

Wiley Series on Methods and
Applications in Data Mining

DATA SCIENCE

USING PYTHON AND R

CHANTAL D. LAROSE | DANIEL T. LAROSE



WILEY

*DATA SCIENCE USING
PYTHON AND R*

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: **Daniel T. Larose**

Practical Text Mining with Perl • Roger Bilisoly

Knowledge Discovery Support Vector Machines • Lutz Hamel

Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data • Darius M. Dziuda

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition
• Daniel T. Larose and Chantal D. Larose

Data Mining and Predictive Analytics • Daniel T. Larose and Chantal D. Larose

Data Mining and Learning Analytics: Applications in Educational Research •
Samira ElAtia, Donald Ipperciel, and Osmar R. Zaiane

Pattern Recognition: A Quality of Data Perspective • Władysław Homenda and
Witold Pedrycz

DATA SCIENCE USING PYTHON AND R

CHANTAL D. LAROSE

*Eastern Connecticut State University
Windham, CT, USA*

DANIEL T. LAROSE

*Central Connecticut State University
New Britain, CT, USA*

WILEY

This edition first published 2019
© 2019 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Chantal D. Larose and Daniel T. Larose to be identified as the authors of this work has been asserted in accordance with law.

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Larose, Chantal D., author. | Larose, Daniel T., author.

Title: Data science using Python and R / Chantal D. Larose, Eastern Connecticut State University, Connecticut, USA, Daniel T. Larose, Central Connecticut State University, Connecticut, USA.

Description: Hoboken, NJ : John Wiley & Sons, Inc, 2019. | Includes index. |

Identifiers: LCCN 2019007280 (print) | LCCN 2019009632 (ebook) | ISBN 9781119526834 (Adobe PDF) | ISBN 9781119526841 (ePub) | ISBN 9781119526810 (hardback)

Subjects: LCSH: Data mining. | Python (Computer program language) |

R (Computer program language) | Big data. | Data structures (Computer science)

Classification: LCC QA76.9.D343 (ebook) | LCC QA76.9.D343 L376 2019 (print) |

DDC 006.3/12–dc23

LC record available at <https://lcn.loc.gov/2019007280>

Cover Design: Wiley

Cover Image: © LumenGraphics/Shutterstock

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

<i>PREFACE</i>	xi
<i>ABOUT THE AUTHORS</i>	xv
<i>ACKNOWLEDGEMENTS</i>	xvii

CHAPTER 1 *INTRODUCTION TO DATA SCIENCE* 1

1.1	Why Data Science?	1
1.2	What is Data Science?	1
1.3	The Data Science Methodology	2
1.4	Data Science Tasks	5
1.4.1	Description	6
1.4.2	Estimation	6
1.4.3	Classification	6
1.4.4	Clustering	7
1.4.5	Prediction	7
1.4.6	Association	7
	Exercises	8

CHAPTER 2 *THE BASICS OF PYTHON AND R* 9

2.1	Downloading Python	9
2.2	Basics of Coding in Python	9
2.2.1	Using Comments in Python	9
2.2.2	Executing Commands in Python	10
2.2.3	Importing Packages in Python	11
2.2.4	Getting Data into Python	12
2.2.5	Saving Output in Python	13
2.2.6	Accessing Records and Variables in Python	14
2.2.7	Setting Up Graphics in Python	15
2.3	Downloading R and RStudio	17
2.4	Basics of Coding in R	19
2.4.1	Using Comments in R	19
2.4.2	Executing Commands in R	20
2.4.3	Importing Packages in R	20
2.4.4	Getting Data into R	21
2.4.5	Saving Output in R	23
2.4.6	Accessing Records and Variables in R	24
	References	26
	Exercises	26

CHAPTER 3 DATA PREPARATION 29

- 3.1 The Bank Marketing Data Set 29
- 3.2 The Problem Understanding Phase 29
 - 3.2.1 Clearly Enunciate the Project Objectives 29
 - 3.2.2 Translate These Objectives into a Data Science Problem 30
- 3.3 Data Preparation Phase 31
- 3.4 Adding an Index Field 31
 - 3.4.1 How to Add an Index Field Using Python 31
 - 3.4.2 How to Add an Index Field Using R 32
- 3.5 Changing Misleading Field Values 33
 - 3.5.1 How to Change Misleading Field Values Using Python 34
 - 3.5.2 How to Change Misleading Field Values Using R 34
- 3.6 Reexpression of Categorical Data as Numeric 36
 - 3.6.1 How to Reexpress Categorical Field Values Using Python 36
 - 3.6.2 How to Reexpress Categorical Field Values Using R 38
- 3.7 Standardizing the Numeric Fields 39
 - 3.7.1 How to Standardize Numeric Fields Using Python 40
 - 3.7.2 How to Standardize Numeric Fields Using R 40
- 3.8 Identifying Outliers 40
 - 3.8.1 How to Identify Outliers Using Python 41
 - 3.8.2 How to Identify Outliers Using R 42
 - References 43
 - Exercises 44

CHAPTER 4 EXPLORATORY DATA ANALYSIS 47

- 4.1 EDA Versus HT 47
- 4.2 Bar Graphs with Response Overlay 47
 - 4.2.1 How to Construct a Bar Graph with Overlay Using Python 49
 - 4.2.2 How to Construct a Bar Graph with Overlay Using R 50
- 4.3 Contingency Tables 51
 - 4.3.1 How to Construct Contingency Tables Using Python 52
 - 4.3.2 How to Construct Contingency Tables Using R 53
- 4.4 Histograms with Response Overlay 53
 - 4.4.1 How to Construct Histograms with Overlay Using Python 55
 - 4.4.2 How to Construct Histograms with Overlay Using R 58
- 4.5 Binning Based on Predictive Value 58
 - 4.5.1 How to Perform Binning Based on Predictive Value Using Python 59
 - 4.5.2 How to Perform Binning Based on Predictive Value Using R 62
 - References 63
 - Exercises 63

CHAPTER 5 PREPARING TO MODEL THE DATA 69

- 5.1 The Story So Far 69
- 5.2 Partitioning the Data 69
 - 5.2.1 How to Partition the Data in Python 70
 - 5.2.2 How to Partition the Data in R 71

- 5.3 Validating your Partition 72
- 5.4 Balancing the Training Data Set 73
 - 5.4.1 How to Balance the Training Data Set in Python 74
 - 5.4.2 How to Balance the Training Data Set in R 75
- 5.5 Establishing Baseline Model Performance 77
 - References 78
 - Exercises 78

CHAPTER 6 *DECISION TREES* 81

- 6.1 Introduction to Decision Trees 81
- 6.2 Classification and Regression Trees 83
 - 6.2.1 How to Build CART Decision Trees Using Python 84
 - 6.2.2 How to Build CART Decision Trees Using R 86
- 6.3 The C5.0 Algorithm for Building Decision Trees 88
 - 6.3.1 How to Build C5.0 Decision Trees Using Python 89
 - 6.3.2 How to Build C5.0 Decision Trees Using R 90
- 6.4 Random Forests 91
 - 6.4.1 How to Build Random Forests in Python 92
 - 6.4.2 How to Build Random Forests in R 92
 - References 93
 - Exercises 93

CHAPTER 7 *MODEL EVALUATION* 97

- 7.1 Introduction to Model Evaluation 97
- 7.2 Classification Evaluation Measures 97
- 7.3 Sensitivity and Specificity 99
- 7.4 Precision, Recall, and F_β Scores 99
- 7.5 Method for Model Evaluation 100
- 7.6 An Application of Model Evaluation 100
 - 7.6.1 How to Perform Model Evaluation Using R 103
- 7.7 Accounting for Unequal Error Costs 104
 - 7.7.1 Accounting for Unequal Error Costs Using R 105
- 7.8 Comparing Models with and without Unequal Error Costs 106
- 7.9 Data-Driven Error Costs 107
 - Exercises 109

CHAPTER 8 *NAÏVE BAYES CLASSIFICATION* 113

- 8.1 Introduction to Naïve Bayes 113
- 8.2 Bayes Theorem 113
- 8.3 Maximum a Posteriori Hypothesis 114
- 8.4 Class Conditional Independence 114
- 8.5 Application of Naïve Bayes Classification 115
 - 8.5.1 Naïve Bayes in Python 121
 - 8.5.2 Naïve Bayes in R 123
 - References 125
 - Exercises 126

CHAPTER 9 *NEURAL NETWORKS* 129

- 9.1 Introduction to Neural Networks 129
- 9.2 The Neural Network Structure 129
- 9.3 Connection Weights and the Combination Function 131
- 9.4 The Sigmoid Activation Function 133
- 9.5 Backpropagation 134
- 9.6 An Application of a Neural Network Model 134
- 9.7 Interpreting the Weights in a Neural Network Model 136
- 9.8 How to Use Neural Networks in R 137
 - References 138
 - Exercises 138

CHAPTER 10 *CLUSTERING* 141

- 10.1 What is Clustering? 141
- 10.2 Introduction to the *K*-Means Clustering Algorithm 142
- 10.3 An Application of *K*-Means Clustering 143
- 10.4 Cluster Validation 144
- 10.5 How to Perform *K*-Means Clustering Using Python 145
- 10.6 How to Perform *K*-Means Clustering Using R 147
 - Exercises 149

CHAPTER 11 *REGRESSION MODELING* 151

- 11.1 The Estimation Task 151
- 11.2 Descriptive Regression Modeling 151
- 11.3 An Application of Multiple Regression Modeling 152
- 11.4 How to Perform Multiple Regression Modeling Using Python 154
- 11.5 How to Perform Multiple Regression Modeling Using R 156
- 11.6 Model Evaluation for Estimation 157
 - 11.6.1 How to Perform Estimation Model Evaluation Using Python 159
 - 11.6.2 How to Perform Estimation Model Evaluation Using R 160
- 11.7 Stepwise Regression 161
 - 11.7.1 How to Perform Stepwise Regression Using R 162
- 11.8 Baseline Models for Regression 162
 - References 163
 - Exercises 164

CHAPTER 12 *DIMENSION REDUCTION* 167

- 12.1 The Need for Dimension Reduction 167
- 12.2 Multicollinearity 168
- 12.3 Identifying Multicollinearity Using Variance Inflation Factors 171
 - 12.3.1 How to Identify Multicollinearity Using Python 172
 - 12.3.2 How to Identify Multicollinearity in R 173
- 12.4 Principal Components Analysis 175
- 12.5 An Application of Principal Components Analysis 175
- 12.6 How Many Components Should We Extract? 176

12.6.1	The Eigenvalue Criterion	176
12.6.2	The Proportion of Variance Explained Criterion	177
12.7	Performing Pca with $K = 4$	178
12.8	Validation of the Principal Components	178
12.9	How to Perform Principal Components Analysis Using Python	179
12.10	How to Perform Principal Components Analysis Using R	181
12.11	When is Multicollinearity Not a Problem?	183
	References	184
	Exercises	184

CHAPTER 13 GENERALIZED LINEAR MODELS 187

13.1	An Overview of General Linear Models	187
13.2	Linear Regression as a General Linear Model	188
13.3	Logistic Regression as a General Linear Model	188
13.4	An Application of Logistic Regression Modeling	189
13.4.1	How to Perform Logistic Regression Using Python	190
13.4.2	How to Perform Logistic Regression Using R	191
13.5	Poisson Regression	192
13.6	An Application of Poisson Regression Modeling	192
13.6.1	How to Perform Poisson Regression Using Python	193
13.6.2	How to Perform Poisson Regression Using R	194
	Reference	195
	Exercises	195

CHAPTER 14 ASSOCIATION RULES 199

14.1	Introduction to Association Rules	199
14.2	A Simple Example of Association Rule Mining	200
14.3	Support, Confidence, and Lift	200
14.4	Mining Association Rules	202
14.4.1	How to Mine Association Rules Using R	203
14.5	Confirming Our Metrics	207
14.6	The Confidence Difference Criterion	208
14.6.1	How to Apply the Confidence Difference Criterion Using R	208
14.7	The Confidence Quotient Criterion	209
14.7.1	How to Apply the Confidence Quotient Criterion Using R	210
	References	211
	Exercises	211

APPENDIX DATA SUMMARIZATION AND VISUALIZATION 215

Part 1:	Summarization 1: Building Blocks of Data Analysis	215
Part 2:	Visualization: Graphs and Tables for Summarizing and Organizing Data	217
Part 3:	Summarization 2: Measures of Center, Variability, and Position	222
Part 4:	Summarization and Visualization of Bivariate Relationships	225

PREFACE

DATA SCIENCE USING PYTHON AND R

Why this Book is Needed

Reason 1. Data Science is Hot. Really hot. *Bloomberg* called data scientist “the hottest job in America.”¹ *Business Insider* called it “The best job in America right now.”² *Glassdoor.com* rated it the best job in the world in 2018 for the third year in a row.³ The *Harvard Business Review* called data scientist “The sexiest job in the 21st century.”⁴

Reason 2: Top Two Open-source Tools. Python and R are the top two open-source data science tools in the world.⁵ Analysts and coders from around the world work hard to build analytic packages that Python and R users can then apply, free of charge.

Data Science Using Python and R will awaken your expertise in this cutting-edge field using the most widespread open-source analytics tools in the world. In *Data Science Using Python and R*, you will find step-by-step hands-on solutions of real-world business problems, using state-of-the-art techniques. In short, *you will learn data science by doing data science.*

Written for Beginners and Non-Beginners Alike

Data Science Using Python and R is written for the general reader, with no previous analytics or programming experience. We know that the information-age economy is making many English majors and History majors retool to take advantage of the great demand for data scientists.⁶ This is why we provide the following materials to help those who are new to the field hit the ground running.

¹ <https://www.bloomberg.com/news/articles/2018-05-18/-sexiest-job-ignites-talent-wars-as-demand-for-data-geeks-soars>.

² <https://www.businessinsider.com/what-its-like-to-be-a-data-scientist-best-job-in-america-2017-9>.

³ <https://www.forbes.com/sites/louiscolombus/2018/01/29/data-scientist-is-the-best-job-in-america-according-glassdoors-2018-rankings/#dd3f65055357>.

⁴ <https://www.hbs.edu/faculty/Pages/item.aspx?num=43110>.

⁵ See, for example, <https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html>.

⁶ For example, in May 2017, IBM projected that yearly demand for “data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.”

Forbes, <https://www.forbes.com/sites/louiscolombus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6b6fde277e3b>

- An entire chapter dedicated to learning the basics of using Python and R, for beginners. Which platform to use. Which packages to download. Everything you need to get started.
- An appendix dedicated to filling in any holes you might have in your introductory data analysis knowledge, called *Data Summarization and Visualization*.
- Step-by-step instructions throughout. Every instruction for every action.
- Every chapter has Exercises, where you may check your understanding and progress.

Those with analytics or programming experience will enjoy having a one-stop-shop for learning how to do data science using both Python and R. Managers, CIOs, CEOs, and CFOs will enjoy being able to communicate better with their data analysts and database analysts. The emphasis in this book on accurately accounting for model costs will help everyone uncover the most profitable nuggets of knowledge from the data, while avoiding the potential pitfalls that may cost your company millions of dollars.

Data Science Using Python and R covers exciting new topics, such as the following:

- Random Forests,
- General Linear Models, and
- Data-driven error costs to enhance profitability.

Data sets and supplemental materials can be found under the Related Resources section at <https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1119526817&bcsId=11765> and <https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1119526817&bcsId=11712>.

Data Science Using Python and R as a Textbook

Data Science Using Python and R naturally fits the role of textbook for a one-semester course or two-semester sequence of courses in introductory and intermediate data science. Faculty instructors will appreciate the *exercises at the end of every chapter*, totaling over 500 exercises in the book. There are three categories of exercises, from testing basic understanding toward more hands-on analysis of new and challenging applications.

- **Clarifying the Concepts.** These exercises test the students' basic understanding of the material, to make sure the students have absorbed what they have read.
- **Working with the Data.** These applied exercises ask the student to work in Python and R, following the step-by-step instructions that were presented in the chapter.
- **Hands-on Analysis.** Here is the real meat of the learning process for the students, where they apply their newly found knowledge and skills to uncover patterns and trends in new data sets. Here is where the students' expertise is challenged, in near real-world conditions. More than half of the exercises in the book consist of *Hands-on Analysis*.

The following supporting materials are also available to faculty adopters of the book at no cost.

- **Full solutions manual**, providing not just the answers, but how to arrive at the answers.
- **Powerpoint presentations of each chapter**, so that you may help the students understand the material, rather than just assigning them to read it.

To obtain access to these materials, contact your local Wiley representation and ask them to email the authors confirming that you have adopted the book for your course.

Data Science Using Python and R is appropriate for advanced undergraduate or graduate-level courses. No previous statistics, computer programming, or database expertise is required. What is required is a desire to learn.

How the Book is Structured

Data Science Using Python and R is structured around the Data Science Methodology.

The Data Science Methodology is a phased, adaptive, iterative, approach to the analysis of data, within a scientific framework.

1. **Problem Understanding Phase.** First, clearly enunciate the project objectives. Then, translate these objectives into the formulation of a problem that can be solved using data science.
2. **Data Preparation Phase.** Data cleaning/preparation is probably the most labor-intensive phase of the entire data science process.
 - Covered in Chapter 3: *Data Preparation*.
3. **Exploratory Data Analysis Phase.** Gain insights into your data through graphical exploration.
 - Covered in Chapter 4: *Exploratory Data Analysis*.
4. **Setup Phase.** Establish baseline model performance. Partition the data. Balance the data, if needed.
 - Covered in Chapter 5: *Preparing to Model the Data*.
5. **Modeling Phase.** The core of the data science process. Apply state-of-the-art algorithms to uncover some seriously profitable relationships lying hidden in the data.
 - Covered in Chapters 6 and 8–14.
6. **Evaluation Phase.** Determine whether your models are any good. Select the best-performing model from a set of competing models.
 - Covered in Chapter 7: *Model Evaluation*.
7. **Deployment Phase.** Interface with management to adapt your models for real-world deployment.

ABOUT THE AUTHORS

Chantal D. Larose, PhD, and Daniel T. Larose, PhD, form a unique father–daughter pair of data scientists. This is their third book as coauthors. Previously, they wrote:

- *Data Mining and Predictive Analytics*, Second Edition, Wiley, 2015.
 - This 800-page tome would be a wonderful companion to this book, for those looking to dive deeper in to the field.
- *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edition, Wiley, 2014.

Chantal D. Larose completed her PhD in Statistics at the University of Connecticut in 2015, with dissertation *Model-Based Clustering of Incomplete Data*. As an Assistant Professor of Decision Science at SUNY, New Paltz, she helped develop the Bachelor of Science in Business Analytics. Now, as an Assistant Professor of Statistics and Data Science at Eastern Connecticut State University, she is helping to develop the Mathematical Science Department’s data science curriculum.

Daniel T. Larose completed his PhD in Statistics at the University of Connecticut in 1996, with dissertation *Bayesian Approaches to Meta-Analysis*. He is a Professor of Statistics and Data Science at Central Connecticut State University. In 2001, he developed the world’s first online Master of Science in Data Mining. This is the 12th textbook that he has authored or coauthored. He runs a small consulting business, <https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1119526817&bcsId=11765> and <https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1119526817&bcsId=11712>. He also directs the online Master of Data Science program at CCSU.

ACKNOWLEDGMENTS

CHANTAL'S ACKNOWLEDGMENTS

Deepest thanks to my father Daniel, for his corny quips when proofreading. His guidance and passion for the craft reflects and enhances my own, and makes working with him a joy. Many thanks to my little sister Ravel, for her boundless love and incredible musical and scientific gifts. My fellow-traveler, she is an inspiration. Thanks to my brother Tristan, for all his hard work in school and letting me beat him at Mario Kart exactly once. Thanks to my mother Debra, for food and hugs. Also, coffee. Many, many thanks to coffee.

CHANTAL D. LAROSE, PH. D.

*Assistant Professor of Statistics & Data Science
Eastern Connecticut State University*

DANIEL'S ACKNOWLEDGMENTS

It is all about family. I would like to thank my daughter Chantal, for her insightful mind, her gentle presence, and for the joy she brings to every day. Thanks to my daughter Ravel, for her uniqueness, and for having the courage to follow her dream and become a chemist. Thanks to my son Tristan, for his math and computer skills, and for his help moving rocks in the backyard. I would also like to acknowledge my stillborn daughter Ellyriane Soleil. How we miss what you would have become. Finally, thanks to my loving wife, Debra, for her deep love and care for all of us, all these years. I love you all very much.

DANIEL T. LAROSE, PH. D.

*Professor of Statistics and Data Science
Central Connecticut State University
www.ccsu.edu/faculty/larose*

INTRODUCTION TO DATA SCIENCE

1.1 WHY DATA SCIENCE?

Data science is one of the fastest growing fields in the world, with 6.5 times as many job openings in 2017 as compared to 2012.¹ Demand for data scientists is expected to increase in the future. For example, in May 2017, IBM projected that yearly demand for “data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.”² <http://InfoWorld.com> reported that the #1 “reason why data scientist remains the top job in America”³ is that “there is a shortage of talent.” That is why we wrote this book, to help alleviate the shortage of qualified data scientists.

1.2 WHAT IS DATA SCIENCE?

Simply put, *data science* is the systematic analysis of data within a scientific framework. That is, data science is the

- adaptive, iterative, and phased approach to the analysis of data,
- performed within a systematic framework,
- that uncovers optimal models,
- by assessing and accounting for the true costs of prediction errors.

¹Forbes, <https://www.forbes.com/sites/louiscolombus/2017/12/11/linkedin-fastest-growing-jobs-today-are-in-data-science-machine-learning/#5b3100f051bd>

²Forbes, <https://www.forbes.com/sites/louiscolombus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6b6fde277e3b>

³<http://Infoworld.com>, <https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html>

Data science combines the

- data-driven approach of statistical data analysis,
- the computational power and programming acumen of computer science, and
- domain-specific business intelligence,

in order to uncover actionable and profitable nuggets of information from large databases.

In other words, data science allows us to extract actionable knowledge from under-utilized databases. Thus, data warehouses that have been gathering dust can now be leveraged to uncover hidden profit and enhance the bottom line. Data science lets people leverage large amounts of data and computing power to tackle complex questions. Patterns can arise out of data which could not have been uncovered otherwise. These discoveries can lead to powerful results, such as more effective treatment of medical patients or more profits for a company.

1.3 THE DATA SCIENCE METHODOLOGY

We follow the *Data Science Methodology* (DSM),⁴ which helps the analyst keep track of which phase of the analysis he or she is performing. Figure 1.1 illustrates the adaptive and iterative nature of the DSM, using the following phases:

- 1. Problem Understanding Phase.** How often have teams worked hard to solve a problem, only to find out later that they solved the wrong problem? Further, how often have the marketing team and the analytics team not been on the same page? This phase attempts to avoid these pitfalls.
 - a. First, clearly enunciate the project objectives,
 - b. Then, translate these objectives into the formulation of a problem that can be solved using data science.
- 2. Data Preparation Phase.** Raw data from data repositories is seldom ready for the algorithms straight out of the box. Instead, it needs to be cleaned or “prepared for analysis.” When analysts first examine the data, they uncover the inevitable problems with data quality that always seem to occur. It is in this phase that we fix these problems. Data cleaning/preparation is probably the most labor-intensive phase of the entire data science process. The following is a non-exhaustive list of the issues that await the data preparer.
 - a. Identifying outliers and determining what to do about them.
 - b. Transforming and standardizing the data.
 - c. Reclassifying categorical variables.
 - d. Binning numerical variables.
 - e. Adding an index field.

⁴Adapted from the Cross-Industry Standard Practice for Data Mining (CRISP-DM). See, for example, *Data Mining and Predictive Analytics*, by Daniel T. Larose and Chantal D. Larose, John Wiley and Sons, Inc, 2015.

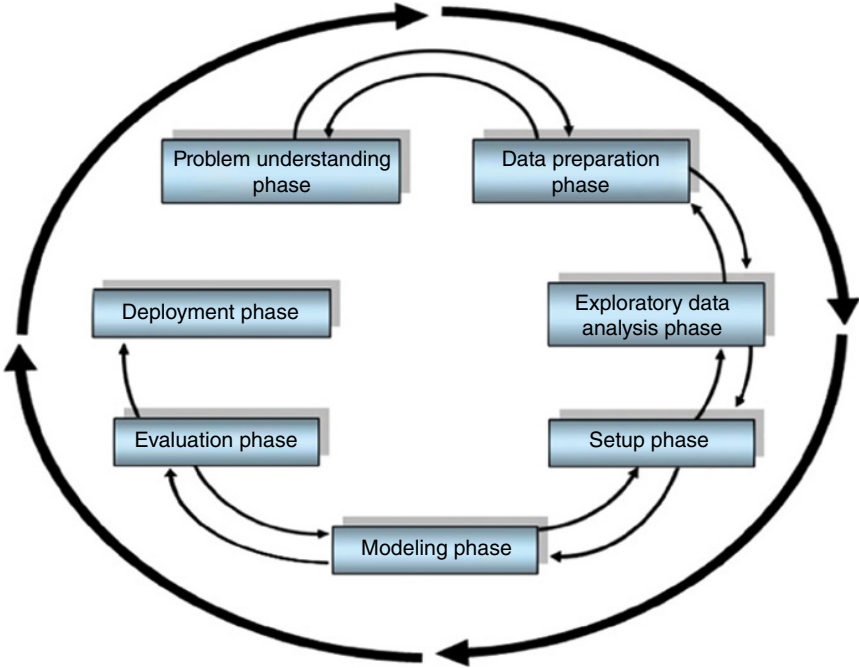


Figure 1.1 Data science methodology: the seven phases.

The data preparation phase is covered in Chapter 3.

3. Exploratory Data Analysis Phase. Now that your data are nice and clean, we can begin to explore the data, and learn some basic information. Graphical exploration is the focus here. Now is not the time for complex algorithms. Rather, we use simple exploratory methods to help us gain some preliminary insights. You might find that you can learn quite a bit just by using these simple methods. Here are some of the ways we can do this.

- a. Exploring the univariate relationships between predictors and the target variable.
- b. Exploring multivariate relationships among the variables.
- c. Binning based on predictive value to enhance our models.
- d. Deriving new variables based on a combination of existing variables.

We cover the exploratory data analysis phase in Chapter 4.

4. Setup Phase. At this point we are nearly ready to begin modeling the data. We just need to take care of a few important chores first, such as the following:

- a. Cross-validation, either twofold or n -fold. This is necessary to avoid data dredging. In addition, your data partitions need to be evaluated to ensure that they are indeed random.
- b. Balancing the data. This enhances the ability of certain algorithms to uncover relationships in the data.

- c. Establishing baseline performance. Suppose we told you we had a model that could predict correctly whether a credit card transaction was fraudulent or not 99% of the time. Impressed? You should not be. The non-fraudulent transaction rate is 99.932%.⁵ So, our model could simply predict that *every* transaction was non-fraudulent and be correct 99.932% of the time. This illustrates the importance of establishing baseline performance for your models, so that we can calibrate our models and determine whether they are any good.

The Setup Phase is covered in Chapter 5.

5. Modeling Phase. The modeling phase represents the opportunity to apply state-of-the-art algorithms to uncover some seriously profitable relationships lying hidden in the data. The modeling phase is the heart of your data scientific investigation and includes the following:

- a. Selecting and implementing the appropriate modeling algorithms. Applying inappropriate techniques will lead to inaccurate results that could cost your company big bucks.
- b. Making sure that our models outperform the baseline models.
- c. Fine-tuning your model algorithms to optimize the results. Should our decision tree be wide or deep? Should our neural network have one hidden layer or two? What should be our cutoff point to maximize profits? Analysts will need to spend some time fine-tuning their models before arriving at the optimal solution.

The modeling phase represents the core of your data science endeavor and is covered in Chapters 6 and 8–14.

6. Evaluation Phase. Your buddy at work may think he has a lock on his prediction for the Super Bowl. But is his prediction any good? That is the question. Anyone can make predictions. It is how the predictions perform against real data that is the real test. In the evaluation phase, we assess how our models are doing, whether they are making any money, or whether we need to go back and try to improve our prediction models.

- a. Your models need to be evaluated against the baseline performance measures from the Setup Phase. Are we beating the monkeys-with-darts model? If not, better try again.
- b. You need to determine whether your models are actually solving the problem at hand. Are your models actually achieving the objectives set for it back in the Problem Understanding Phase? Has some important aspect of the problem not been sufficiently accounted for?

⁵The Alaric Fraud Report, 2015, https://www.paymentscardsandmobile.com/wp-content/uploads/2015/03/PCM_Alaric_Fraud-Report_2015.pdf

- c. Apply error costs intrinsic to the data, because data-driven cost evaluation is the best way to model the actual costs involved. For instance, in a marketing campaign, a false positive is not as costly as a false negative. However, for a mortgage lender, a false positive is much more costly.
- d. You should tabulate a suite of models and determine which model performs the best. Choose either a single best model, or a small number of models, to move forward to the Deployment Phase.

The Evaluation Phase is covered in Chapter 7.

- 7. **Deployment Phase.** Finally, your models are ready for prime time! Report to management on your best models and work with management to adapt your models for real-world deployment.
 - a. Writing a report of your results may be considered a simple example of deployment. In your report, concentrate on the results of interest to management. Show that you solved the problem and report on the estimated profit, if applicable.
 - b. Stay involved with the project! Participate in the meetings and processes involved in model deployment, so that they stay focused on the problem at hand.

It should be emphasized that the DSM is iterative and adaptive. By *adaptive*, we mean that sometimes it is necessary to return to a previous phase for further work, based on some knowledge gained in the current phase. This is why there are arrows pointing both ways between most of the phases. For example, in the Evaluation Phase, we may find that the model we crafted does not actually address the original problem at hand, and that we need to return to the Modeling Phase to develop a model that will do so.

Also, the DSM is *iterative*, in that sometimes we may use our experience of building an effective model on a similar problem. That is, the model we created serves as an input to the investigation of a related problem. This is why the outer ring of arrows in Figure 1.1 shows a constant recycling of older models used as inputs to examining new solutions to new problems.

1.4 DATA SCIENCE TASKS

The most common data science tasks are the following:

- Description
- Estimation
- Classification
- Clustering
- Prediction
- Association

Next, we describe what each of these tasks represent and in which chapters these tasks are covered.

1.4.1 Description

Data scientists are often called upon to *describe* patterns and trends lying within the data. For example, a data scientist may describe a cluster of customers most likely to leave our company's service as those with high-usage minutes and a high number of customer service calls. After describing this cluster, the data scientist may explain that the high number of customer service calls indicates perhaps that the customer is unhappy. Working with the marketing team, the analyst can then suggest possible interventions to explore to retain such customers.

The description task is in widespread use around the world by specialists and nonspecialists alike. For example, when a sports announcer states that a baseball player has a lifetime batting average (hits/at-bats) of 0.350, he or she is describing this player's lifetime batting performance. This is an example of *descriptive statistics*,⁶ further examples of which may be found in the Appendix: Data Summarization and Visualization. Nearly every chapter in the book contains examples of the description task, from the graphical EDA methods of Chapter 4, to the descriptions of data clusters in Chapter 10, to the bivariate relationships in Chapter 11.

1.4.2 Estimation

Estimation refers to the approximation of the value of a numeric target variable using a collection of predictor variables. Estimation models are built using records where the target values are known, so that the models can learn which target values are associated with which predictor values. Then, the estimation models can estimate the target values for new data, for which the target value is unknown. For example, the analyst can estimate the mortgage amount a potential customer can afford, based on a set of personal and demographic factors. This estimate is based on a model built by looking at past models of how much previous customers could afford. Estimation requires that the target variable be numeric. Estimation methods are covered in Chapters 9, 11, and 13.

1.4.3 Classification

Classification is similar to estimation, except that the target variable is categorical rather than continuous. Classification represents perhaps the most widespread task in data science, and the most profitable. For instance, a mortgage lender would be interested in determining which of their customers is likely to default on their

⁶For example, see *Discovering Statistics*, by Daniel T. Larose, W.H. Freeman, 2016.