data Science Tasks and Techniques. 289
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.
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.
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

www.it-ebooks.info

Examine Data Science Case Studies 323
Be Ready to Accept Creative Ideas from Any Source 324
Be Ready to Evaluate Proposals for Data Science Projects 324
Example Data Mining Proposal 325
Flaws in the Big Red Proposal 326
A Firm's Data Science Maturity 327

14. Conclusion. 331
The Fundamental Concepts of Data Science 331
Applying Our Fundamental Concepts to a New Problem: Mining Mobile Device Data 334
Changing the Way We Think about Solutions to Business Problems 337
What Data Can't Do: Humans in the Loop, Revisited 338
Privacy, Ethics, and Mining Data About Individuals 341
Is There More to Data Science? 342
Final Example: From Crowd-Sourcing to Cloud-Sourcing 343
Final Words 344

A. Proposal Review Guide. 347

B. Another Sample Proposal. 351

Glossary. 355

Bibliography. 359

Index. 367

www.it-ebooks.info