Finding Ghosts in Your Data

Anomaly Detection Techniques with Examples in Python

Kevin Feasel

Apress®

Finding Ghosts in Your Data: Anomaly Detection Techniques with Examples in Python

Kevin Feasel
DURHAM, NC, USA

ISBN-13 (pbk): 978-1-4842-8869-6 ISBN-13 (electronic): 978-1-4842-8870-2
https://doi.org/10.1007/978-1-4842-8870-2

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Jonathan Gennick
Development Editor: Laura Berendson
Coordinating Editor: Jill Balzano

Cover photo by Pawel Czerwinski on Unsplash

Distributed to the book trade worldwide by Springer Science+Business Media LLC, 1 New York Plaza, Suite 4600, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

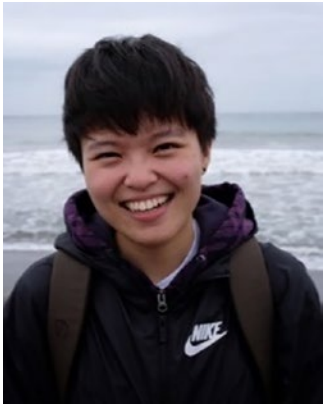
To Mom and Dad, who know a thing or four about anomalies.

Table of Contents

About the Author

Kevin Feasel is a Microsoft Data Platform MVP and CTO at Faregame Inc., where he specializes in data analytics with T-SQL and R, forcing Spark clusters to do his bidding, fighting with Kafka, and pulling rabbits out of hats on demand. He is the lead contributor to Curated SQL, president of the Triangle Area SQL Server Users Group, and author of *PolyBase Revealed*. A resident of Durham, North Carolina, he can be found cycling the trails along the triangle whenever the weather's nice enough.

About the Technical Reviewer



Yin-Ting (Ting) Chou is currently a Data Engineer/Full-Stack Data Scientist at ChannelAdvisor. She has been a key member on several large-scale data science projects, including demand forecasting, anomaly detection, and social network analysis. Even though she is keen on data analysis, which drove her to obtain her master's degree in statistics from the University of Minnesota, Twin Cities, she also believes that the other key to success in a machine learning project is to have an efficient and effective system to support the whole model productizing process. To create the system, she is currently diving into the fields of MLOps and containers. For more information about her, visit www.yintingchou.com.

Introduction

Welcome to this book on anomaly detection! Over the course of this book, we are going to build an anomaly detection engine in Python. In order to do that, we must first answer the question, "What is an anomaly?" Such a question has a simple answer, but in providing the simple answer, we open the door to more questions, whose answers open yet more doors. This is the joy and curse of the academic world: we can always go a little bit further down the rabbit hole.

Before we start diving into rabbit holes, however, let's level-set expectations. All of the code in this book will be in Python. This is certainly not the only language you can use for the purpose—my esteemed technical reviewer, another colleague, and I wrote an anomaly detection engine using a combination of C# and R, so nothing requires that we use Python. We do cover language and other design choices in the book, so I'll spare you the rest here. As far as your comfort level with Python goes, the purpose of this book is not to teach you the language, so I will assume some familiarity with the language. I do, of course, provide context to the code we will write and will spend extra time on concepts that are less intuitive. Furthermore, all of the code we will use in the book is available in an accompanying GitHub repository at `https://github.com/Apress/finding-ghosts-in-your-data`.

My goal in this book is not just to write an anomaly detection engine—it is to straddle the line between the academic and development worlds. There is a rich literature around anomaly detection, but much of the literature is dense and steeped with formal logic. I want to bring you some of the best insights from that academic literature but expose it in a way that makes sense for the large majority of developers. For this reason, each part in the book will have at least one chapter dedicated to theory. In addition, most of the code-writing chapters also start with the theory because it isn't enough simply to type out a few commands or check a project's readme for a sample method call; I want to help you understand why something is important, when an approach can work, and when the approach may fail. Furthermore, should you wish to take your own dive into the literature, the bibliography at the end of the book includes a variety of academic resources.

Before I sign off and we jump into the book, I want to give a special thank you to my colleague and technical editor, Ting Chou. I have the utmost respect for Ting's skills, so much so that I tried to get her to coauthor the book with me! She did a lot to keep me on the right path and heavily influenced the final shape of this book, including certain choices of algorithms and parts of the tech stack that we will use. That said, any errors are, of course, mine and mine alone. Unfortunately.

If you have thoughts on the book or on anomaly detection, I'd love to hear from you. The easiest way to reach out is via email: feasel@catallaxyservices.com. In the meantime, I hope you enjoy the book.