Walter R. Paczkowski

Business Analytics

Data Science for Business Problems

Business Analytics

Walter R. Paczkowski

Business Analytics

Data Science for Business Problems

Springer

Walter R. Paczkowski
Data Analytics Corp.
Plainsboro, NJ, USA

ISBN 978-3-030-87022-5 ISBN 978-3-030-87023-2 (eBook)
https://doi.org/10.1007/978-3-030-87023-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

I analyze business data—and I have been doing this for a long time. I was an analyst and department head, a consultant and trainer, worked on countless problems, written many books and reports, and delivered numerous presentations to all levels of management. I learned a lot. This book reflects insights I gained from this experience about *Business Data Analytics* that I want to share.

There are three questions you should quickly ask about this sharing. The first is obvious: "*Share what*?" The second logically follows: "*Share with whom*?" The third is more subtle: "*How does this book differ from other data analytic books*?" The first is about focus, the second is about target, and the third is about competitive comparison. So, let me address each question.

The Book's Focus

My experience has been with practical business problems. When I finished my academic training with a Ph.D. in economics and a heavy statistics exposure, I immediately started my professional career with an AT&T internal consulting group, *The Analytical Support Center* (*ASC*). I quickly learned that I needed both a theoretical, technical understanding of quantitative work—how to estimate a regression model, for example—and an understanding of how to deal with messy data beyond the nice, clean data sets I used as a graduate student. My time at the *ASC* was a great learning experience that I carried throughout my professional career at AT&T, including Bell Labs, and into my own consulting business. The lessons I learned were that good, solid data analysis for practical business problems requires:

1. A theoretical understanding of statistical, econometric, and (in the current era) machine learning methods
2. Data handling capabilities encompassing data organizing, preprocessing, and wrangling
3. Programming knowledge in at least one software language

These three components form a synergistic whole, a unifying approach if you wish, for doing business data analytics, and, in fact, any type of data analysis. This synergy implies that one part does not dominate any of the other two. They work together, feeding each other with the goal of solving only one overarching problem: how to provide decision makers with rich information extracted from data. Recognizing this problem was the most valuable lesson of all. All the analytical tools and know how must have a purpose and solving this problem is that purpose—there is no other.

I show this problem and the synergy of the three components for solving it as a triangle in Fig. 1. This triangle represents the almost philosophical approach I take for any form of business data analysis and is the one I advocate for all data analyses.

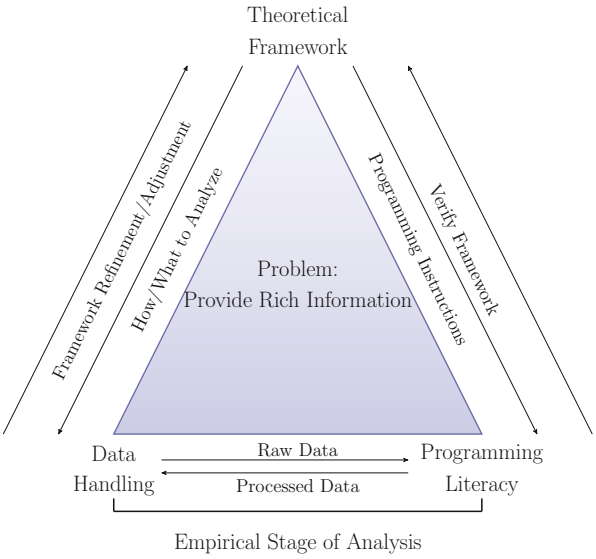

Fig. 1 The synergistic connection of the three components of effective data analysis for the overarching problem is illustrated in this triangular flow diagram. Every component is dependent on the others and none dominates the others. Regardless of the orientation of the triangle, the same relationships will hold

The overarching problem at the center of the triangle is not obvious. It is subtle. But because of its preeminence in the pantheon of problems any decision maker faces, I decided to allocate the entire first chapter to it. Spending so much space talking about information in a data analytics book may seem odd, but it is very important to understand why we do what we do, which is to analyze data to extract that rich information from data.

The theoretical understanding should be obvious. You need to know not just the methodologies but also their limitations so you can effectively apply them to solve a problem. The limitations may hinder you or just give you the wrong answers. Assume you were hired or commissioned by a business decision maker (e.g., a

CEO) to provide actionable, insightful, and useful rich information relevant for their problem. If the limitations of a methodology prevent you from accomplishing your charge, then your life as an analyst will be short-lived, to say the least. This will hold if you either do not know these limitations or simply choose to ignore them. Another methodological approach might be better, one that has fewer problems, or is just more applicable.

There is a dichotomy in methodology training. Most graduate-level statistics and econometric programs, and the newer Data Science programs, do an excellent job instructing students in the theory behind the methodologies. The focus of these academic programs is largely to train the next generation of academic professionals, not the next generation of business analytical professionals. Data Science programs, of which there are now many available online and "in person," often skim the surface of the theoretical underpinnings since their focus is to prepare the next generation of business analysts, those who will tackle the business decision makers' tough problems, and not the academic researchers. Something in between the academic and data science training is needed for successful business data analysts.

Data handling is not as obvious since it is infrequently taught and talked about in academic programs. In those programs, beginner students work with clean data with few problems and that are in nice, neat, and tidy data sets. They are frequently just given the data. More advanced students may be required to collect data, most often at the last phase of training for their thesis or dissertation, but these are small efforts, especially when compared to what they will have to deal with post training. The post-training work involves:

- Identifying the required data from diverse, disparate, and frequently disconnected data sources with possibly multiple definitions of the same quantitative concept
- Dealing with data dictionaries
- Dealing with samples of a very large database—how to draw the sample and determine the sample size
- Merging data from disparate sources
- Organizing data into a coherent framework appropriate for the statistical/econometric/machine learning methodology chosen
- Visualizing complex multivariate data to understand relationships, trends, patterns, and anomalies inside the data sets

This is all beyond what is provided by most training programs.

Finally, there is the programming. First, let me say that there is programming and then there is programming. The difference is scale and focus. Most people, when they hear about programming and programming languages, immediately think about large systems, especially ones needing a considerable amount of time (years?) to fully specify, develop, test, and deploy. They would be correct regarding large-scale, complex systems that handle a multitude of interconnected operations. Online ordering systems easily come to mind. Customer interfaces, inventory management, production coordination, supply chain management, price maintenance and dynamic pricing platforms, shipping and tracking, billing, and

collections are just a few components of these systems. The programming for these is complex to say the least.

As a business data analyst, you would not be involved in this type of programming although you might have to know about and access the subsystems of one or more of these larger systems. And major businesses are composed of many larger systems! You might have to write code to access the data, manipulate the retrieved data, and so forth, basically write programming code to do all the data handling I described above. And for this you need to know programming and languages.

There are many programming languages available. Only a few are needed for most business data analysis problems. In my experience, these are:

- *SQL*
- Python
- R

Julia should be included because it is growing in popularity due to its performance and ease of use. For this book, I will use Python because its ecosystem is strongly oriented toward machine learning with strong modeling, statistics, data visualization, and programming functionalities. In fact, its programming paradigm is clear to use, which is a definite advantage over other languages.

The Target Audience

The target audience for this book consists of business data analysts, data scientists, and market research professionals, or those aspiring to be any of these, in the private sector. You would be involved in or responsible for a myriad of quantitative analyses for business problems such as, but not limited to:

- Demand measurement and forecasting
- Predictive modeling
- Pricing analytics including elasticity estimation
- Customer satisfaction assessment
- Market and advertisement research
- New product development and research

To meet these tasks, you will have a need to know basic data analytical methods and some advanced methods, including data handling and management. This book will provide you with this needed background by:

- Explaining the intuition underlying analytic concepts
- Developing the mathematical and statistical analytic concepts
- Demonstrating analytical concepts using Python
- Illustrating analytical concepts with case studies

This book is also suitable for use in colleges and universities offering courses and certifications in business data analytics, data sciences, and market research. It could be used as a major or supplemental textbook.

Since the target audience consists of either current or aspiring business data analysts, it is assumed that you have or are developing a basic understanding of fundamental statistics at the "Stat 101" level: descriptive statistics, hypothesis testing, and regression analysis. Knowledge of econometric and market research principles, while not required, would be beneficial. In addition, a level of comfort with calculus and some matrix algebra is recommended, but not required. Appendices will provide you with some background as needed.

The Book's Competitive Comparison

There are many books on the market that discuss the three themes of this book: analytic methods, data handling, and programming languages. But they do them separately as opposed to a synergistic, analytic whole. They are given separate treatment so that you must cover a wide literature just to find what is needed for a specific business problem. Also, once found, you must translate the material into business terms. This book will present the three themes so you can more easily master what is needed for your work.

The Book's Structure

I divided this book into three parts. In Part I, I cover the basics of business data analytics including data handling, preprocessing, and visualization. In some instances, the basic analytic toolset is all you need to address problems raised by business executives. Part II is devoted to a richer set of analytic tools you should know at a minimum. These include regression modeling, time series analysis, and statistical table analysis. Part III extends the tools from Part II with more advanced methods: advanced regression modeling, classification methods, and grouping methods (*a.k.a., clustering*).

The three parts lead naturally from basic principles and methods to complex methods. I illustrate this logical order in Fig. 2.

Embedded in the three parts are case study examples of business problems using (albeit, fictitious, fake, or simulated) business transactions data designed to be indicative of what business data analysts use every day. Using simulated data

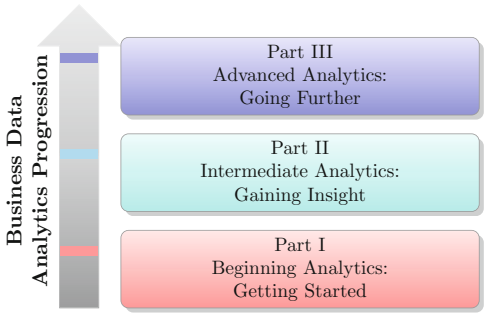

Fig. 2 This is a flow chart of the three parts of this book. The parts move progressively from basics to advanced. At the end of Part I, you should be able to do basic analyses of business data. At the end of Part II, you should be able to do regression and times series analysis. At the end of Part III, you should be able to do advanced machine learning work

for instructional purposes is certainly not without precedence. See, for example, Gelman et al. (2021). Data handling, visualization, and modeling are all illustrated using Python. All examples are in Jupyter notebooks available on Github.

Plainsboro, NJ, USA Walter R. Paczkowski

Acknowledgments

In my last book, I noted the support and encouragement I received from my wonderful wife, Gail, and my two daughters, Kristin and Melissa, and my son-in-law, David. As before, my wife Gail encouraged me to sit down and just write, especially when I did not want to, while my daughters provided the extra set of eyes I needed to make this book perfect. They provided the same support and encouragement for this book, so I owe them a lot, both then and now. I would also like to say something about my two grandsons who, now at 5 and 9, obviously did not contribute to this book but who, I hope, will look at this one in their adult years and say "My grandpa wrote this book, too."

Contents

List of Figures