

Confident Data Skills

Second Edition

Confident Data Skills

How to work with data
and futureproof your career

Kirill Eremenko



Publisher's note

Every possible effort has been made to ensure that the information contained in this book is accurate at the time of going to press, and the publishers and authors cannot accept responsibility for any errors or omissions, however caused. No responsibility for loss or damage occasioned to any person acting, or refraining from action, as a result of the material in this publication can be accepted by the publisher or the author.

First published in Great Britain and the United States in 2018 by Kogan Page Limited

Second edition 2020

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licences issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned addresses:

2nd Floor, 45 Gee Street
London
EC1V 3RS
United Kingdom

122 W 27th Street
New York, NY 10001
USA

4737/23 Ansari Road
Daryaganj
New Delhi 110002
India

Kogan Page books are printed on paper from sustainable forests.

© Kirill Eremenko 2018, 2020

The right of Kirill Eremenko to be identified as the author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

ISBNs

Hardback 978 1 78966 441 6

Paperback 978 1 78966 438 6

Ebook 978 1 78966 439 3

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library.

Library of Congress Control Number

2020941850

Typeset by Integra Software Services, Pondicherry

Print production managed by Jellyfish

Printed and bound in Great Britain by CPI Group (UK) Ltd, Croydon CR0 4YY

*To my parents, Alexander and Elena Eremenko,
who taught me the most important thing in life –
how to be a good person*

CONTENTS

List of figures xi

Bonus for readers xiv

Acknowledgments xv

Introduction 1

PART ONE ‘What is it?’ Key principles 3

Striding out 4

The future is data 5

Arresting developments 6

01 **Defining data** 7

Data is everywhere 8

Big (data) is beautiful 10

Storing and processing data 12

Data can generate content 14

Using data 17

Why data matters now 18

Worrying achieves nothing 20

References 24

Notes 25

02 **How data fulfils our needs** 26

The permeation of data 26

Data science and physiology 27

Data science and safety 30

Data science and belonging 32

Data science and esteem 38
Data science and self-actualization 38
Some final thoughts 40
References 40
Notes 42

03 AI and our future 43

What is artificial intelligence? 43
Artificial general intelligence 44
Narrow AI 45
Robotic process automation 46
Computer vision 47
Natural language processing 49
Reinforcement learning and deep learning 52
The dark side of AI 53
Prepare for Part Two 63
References 64
Notes 66

PART TWO ‘When and where can I get it?’ Data gathering and analysis 69

The Data Science Process 70
Getting started 74

04 Identify the question 77

Look, ma, no data! 78
How do you solve a problem like... 79
Timekeeping 94
The art of saying no 95
Onward! 96
References 97
Notes 97

- 05 Data preparation 98**
Encouraging data to talk 98
With great power comes great responsibility 99
Preparing your data for a journey 102
Reference 120
Notes 120
- 06 Data analysis: Part 1 122**
Don't skip this step 123
Classification and clustering 124
Classification 125
Decision trees 126
Random forest 129
K-nearest neighbours (K-NN) 133
Naive Bayes 137
Classification with Naive Bayes 144
Logistic regression 151
Clustering 160
K-means clustering 160
Hierarchical clustering 170
References 176
Notes 176
- 07 Data analysis: Part 2 178**
Reinforcement learning 178
The multi-armed bandit problem 180
Upper confidence bound 186
Thompson sampling 194
Upper confidence bound vs Thompson sampling:
Which is preferable? 202
Deep learning 204
Weight training: How neural networks learn 217
The future for data analysis 220
References 220
Notes 221

PART THREE 'How can I present it?' Communicating data 223

Looking good 224
You're not finished yet! 224
The career shaper 225

08 Data visualization 227

What is visual analytics? 227
What is data visualization? 233
Speaking a visual language 235
Steps to creating compelling visuals 237
Final thoughts 248
References 248
Notes 248
• *Looking further: Types of chart* 249

09 Data presentation 261

The importance of storytelling 261
Creating data advocacy 263
How to create a killer presentation 265
The end of the process 276
References 276
Notes 277

10 Your career in data science 278

Breaking in 279
Applying for jobs 290
Preparing for an interview 291
Interviewing them 293
Nurturing your career within a company 294
References 296
Notes 297

Index 299

LIST OF FIGURES

- 2.1** Maslow's hierarchy of needs 27
- II.1** The Data Science Process 70
- 5.1** Importing spreadsheets containing missing data 109
- 5.2** Missing data in a spreadsheet to show start-up growth 114
- 6.1** Scatter plot to show customer data organized by age and time spent gaming 127
- 6.2** Scatter plot to show customer data, divided into leaves 128
- 6.3** Flowchart to show the decision tree process 129
- 6.4** Measuring Euclidean distance 135
- 6.5** The K-NN algorithm 136
- 6.6** Success of wine harvests according to hours of sunlight and rainfall 1 145
- 6.7** Success of wine harvests according to hours of sunlight and rainfall 2 146
- 6.8** Calculating marginal likelihood for Naive Bayes 147
- 6.9** Calculating likelihood for Naive Bayes 148
- 6.10** Scatter diagram to show respondents' salaries based on experience 152
- 6.11** Scatter diagram to show whether or not respondents opened our email, based on their age 153
- 6.12** Straight linear regression to show the likelihood of respondents opening our email, based on their age 154
- 6.13** Adapted linear regression to show the likelihood of respondents opening our email, based on their age 154
- 6.14** Linear regression transposed to a logistic regression function 155
- 6.15** Graph containing categorical variables 156
- 6.16** Logistic regression line 156
- 6.17** Logistic regression line (without observations) 157

- 6.18** Logistic regression including fitted values 158
- 6.19** Logistic regression line (segmented) 159
- 6.20** Dataset for e-commerce company customers 162
- 6.21** Scatter plot for two independent variables 163
- 6.22** K-means assigning random centroids 163
- 6.23** Assigning data points to closest centroid 164
- 6.24** Re-computing centroids 164
- 6.25** Reassigning data points to closest centroid 165
- 6.26** Centroids are displaced 165
- 6.27** Repeat re-computation of centroids 166
- 6.28** K-means, visualized 166
- 6.29** Three clusters, identified by K-means 167
- 6.30** Finding the distance between data points in one cluster 168
- 6.31** Finding the distance between data points in two clusters 169
- 6.32** WCSS values against number of clusters 170
- 6.33** Steps to agglomerative hierarchical clustering 172
- 6.34** The process of agglomerative clustering, as shown in a dendrogram 173
- 6.35** Splitting into four clusters 174
- 6.36** Identifying the largest vertical segment 175
- 6.37** Splitting into two clusters 175
- 7.1** Multi-armed bandit distributions 183
- 7.2** Multi-armed bandit distributions with expected returns 184
- 7.3** Expected returns vertical axis 187
- 7.4** Starting line 188
- 7.5** Starting confidence bound 189
- 7.6** Real-life scenario 189
- 7.7** Machine D3 after trial rounds 191
- 7.8** All machines after trial rounds 192
- 7.9** Shifting and narrowing of the D4 confidence bound 193
- 7.10** Thompson sampling multi-armed bandit distributions 195

- 7.11** Thompson sampling expected returns 195
- 7.12** Distribution for machine M3 after trial rounds 196
- 7.13** Distribution for machine M3 with true expected returns 197
- 7.14** Distributions for all three machines after trial runs 197
- 7.15** Randomly generated bandit configuration 198
- 7.16** Machine M2 updated distribution 200
- 7.17** Thompson sampling refined distribution curves 201
- 7.18** Parts of a neuron 205
- 7.19** A neuron's input and output values 205
- 7.20** A threshold activation function 208
- 7.21** A rectified linear unit 209
- 7.22** A sigmoid activation function 210
- 7.23** A perceptron to estimate housing value 211
- 7.24** A complex neural network to estimate housing value 212
- 7.25** Impact of features on N1 213
- 7.26** Impact of features on N2 213
- 7.27** Impact of features on N3 214
- 7.28** Calculating output with a neural network 215
- 7.29** A neural network with three hidden layers 217
- 8.1** Napoleon's retreat from Moscow (the Russian Campaign 1812–13) 246

BONUS FOR READERS

Thank you for picking up this book. You've made a huge step in your journey into data science.

All readers gain complimentary access to my Data Science A–Z course. Just go to www.superdatascience.com/bookbonus and use the password `datarockstar`.

You can download a guide to using colour in visualizations at www.superdatascience.com/cds.

Happy analysing!

ACKNOWLEDGEMENTS

I would like to thank my father, Alexander Eremenko, whose love and care have shaped me into the person I am today, and who has shown me through his firm guidance how I can take hold of life's opportunities. Thank you to my warm-hearted mother, Elena Eremenko, for always lending an ear to my crazy ideas and for encouraging my brothers and me to take part in the wider world – through music, language, dance and so much more. Were it not for her wise counsel, I would never have immigrated to Australia.

Thanks to my brother, Mark Eremenko, for always believing in me, and for his unshakeable confidence. His fearlessness in life continues to fuel so many of the important decisions I make. Thank you to my brother, Ilya Eremenko, wise beyond his years, for his impressive business ideas and well-considered ventures. I am certain that fame and fortune will soon be knocking on his door.

Thank you to my grandmother Valentina, aunt Natasha and cousin Yura, for their endless love and care. And thank you to the entire Tanakovic and Svoren families, including my brothers Adam and David for all the dearest moments we share.

Thank you to my students and to the thousands of people who listen to the SuperDataScience podcast. My audience inspires me to keep going!

Several key people helped to make this book a reality. I would like to thank my writing partner Zara Karschay for capturing my voice. Thank you to my commissioning editor Rebecca Bush and production editor Stefan Leszczuk, along with Anna Moss, whose feedback and guidance were fundamental to the writing process, and to Kogan Page's editorial team for their rigour. Thank you to my friend and business partner Hadelin de Ponteves for inspiring me and for being a great source of support in tackling some of the most challenging subject matters in the field of data science, as well as for his help in reviewing the technical aspects of this book.

Thank you to my friend and executive assistant Mitja Bosnič for his tireless efforts in making this second edition possible. My thanks also go to the talented team at SuperDataScience for taking on additional responsibilities so that I could write this book. Thank you to the hardworking team at Udemy, including my supportive account managers Lana Martinez and Erin Adams.

Thank you to my friend and mentor Artem Vladimirov, whose admirable work ethic and knowledge lay the foundations upon which I have built everything I know about data science. Many thanks to Vitaly Dolgov, Ivor Lok, Richard Hopkins, Tracy Crossley and Herb Kanis for being excellent role models, for believing in me, for always being there when I needed help, and for guiding me through times both good and bad. Thank you to Katherina Andryskova – I am grateful that she was the first person to read and offer invaluable feedback on *Confident Data Skills*.

I give my express thanks to the people whose work contributed to the case studies in this book: Alberto Cairo, Samuel Hinton, Richard Hopkins, Kristen Kehrer, Raul Popa, Caroline McColl, Ulf Morys, Daniel and Leigh Pullen, Dominic Roe, Adrian Rosebrock, Matthew Rosenquist, Dan Shiebler, Ben Taylor, Artem Vladimirov, and Stephen Welch.

The motivational teachers at Moscow School 54, the Moscow Institute of Physics and Technology and the University of Queensland have my thanks for giving me such beneficial education. To all my past colleagues at Deloitte and Sunsuper, thank you for giving me the professional development I needed for creating my data science toolkit.

Most of all, I want to thank you, the reader, for giving me your valuable time. It has been my foremost hope that this book will encourage those who wish to understand and implement data science in their careers.