Smart Data Discovery Using SAS® Viya®

Powerful Techniques for Deeper Insights

Felix Liao

The correct bibliographic citation for this manual is as follows: Liao, Felix. 2020. *Smart Data Discovery Using SAS® Viya®: Powerful Techniques for Deeper Insights*. Cary, NC: SAS Institute Inc.

Smart Data Discovery Using SAS® Viya®: Powerful Techniques for Deeper Insights

ISBN 978-1-64295-803-4 (Hardcover)
ISBN 978-1-63526-259-9 (Paperback)
ISBN 978-1-63526-726-6 (Web PDF)
ISBN 978-1-63526-724-2 (EPUB)
ISBN 978-1-63526-725-9 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

August 2020

Contents

Preface

Analytics is playing an increasingly strategic role in the ongoing digital transformation of organizations today. To succeed on your digital transformation journey, however, it is critical to enable analytics skills at all tiers of your organization and scale beyond the traditional data science team. It is through people with both strong business domain knowledge and analytics skills that you can often find the most valuable insights and make the biggest impact.

At SAS, we believe analytics can and should be for everyone and SAS Viya was built from the ground up to fulfill this vision of democratizing analytics. With a visual-based approach that supports the end-to-end analytics life cycle, SAS Viya supports the needs of traditional programmers as well as supporting a low-code and no-code approach to programming. By leveraging augmented analytics capabilities and machine learning based automation, we are making analytics easier and more accessible for everyone within an organization.

In this book, Felix Liao takes you on a tour of how SAS Viya empowers any user to uncover deeper insights using powerful analytics techniques. Felix reminds us that there is so much more to visualization today beyond just using traditional charts and graphs. Through step-by-step examples using real world data, the book guides the reader through how to apply statistical and machine learning techniques using a visual framework in order to answer complex business questions and extract valuable insights.

Shadi Shahin
Vice President, Product Strategy
SAS

About This Book

What Does This Book Cover?

This book focuses on how smart data discovery can empower everyone in an organization to leverage data in powerful ways and derive valuable insights. By leveraging the powerful visual interface of SAS Viya and more advanced analytics and machine learning-based techniques, the book demonstrates how to analyze business problems in new ways as well as derive actionable insights quickly and easily.

The main topics covered in this book includes the benefits of smart data discovery, the overall approach to smart data discovery, as well as how to apply specific smart data discovery techniques to solve business problems using SAS Viya.

This book does not cover how SAS Viya can be used for data discovery using programming techniques nor does it cover advanced modeling concepts such as pipeline modeling or model management.

Is This Book for You?

The intended audience of this book consists of business users and analysts who want to leverage data and analytics to drive actionable insights across multiple business functions.

It is for people who are familiar with traditional reporting or data visualization techniques but want to tap into the power of more advanced analytics and machine learning techniques in order to tackle more complex business questions.

What Are the Prerequisites for This Book?

While it would be helpful to have some familiarization with SAS Viya, there are no real prerequisites that are necessary in order to benefit from reading this book. This book introduces foundational knowledge that is needed in order to leverage more advanced statistical and machine learning techniques. Because it focuses on a visual approach to insight discovery and modeling, there are also no knowledge requirements around any programming languages.

What Should You Know about the Examples?

This book includes step-by step examples that you can follow to gain hands-on experience with SAS Viya. These examples use real data and demonstrate how specific analytics techniques can be applied in order to answer complex business questions.

Software Used to Develop the Book's Content

The examples used throughout the book leverage SAS Viya 3.5.

Example Code and Data

The examples in the book use the World Development Indicators data set published by the World Bank. You can access the data via the following link:

http://datatopics.worldbank.org/world-development-indicators/

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or https://support.sas.com/author_feedback.

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: sas.com/books.

About The Author

Felix Liao is a manager within the customer advisory team at SAS and is also responsible for the analytics platform product portfolio for SAS Australia and New Zealand. He has over 15 years of experience working in the Australian and New Zealand analytics market. Felix was responsible for the regional launch of SAS Viya and was also responsible for the successful launch of SAS Visual Analytics in Australia and New Zealand in 2012. He is a regular speaker and blogger on the topic of analytics, data visualization, and machine learning. A computer engineer from his undergraduate study, Felix obtained his MBA in 2009 from Macquarie University, and he is also a SAS certified data scientist. His diverse background allows him to bring a wide set of views and perspectives, which are critical in modern analytics and machine learning projects and initiatives.

Learn more about this author by visiting his author page at http://support.sas.com/liao. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

Chapter 1: **Why Smart Data Discovery?**

Introduction

As information workers, our ability to leverage data and extract insights in order to make critical business decisions is fundamental to our success as individuals and the organizations that we work for. Regardless of whether you are an executive, departmental decision maker, or an analyst, the need to leverage data and analytical techniques effectively in order make business decisions is now pervasive throughout every part of an organization.

From the organization's perspective, the historical approach of relying solely on statisticians or decision support specialists to prepare and analyze data is no longer a workable approach. Organizations today must involve everyone in their analytics efforts – especially those closest to core business functions – to truly leverage data for maximum strategic and tactical advantage.

The good news for both us as individuals and the organizations that we work for is that tremendous advancements in computer hardware and software in recent years have allowed us to collect more data than ever before. Furthermore, with the addition of advanced analytics and machine learning capabilities, modern analytics tools are now easier to use and have never been more powerful. These shifts have made data more accessible and true self-service analytics a reality today.

One area of analytics that has made a significant impact in recent years is self-service data visualization. These easy-to-use data visualization and exploration tools enable any information workers today to assemble data rapidly, explore hypotheses visually, and find new insights quickly. Data visualization tools empower business users and accelerate the process of insight discovery by reducing the need for statisticians, data modelers, or IT specialists. By shifting the process of insight discovery closer to the business and subject matter experts, it has enabled more timely and relevant insights to be discovered and acted upon.

> "The greatest value of a picture is when it forces us to notice what we never expected to see."
>
> – John Tukey

Not only have these new data visualization tools accelerated the process of insight discovery, they have also allowed business users to ask more complex and forward-looking questions. These powerful, visual-based data discovery tools have revolutionized the traditional business intelligence solution space and led the way in terms of self-service analytics. Never before has it been easier for individual users to explore and visualize data for powerful insight with such ease.

With growing awareness and understanding of advanced data visualizations techniques, business users are now increasingly asking more complex questions, conduct more forward-looking analysis and eager to move beyond basic charts and graphs for answers. Enter the era of smart data discovery and the rise of citizen data scientist. Smart data discovery extends beyond the realm of traditional charts and visualization techniques with embedded machine learning techniques and algorithms. This new, augmented approach to data discovery leverages new visualization frameworks and automated machine learning capabilities to empower a new generation of users often described as citizen data scientists. Smart data discovery enables deeper insight discovery and empowers these new citizen data scientists to conduct deeper investigation, ask forward-looking questions, and develop valuable predictive insights.

What Is a Citizen Data Scientist?

Gartner defines a citizen data scientist as "a person who creates or generates models that leverage predictive or prescriptive analytics, but whose primary job function is outside of the field of statistics and analytics."

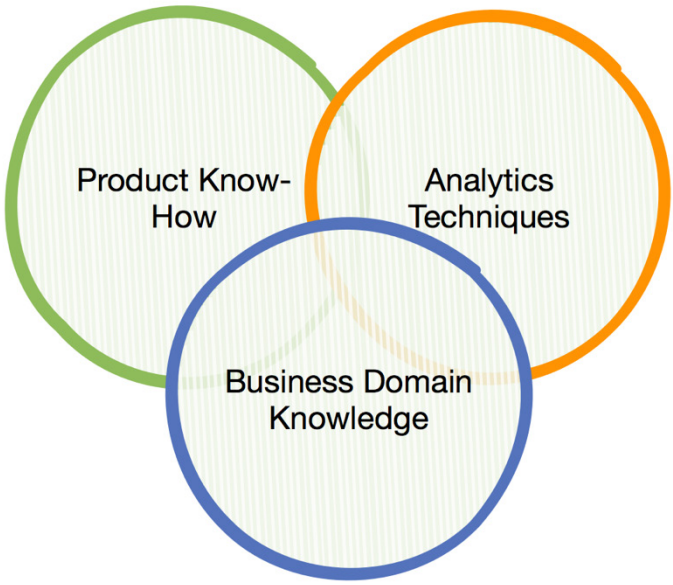
According to Gartner and other industry analysts, citizen data scientists are "power users" who can perform both simple and moderately sophisticated analytical tasks that would previously have required more expertise. They provide a complementary role to expert data scientists.

While smart data discovery holds tremendous promise, a new approach is needed in order to fully release its potential. At the core of this new approach is the recognition that citizen data scientists will need to be equipped with a new set of knowledge and skills, including the following:

1. **Business Domain Knowledge –** Smart data discovery needs to be built on a deeper understanding of the relevant business context and problem domains. Smarter insights can only come from asking more relevant and intelligent questions.
2. **Analytics Techniques** – A high level of familiarity with various analytical techniques and principles is required. While a PhD in Statistics is not necessary, a sound understanding of fundamental statistics and machine learning principles and techniques will be needed.
3. **Product Know-How** – Finally, it is about having access to the right tools and the necessary skills to bring it all together in a timely manner.

As depicted in Figure 1.1 below, these three skill sets in isolation are valuable, but when combined to solve a specific business problem, a new level of analytics and insight can be achieved – in this case, the sum is greater than its parts.

Figure 1.1: Smart Data Discovery Requirements



This book will help you navigate these intersecting knowledge domains and empower you to ask more complex questions by illustrating the key components of a smart data discovery process. We will highlight fundamental statistical concepts and how to leverage the relevant features. Most importantly, we will also be using real examples and applications to bring these concepts together.

SAS Viya, the latest evolution of the SAS platform will be used to demonstrate these examples throughout this book. An introduction to relevant features and functionalities that are needed in a smart data discovery process will be provided, followed by an explanation on how to leverage and interpret the various charts and outputs from SAS Viya.

Smart data discovery has the potential to shift an organization's overall analytic maturity, accelerate its analytical efforts, and create a much bigger analytics workforce. From an individual perspective, it has the potential to transform the way you view data, conduct data discovery processes, and think about how complex business problems can be solved. In many ways, we are just at the start of this revolution, and I am hopeful that this book will help you and your organization lead the way in terms of realizing the true potential of data and analytics.

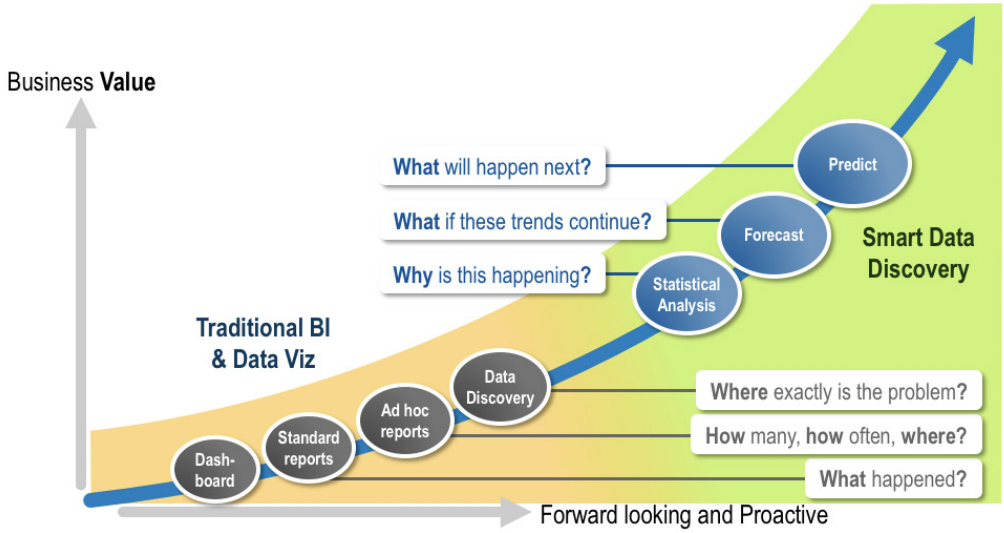
Why Smart Data Discovery Now?

Traditional Business Intelligence (BI) and data visualization tools do a great job of slicing and dicing data in order to help answer questions such as what happened and what is happening. These tools can also provide valuable dashboards and reports for the purpose of insights sharing and communication. However, they cannot easily identify correlating factors or help predict

future outcomes. These outcomes might include: which customers will respond to a promotion? Who will churn to a competitor? And when will a piece of equipment fail? In a modern environment where businesses are no longer content with simply analyzing data from the past, but instead would also like to gain insights into the future, a new approach is clearly needed.

Smart data discovery extends the traditional data visualization paradigm by marrying intuitive visualizations with predictive analytics techniques and machine learning algorithms. As illustrated in Figure 1.2, the use of these more advanced techniques changes the nature of analysis from reactive to proactive. It changes the perspective from backward-looking to forward-looking, and it empowers the analytics professionals to dig deeper, investigate root causes, and predict future outcomes.

Figure 1.2: From Reactive to Forward-Looking and Proactive



What is not widely known is that data visualization has always been a fundamental tool for expert data scientists. Visualization techniques are often used by expert data scientists for correlation analysis, model feature selection, and testing different hypotheses quickly. Techniques range from the use of box plots and histograms for general exploration to the use of decision tree and scatter plots for feature selection. These exploration steps often act as the precursor to more complex predictive modeling techniques. Smart data discovery extends these analytical techniques to a much broader audience using an intuitive and visual-oriented approach that does not require complex programming or deep statistical knowledge.

From an organization's perspective, smart data discovery has the potential to minimize the challenges associated with resourcing and staffing that most organizations face today. In an environment where there is still severe shortage of experienced, expert data scientists, smart data discovery not only has the potential to empower more users to leverage advanced machine learning techniques, it can also reduce the friction between the expert data scientist and the broader analytics professional community.

Data science is widely recognized as a highly collaborative process that require inputs from multiple teams and different personas. Expert data scientists are important and valuable, but they are only one part of the puzzle. Expert data scientists typically have a strong math and programming background and have a high-level understanding of the business process and functions. They normally do not have in-depth knowledge of the functional parts of a business versus someone who is directly involved in a particular line of business. Problems associated with marketing, fraud, customer service, and the supply chain all require deep and relevant domain knowledge in order to solve. This is where citizen data scientists fit in. They are typically used in lines of business and are intimately involved with core functional business areas. As a result, they tend to have a greater understanding and appreciation of the challenges being faced and how to solve them. This deep domain knowledge is critical in the success of smart data discovery processes. Armed with a new approach to data exploration and a powerful solution, smart data discovery can not only empower these citizen data scientists to solve more complex business problems, it can also bring business communities closer to the expert data science teams, which can only be a good thing.

Empowering and developing the citizen data scientist community will also benefit the expert data scientist community in many ways. As citizen data scientists improve their ability to ask more complex questions and test hypothesis more quickly, they can then communicate their findings with the expert data scientist community in a timelier manner and direct them to the more relevant, high-value problem domain areas. The expert data scientists will appreciate the deeper insights and analysis generated by the citizen data scientists, which will allow them to better prioritize and focus their time and efforts. This paradigm shift will lift an organization's overall analytical capabilities as well as create a more analytics-focused culture.

Who Is This Book For?

Whether it be basic descriptive analysis or sophisticated diagnostic and predictive analysis, the need to leverage data and analytical techniques in order to make important business decisions is everyone's business today. From that perspective, this book is really for everyone in every part of an organization. Having said that, this book takes a pragmatic approach and is targeted squarely at any analytics professional who need to bring in data, explore, prototype, and move the needle forward in terms of finding new insight and create insight value.

The SAS products that we will be using throughout this book are SAS Visual Analytics and SAS Visual Statistics. These two products form the foundation of SAS Viya and are targeted at non-programmers and business users. If you are an existing SAS Visual Analytics or SAS Visual Statistics user, you will benefit from the focus around advanced visualization and machine learning techniques covered in this book. We will be covering visualizations and features that are often unused or put into the "too hard" basket. The goal of this book is to help you realize the true potential and value of the tools that you and your organization have invested in.

On the other hand, if you are not familiar with the world of SAS Visual Analytics and SAS Visual Statistics, we will introduce the basic elements of each tool. You will find a dedicated chapter (Chapter 3) where we will provide you with a high-level overview of the key features and functionalities needed in a smart data discovery process. It is important to note that this book should not be treated as a training manual for SAS Visual Analytics or SAS Visual Statistics. We

will not be covering every single feature or control offered by these tools. If that is what you want, there are many excellent books and training resources that cover these two tools in lot more depth, especially around the reporting capabilities that are really not the focus of this book.

While SAS Visual Analytics and SAS Visual Statistics offer powerful programming capabilities and user interfaces targeted at the programmers, those will also not be the focus of this book. We will primarily be focusing on the graphical interface. Once again, you can find out more about these programming capabilities via standard product documentations and SAS training courses.

What if you are intimidated by math? If you are in this group, then I have some good news for you: in order to get started leveraging machine learning techniques and build basic predictive models, you need less math background and knowledge than you think (and almost certainly less math than you have been told that you need!). While the role of a citizen data scientist does require a sound understanding of statistics and machine learning principles, you generally do not need to understand the mathematical underpinnings used to construct these algorithms in the first place. Modern statistical and automated machine learning software such as SAS Visual Analytics and SAS Visual Statistics take care of much of the mathematics for you, enabling you to focus on interpreting the outputs of these advanced techniques instead.

The only thing you are required to bring is your relevant business domain knowledge. Someone who has the relevant domain knowledge and is always generating interesting business problems and hypotheses will benefit greatly from this book. Ultimately, if you believe in the power of data and analytics and are curious as to how they can help you answer some of your most difficult questions, then this book is for you. I am hopeful that the combination of relevant foundational knowledge and practical advice in the following chapters will help unleash the true citizen data scientist in you.

Chapter Overview

While it might be tempting to jump into the chapter on prediction using decision trees to tackle your current business requirement, this book is written in such a way that it builds on the knowledge developed in prior chapters. The recommended way to read this book is to start from the beginning. Once you have finished the book, you can then use each chapter as a reference guide for specific challenges.

This book is loosely arranged into three main parts. The first part (Chapters 1 and 2) explains the what and why of smart data discovery and the role it plays for a citizen data scientist.

The second part (Chapters 3 and 4) provides foundational knowledge on the tools and data needed in a smart data discovery process. Chapter 3 introduces the relevant user interface components as well as key capabilities of SAS Visual Analytics and SAS Visual Statistics. Chapter 4 highlights the role data plays in a smart data discovery process and how to manage it effectively.

The third part (Chapter 5 through Chapter 10) dives into specific techniques as well as how these techniques can be applied to solve real business problems and extract valuable insights. While we do not delve into complex mathematical equations, various foundational statistical concepts are introduced throughout these chapters to support the examples and technique used. Chapter 10 concludes the book, but it introduces areas that you can explore further in order to build on the knowledge you have gained.