

The Art of Feature Engineering

When working with a data set, machine learning engineers might train a model but find that the results are not as good as they need. To get better results, they can try to improve the model or collect more data, but there is another avenue: feature engineering. The feature engineering process can help improve results by modifying the data's features to better capture the nature of the problem. This process is partly an art and partly a palette of tricks and recipes. This practical guide to feature engineering is an essential addition to any data scientist's or machine learning engineer's toolbox, providing new ideas on how to improve the performance of a machine learning solution. Beginning with the basic concepts and techniques of feature engineering, the text builds up to a unique cross-domain approach that spans data on graphs, texts, time series and images, with fully worked-out case studies. Key topics include binning, out-of-fold estimation, feature selection, dimensionality reduction and encoding variable-length data. The full source code for the case studies is available on a companion website as Python Jupyter notebooks.

PABLO DUBOUE is the director of Textualization Software Ltd. and is passionate about improving society through technology. He has a PhD in computer science from Columbia University and was part of the IBM Watson team that beat the "Jeopardy!" champions in 2011. He splits his time between teaching machine learning, doing open research, contributing to free software projects, and consulting for start-ups. He has taught in three different countries and done joint research with more than 50 coauthors. Recent career highlights include a best paper award in the Canadian AI Conference industrial track and consulting for a start-up acquired by Intel Corp.

Cambridge University Press
978-1-108-70938-5 — The Art of Feature Engineering
Pablo Duboue
Frontmatter
[More Information](#)

The Art of Feature Engineering
Essentials for Machine Learning

PABLO DUBOUE
Textualization Software Ltd



CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108709385

DOI: 10.1017/9781108671682

© Pablo Duboue 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2020

Printed in the United Kingdom by TJ International, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Duboue, Pablo, 1976– author.

Title: The art of feature engineering : essentials for machine learning / Pablo Duboue, Textualization Software Ltd.

Description: First edition. | Cambridge ; New York, NY : Cambridge University Press, 2020. |

Includes bibliographical references and index.

Identifiers: LCCN 2019060140 (print) | LCCN 2019060141 (ebook) | ISBN 9781108709385 (paperback) | ISBN 9781108671682 (epub)

Subjects: LCSH: Machine learning. | Python (Computer program language)

Classification: LCC Q325.5 .D83 2020 (print) | LCC Q325.5 (ebook) | DDC 006.3/1–dc23

LC record available at <https://lcn.loc.gov/2019060140>

LC ebook record available at <https://lcn.loc.gov/2019060141>

ISBN 978-1-108-70938-5 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

A la Universidad Nacional de Córdoba, que me formó en virtud y letras.[†]

[†] Dedicated to National University of Córdoba, which taught me character and knowledge.

Cambridge University Press
978-1-108-70938-5 — The Art of Feature Engineering
Pablo Duboue
Frontmatter
[More Information](#)

Contents

<i>Preface</i>	<i>page xi</i>
PART ONE FUNDAMENTALS	1
1 Introduction	3
1.1 Feature Engineering	6
1.2 Evaluation	10
1.3 Cycles	15
1.4 Analysis	20
1.5 Other Processes	25
1.6 Discussion	30
1.7 Learning More	32
2 Features, Combined: Normalization, Discretization and Outliers	34
2.1 Normalizing Features	35
2.2 Discretization and Binning	43
2.3 Descriptive Features	50
2.4 Dealing with Outliers	54
2.5 Advanced Techniques	56
2.6 Learning More	57
3 Features, Expanded: Computable Features, Imputation and Kernels	59
3.1 Computable Features	60
3.2 Imputation	67
3.3 Decomposing Complex Features	70
3.4 Kernel-Induced Feature Expansion	73
3.5 Learning More	78

4	Features, Reduced: Feature Selection, Dimensionality Reduction and Embeddings	79
4.1	Feature Selection	80
4.2	Regularization and Embedded Feature Selection	94
4.3	Dimensionality Reduction	99
4.4	Learning More	111
5	Advanced Topics: Variable-Length Data and Automated Feature Engineering	112
5.1	Variable-Length Feature Vectors	112
5.2	Instance-Based Engineering	124
5.3	Deep Learning and Feature Engineering	127
5.4	Automated Feature Engineering	130
5.5	Learning More	135
	PART TWO CASE STUDIES	137
6	Graph Data	139
6.1	WikiCities Dataset	142
6.2	Exploratory Data Analysis (EDA)	144
6.3	First Feature Set	150
6.4	Second Feature Set	158
6.5	Final Feature Sets	160
6.6	Learning More	162
7	Timestamped Data	163
7.1	WikiCities: Historical Features	166
7.2	Time Lagged Features	169
7.3	Sliding Windows	172
7.4	Third Featurization: EMA	173
7.5	Historical Data as Data Expansion	174
7.6	Time Series	176
7.7	Learning More	183
8	Textual Data	186
8.1	WikiCities: Text	189
8.2	Exploratory Data Analysis	190
8.3	Numeric Tokens Only	194
8.4	Bag-of-Words	196
8.5	Stop Words and Morphological Features	200
8.6	Features in Context	203
8.7	Skip Bigrams and Feature Hashing	204

8.8	Dimensionality Reduction and Embeddings	205
8.9	Closing Remarks	208
8.10	Learning More	211
9	Image Data	212
9.1	WikiCities: Satellite Images	215
9.2	Exploratory Data Analysis	216
9.3	Pixels as Features	217
9.4	Automatic Dataset Expansion	222
9.5	Descriptive Features: Histograms	223
9.6	Local Feature Detectors: Corners	225
9.7	Dimensionality Reduction: HOGs	227
9.8	Closing Remarks	228
9.9	Learning More	231
10	Other Domains: Video, GIS and Preferences	233
10.1	Video	234
10.2	Geographical Features	239
10.3	Preferences	242
	<i>Bibliography</i>	246
	<i>Index</i>	270

Cambridge University Press
978-1-108-70938-5 — The Art of Feature Engineering
Pablo Duboue
Frontmatter
[More Information](#)

Preface

There are as many reasons to write books as there are books written. In the case of this book, it was driven by a desire to both help practitioners and structure information scattered in a variety of formats. The end result is yours to judge. Preliminary feedback indicates it errs on the side of too much material rather than covering only material for which there are clear intuitions or consensus. It seems that, as a matter of personal taste, I would rather have more tools in a toolbox than a few general tools. The case studies are a different matter. Some reviewers have liked them quite a bit. Others have not seen the point of having them. All I can say is that there was a substantive effort behind putting them together. People have different learning styles. Some learn by seeing others do things. If you are in that group, here you have many end-to-end examples of somebody doing feature engineering. Hopefully, it will help the concepts click.

If you think you would have written about the topic in a different way; feature engineering needs more detailed treatment beyond the anecdotal. If you leave these pages with the intention of writing your own material, I will consider the effort of putting together this book to be a success.

My interest with feature engineering started while working together with David Gondek and the rest of the “Jeopardy!” team at IBM TJ Watson Research Center in the late 2000s. The process and ideas in this book draw heavily from that experience. The error analysis sessions chaired by David Ferrucci were a gruelling two days of looking at problem after problem and brainstorming ideas of how to address them. It was a very stressful time; hopefully, this book will help you profit from that experience without having to endure it. Even though it has been years since we worked together, this book exists thanks to their efforts that transcend the show itself.

After leaving IBM, during the years I have been consulting, I have seen countless professionals abandon promising paths due to lack of feature engineering tools. This book is for them.

My students from the 2014 course on Machine Learning over Large Datasets in the National University of Córdoba were also instrumental in the creation of this book, and the students of the 2018 graduate course in feature engineering tested an earlier version of the material. The fantastic data science community in Vancouver, particularly the one centred around the paper reading LearnDS meetup, also proved very helpful with comments, suggestions and as a source of reviewers.

This book has benefited from more than 30 full-book and book-chapter reviewers. In alphabetical order, I would like to extend my unreserved gratitude to Sampoorna Biswas, Eric Brochu, Rupert Brooks, Gavin Brown, Steven Butler, Pablo Gabriel Celayes, Claudio Conejero, Nelson Correa, Facundo Deza, Michel Galley, Lulu Huang, Max Kanter, Rahul Khopkar, Jessica Kuo, Newick Lee, Alice Liang, Pierre Louarn, Ives Macedo, Aanchan Mohan, Kenneth Odoh, Heri Rakotomalala, Andriy Redko, David Rowley, Ivan Savov, Jason Smith, Alex Strackhan, Adrián Tichno, Chen Xi, Annie Ying, and Wlodek Zadrozny. The anonymous reviewers provided by the publisher helped grow the manuscript in a direction that better fits with existing instructional material. I thank them for their help and apologize for the lack of explicit acknowledgement. Anonymity in reviews is the cornerstone of quality scientific publishing.

The publisher has also helped move this book from concept to finished product. Kaitlin Leach's help made all the difference while navigating the intricacies of book publishing, together with Amy He, who helped me stay on track.

A book always takes a deep toll on a family. This book is unusually intertwined with family, as my wife, Annie Ying, is also a data scientist.[†] This book started during a personal sabbatical while I was accompanying her in New York on a spousal visa. She proofread every chapter and helped keep my sanity during the dark times of writing. This book would not exist without her. It would have not even been started. Annie, muchas gracias, de corazón.

[†] You might say we are a couple trying to beat the odds.