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.

The Art of Feature Engineering

Essentials for Machine Learning

PABLO DUBOUE
Textualization Software Ltd

CAMBRIDGE
UNIVERSITY PRESS

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

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://lccn.loc.gov/2019060140
LC ebook record available at https://lccn.loc.gov/2019060141

ISBN 978-1-108-70938-5 Paperback

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.

Contents

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, Newvick 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. Annonymity 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.