Applied Analytics through Case Studies Using SAS and R

Implementing Predictive Models and Machine Learning Techniques

Deepti Gupta

Apress®

Applied Analytics through Case Studies Using SAS and R

Implementing Predictive Models and Machine Learning Techniques

Deepti Gupta

Apress®

Applied Analytics through Case Studies Using SAS and R

Deepti Gupta
Boston, Massachusetts, USA

ISBN-13 (pbk): 978-1-4842-3524-9 ISBN-13 (electronic): 978-1-4842-3525-6
https://doi.org/10.1007/978-1-4842-3525-6

Library of Congress Control Number: 2018952360

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestin John
Development Editor: James Markham
Coordinating Editor: Divya Modi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. 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 rights@apress.com, or visit http://www.apress.com/rights-permissions.

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 via the book's product page, located at www.apress.com/978-1-4842-3524-9. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

I am dedicating this book to my Family.

Table of Contents

About the Author

Deepti Gupta completed her MBA in Finance & PGPM in Operation Research in 2010. She has worked with KPMG and IBM private limited as a Data Scientist and is currently working as a data science freelancer. Deepti has extensive experience in predictive modeling and machine learning and her expertise is in SAS and R. Deepti has developed data science courses and delivered data science trainings and conducted workshops in both corporate and academic institutions. She has written multiple blogs and white papers. Deepti has a passion for mentoring budding data scientists.

About the Contributor

Dr. Akshat Gupta is currently working as a Senior Applications Engineer at MilliporeSigma in Applications Engineering, Global Manufacturing Sciences and Technology (MSAT) group. He authored the health-care case study (Chapter5) of this book. His focal area of research is cell culture clarification and tangential flow filtration. Dr. Gupta has extensive experience in Design of Experiments (DOE) and statistical analysis. He holds a Bachelor of Technology (B.Tech) degree in Chemical Engineering from the Vellore Institute of Technology, and a Master of Science (MS) and Doctor of Philosophy (Ph.D.) in Chemical Engineering from the University of Massachusetts Lowell. He also has graduate certificates in Modeling and Simulation, and Nanotechnology.

About the Technical Reviewer

Preeti Pandhu has a Master of Science degree in Applied (Industrial) Statistics from the University of Pune. She is SAS certified as a base and advanced programmer for SAS 9 as well as a predictive modeler using SAS Enterprise Miner 7. Preeti has more than 18 years of experience in analytics and training.

She started her career as a lecturer in statistics and began her journey into the corporate world with IDeaS (now a SAS company), where she managed a team of business analysts in the optimization and forecasting domain. She joined SAS as a corporate trainer before stepping back into the analytics domain to contribute to a solution-testing team and research/consulting team. She was with SAS for 9 years. Preeti is currently passionately building her analytics training firm, DataScienceLab (`www.datasciencelab.in`).

Acknowledgments

Book writing is one of the most interesting and challenging attempt one can take up. This book could not have been completed without the encouragement, guidance, and support of my family. I would like to thank Dr. Akshat Gupta, Ved Prakash Garg, Col. Atul Gupta, Dr. Anvita Garg, Ayush Gupta, RS Miyan, Ansi Miyan, Dr. James Chrostowski, and my colleagues and friends for their productive discussions and suggestions. My special thanks to Celestin John who provided great help on everything ranging from technical support to answering my queries. I appreciate the thoughtful and insightful comments from the editor and the reviewers. Thanks to the Apress team, especially to Divya Modi for all the patience, support, and guidance in completing this project.

Introduction

Analytics is a big buzz and a need for today's industries to solve their business problems. Analytics helps in mining the structured and unstructured data in order to withdraw the effective insights from the data, which will help to make effective business decisions. SAS and R are highly used tools in analytics across the globe by all industries for data mining and building machine learning and predictive models. This book focuses on industrial business problems and a practical analytical approach to solve those problems by implementing predictive models and machine learning techniques using SAS and R analytical languages.

The primary objective of this book is to help statisticians, developers, engineers, and data analysts who are well versed in writing codes; have a basic understanding of data and statistics; and are planning to transition to a data scientist profile. The most challenging part is practical and hands-on knowledge of building predictive models and machine learning algorithms and deploying them in industries to address industrial business problems. This book will benefit the reader in solving the business problems in various industrial domains by sharpening their analytical skills in getting practical exposure to various predictive model and machine learning algorithms in six industrial domains.

What's in This Book

This book focuses on industrial business problems and practical analytical approaches to solve those problems by implementing predictive models and machine learning techniques using SAS Studio and R analytical languages. **This book contains six industrial case studies of various domains with data and all the codes in SAS Studio and R languages, which would benefit all readers to practice and implement these models in their own business cases.**

In Chapter 1 the general outline about analytics, the role of analytics in various industries, and a few popular data science and analytical tools are discussed. Chapter 2 describes the role of analytics in the banking industry with a detailed explanation of predicting a bank loan default case study in R and SAS. Chapter 3

describes how analytics contribute in the retail industry and offers a detailed explanation of forecasting a case study in R and SAS. Chapter 4 describes how analytics is reshaping the telecommunications industry and gives a detailed explanation of a case study on predicting customer churn in R and SAS. Chapter 5 describes the application of analytics in the healthcare industry and gives a clear explanation of a case study on predicting the probability of benign and malignant breast cancer using R and SAS. Chapter 6 describes the role of analytics in the airline industry and provides a case study on predicting flight arrival delays (minutes) in R and SAS. Chapter 7 describes the application of analytics in the FMCG industry with a detailed explanation of a business case study on customer segmentation based on their purchasing history using R and SAS.

Who's the Target Audience?

- Data Scientists who would like to implement machine learning techniques with a practical analytical approach toward a particular industrial problem.
- Statistician, Engineers, and Researchers with a great theoretical understanding of data and statistics and would like to enhance their skills by getting practical exposure to data modeling.
- Data analysts who know about data mining but would like to implement predictive models and machine learning techniques.
- Developers who are well versed with coding but would like to transition to a career in data science.

What You Will Learn

- Introduction to analytics and data understanding.
- How to approach industrial business problems with an analytical approach.
- Practical and hands-on knowledge in building predictive model and machine learning techniques.
- Building the analytical strategies.

CHAPTER 1

Data Analytics and Its Application in Various Industries

Data analytics has become part and parcel of any business in today's world. In fact, it has evolved into an industry in itself. Vast numbers of software platforms are available for data extraction, scrubbing, analysis, and visualization. Some of these platforms are specialized for carrying out one of the above-listed aspects of data analytics, while others offer a generalist tool to carry out almost all tasks ranging from data scrubbing to visualization. Of these platforms, SAS® and R are the most popular for data analytics with a large global clientele.

In 1967, Statistical Analysis System (SAS) started as a federal funded project for graduate students to track agriculture data at North Carolina State University.[1] Today it has become a global leader in data analysis software market with customers spanning over 148 countries.[2] Ninety-six of the top 100 Fortune Global 500® companies use SAS. R, which originally was a statistical computing language, has advanced significantly over the years. R Studio is an Integrated Development Environment (IDE) for R[3] and offers a free, user-friendly platform for data analytics. Both SAS® and R offer vast capabilities but have certain contrasting advantages that are discussed later in more detail.

A broad array of companies ranging from the largest global banks to regional transport firms are using data analytics to solve diverse sets of problems These diverse applications have one commonality: using data and statistics as the basis for decision making.

In this chapter, certain key aspects related to data analytics will be introduced.

D. Gupta, *Applied Analytics through Case Studies Using SAS and R*,
https://doi.org/10.1007/978-1-4842-3525-6_1

What Is Data Analytics?

Analytics is defined as the process of developing the actionable insights through the application of statistical model and analysis from the data.[4] Applying data analytics for decision making is a systematic process. It starts with understanding the nature of industry, general functionality, bottlenecks, and challenges specific to the industry. It is also helpful to know who the key companies are, size of industry, and in some cases general vocabulary and terms associated with operations. After that we take a deeper dive in to the area specific to the application or a business case to which data analytics needs to be applied. A thorough understanding of the application, associated variables, sources of data, and knowledge of the reliability of different data sources are very important.

Data analytics firms pay a lot of attention to these aspects and often employ a vast number of subject-matter experts specific to industries and at times even specific to certain key applications. Business research consultants are also employed for gaining understanding and insights in certain cases. During the preliminary phase of a project, data analytics firms perform elaborate surveys and conduct series of interviews to gain more information about the company and the business problem.[5] A good understanding of industry and the application can result in significant cost saving and can improve accuracy, performance, and practicality of the model.

Once the application or the problem statement is well understood, then the implementation process starts. The core methodology of implementing data analytics for solving a business problem is shown in Figure 1-1.[6]

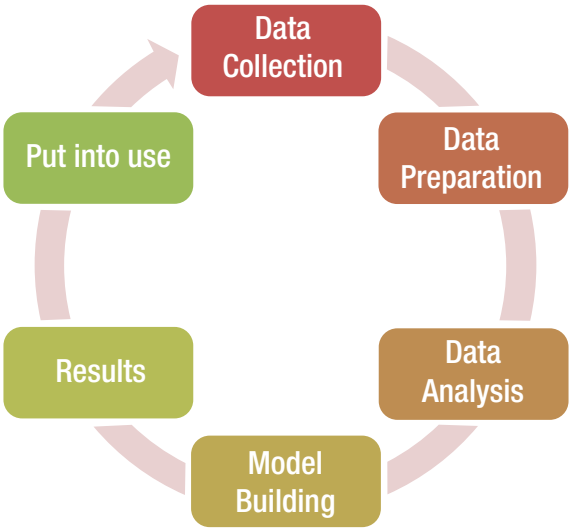

Figure 1-1. *Data Analytics Methodology*

Data Collection

The first step in the process is data collection. Data relevant to the applicant is collected. The quality, quantity, validity, and nature of data directly impact the analytical outcome. A thorough understanding of the data on hand is extremely critical.

It is also useful to have an idea about some other variables that may not directly be sourced from the industry or the specific application itself but may have a significant impact if included into the model. For example, when developing a model to predict flight delays, weather can be a very important variable, but it might have to be obtained from a different source then the rest of the data set. Data analytics firms also have ready access to certain key global databases including weather, financial indices, etc. In recent years, data mining of digital social media like Twitter and Facebook is also becoming very popular.[7] This is particularly helpful in understanding trends related to customer satisfaction with various services and products. This technique also helps reduce the reliance on surveys and feedbacks. Figure 1-2 shows a Venn diagram of various sources of data that can be tapped into for a given application.

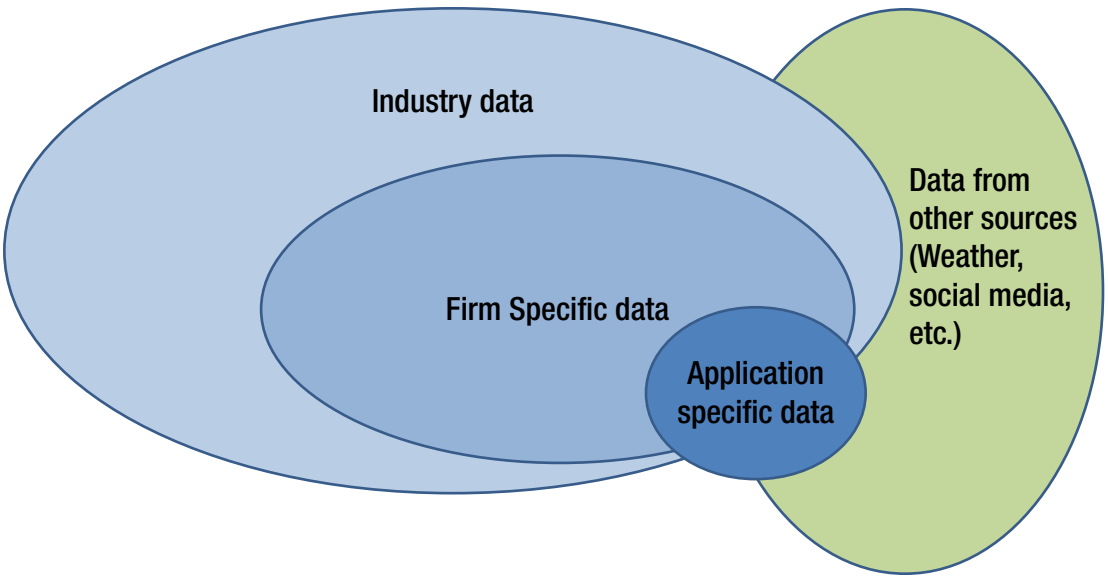

Figure 1-2. *Venn diagram of data sources*

Data Preparation

The next step is data preparation. Usually raw data is not in a format that can be directly used to perform data analysis. In very simple terms, most platforms require data to be in a matrix form with the variables being in different columns and rows representing various observations. Figure 1-3 shows an example of structured data.

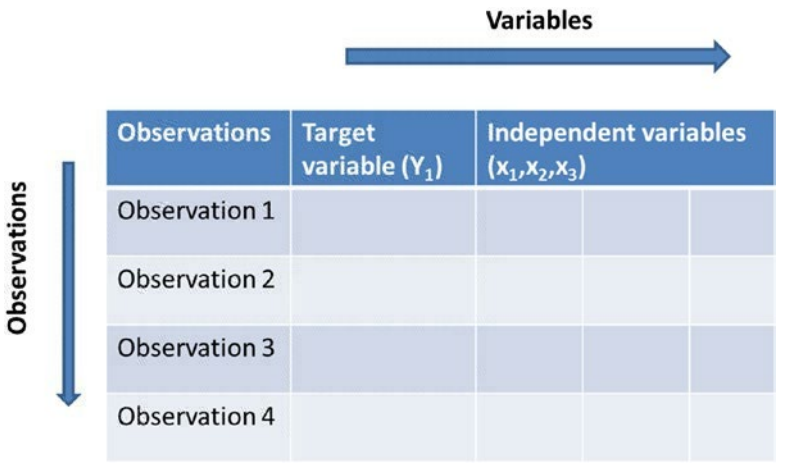

Figure 1-3. *Format of structured data*

Data may be available in structured, semi-structured, and unstructured form. A significant effort is needed to align semi-structured and unstructured data into a usable form as shown in Figure 1-3. Once the data is brought together in a structured form, the next stage in data preparation is data cleansing or scrubbing. Data scrubbing encompass processes that help remove inconsistencies, errors, missing values, or any other issues that can pose challenges during data analysis or model building with a given data set.[8] Work at this stage can be as simple as changing the format of a variable, to running advanced algorithms to estimate suitable estimates for missing values. This task is significantly more involved when it comes to big data.

Data Analysis

Once data is converted into a structured format, the next stage is to perform data analysis. At this stage underlying trends in the data are identified. This step can include fitting a linear or nonlinear regression model, performing principal component analysis or cluster analysis, identifying if data is normally distributed or not. The goal is to identify

what kind of information can be extracted from the data and if there are underlying trends that can be useful for a given application. This phase is also very useful for scoping out the models that can be most useful to capture the trends in data and if the data satisfies underlying assumptions for the model. One example would be to see if the data is normally distributed or not to identify if parametric models can be used or a non-parametric model is required.

Model Building

Once the trends in data are identified, the next step is to put the data to work and build a model that will help with the given application or help solve a business problem. A vast number of statistical models are available that can be used, and new models are being developed every day. Models can significantly vary in terms of complexity and can range from simple univariate linear regression models to complex machine learning algorithms. Quality of a model is not governed by complexity but rather by its ability to account for real trends and variations in data and sift information from noise.

Results

Results obtained from the models are validated to ensure accuracy and model robustness. This can be done two ways; the first is by splitting the original data set into training and validation data sets. In this approach, part of the data is used for model building and the remaining part is used for validation. The other approach is to validate data against real-time data once the model is deployed. In some cases, the same data is used to build multiple different types of models to confirm if the model outputs are real and not statistical artifacts.

Put into Use

Once the model is developed it is deployed in a real-time setting for a given application. As shown in the Figure 1-1, the overall process is somewhat iterative in nature. Many times, the models have to be corrected and new variables added or some variables removed to enhance model performance. Additionally, models need to be constantly recalibrated with fresh data to keep them current and functional.

Types of Analytics

Analytics can be broadly classified under three categories: descriptive analytics, predictive analytics, and prescriptive analytics.[9] Figure 1-4 shows the types and descriptions of types of analytics.

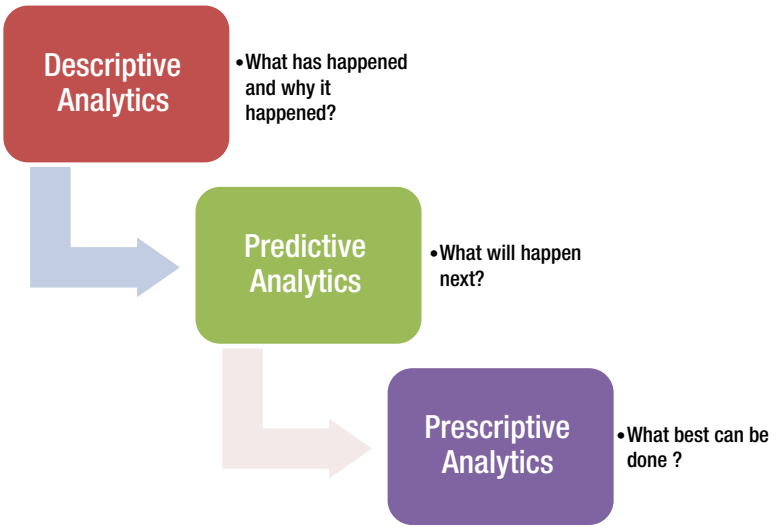

Figure 1-4. *Types of Analytics*

Different types of information can be obtained by applying the different categories of analytics. This will be explained in the following section.

1. **Descriptive Analytics**: Most of the organizations use descriptive analytics in order to know about their company performance. Example, management at a retail firm can use descriptive analytics to know the trends of sales in past years, or inferring trends of operation cost, product, or service performance.

2. **Predictive Analytics**: In case of predictive analytics, historical trends coupled with other variables are used to see what could happen in the future to the firm. Example, Management at the same retail firm can use the sales trends from previous years to forecast sales for the coming year.

3. **Prescriptive Analytics**: In prescriptive analytics, the objective is to identify factors or variables that are impacting trends. Once the responsible variables are identified, strategies and recommendations are made to improve the outcome. For example, Management at the same retail firm identifies that the operation cost is significantly high due to overstocking at certain stores. Based on this insight, an improved inventory management would be recommended to the given locations.

Understanding Data and Its Types

Data is a collection of variables, facts, and figures that serves as raw material to create information and generate insights. The data needs to be manipulated, processed, and aligned in order to withdraw useful insights. Data is divided into two broad forms: qualitative and quantitative data.[10]

1. **Qualitative data**: The data that is expressed in words and descriptions like text, images, etc. is considered as qualitative data. Qualitative data collection uses unstructured and semi-structured techniques. There are various common methods to collect qualitative data like conducting interviews, diary studies, open-ended questionnaires, etc. Examples of qualitative data are gender, demographic details, colors, etc. There are three main types of qualitative data:

    - **Nominal**: Nominal data can have two or more categories but there is no intrinsic rank or order to the categories. For example, gender and marital status (single, married) are categorical variables having two categories and there is no intrinsic rank or order to the categories.

    - **Ordinal**: In ordinal data, the items are assigned to categories and there is an intrinsic rank or order to the categories. For example, age group: Infant, Young, Adult, and Senior Citizen.

    - **Binary**: Binary data can take only two possible values. For example, Yes/No, True/False.

2. **Quantitative data**: The data that is in numerical format is considered as quantitative data. Such a type of data is used in conducting quantitative analysis. Quantitative data collection uses much more structured techniques. There are various common methods to collect quantitative data like surveys, online polls, telephone interviews, etc. Examples of quantitative data are height, weight, temperature, etc. There are two types of quantitative data:

    - **Discrete Data**: Discrete data is based on count and it can only take a finite number of values. Typically it involves integers. For example, the number of students in data science class is discrete data because you are counting a whole and it cannot be subdivided. It is not possible to have 8.3 students.
    - **Continuous Data**: Continuous data can be measured, take any numeric values, and be subdivided meaningfully into finer and finer levels. For example, the weights of the data science students can be measured at a more precise scale – kilograms, grams, milligrams, etc.

While on the topic of data, it is a good time to get a basic understanding of "Big Data."

Big Data is not just a buzzword but is fast becoming a critical aspect of data analytics. It is discussed in more detail in the following section.

What Is Big Data Analytics?

The term "big data" is defined as the huge volume of both structured and unstructured data that is so large that it is not possible to process such data using traditional databases and software. As a result, many organizations that collect, process, and conduct big data analysis turn to specialized big data tools like NoSQL databases, Hadoop, Kafka, Mapreduce, Spark, etc. Big data is a huge cluster of numbers and words. Big data analytics is the process of finding the hidden patterns, trends, correlations, and other effective insights from those large stores of data. Big data analytics helps organizations

harness their data to use it for finding new opportunities, faster and better decision making, increased security, and competitive advantages over rivals, such as higher profits and better customer service. Characteristics of Big data are often described using 5 Vs, which are velocity, volume, value, variety, and veracity.[11] Figure 1-5 illustrates 5 Vs related to the big data.

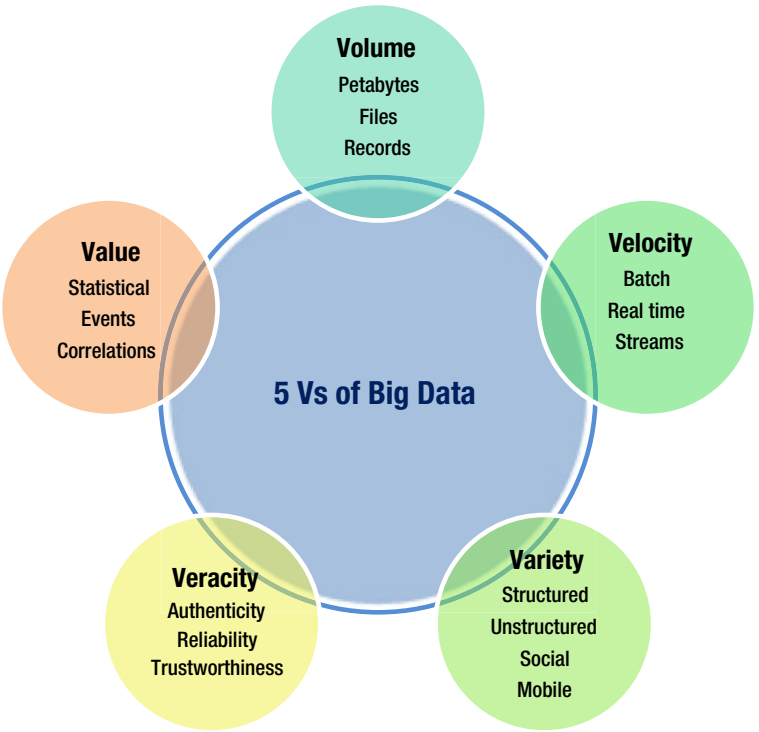

Figure 1-5. *5 Vs of Big Data*

Big Data analytics applications assist data miners, data scientists, statistical modelers, and other professionals to analyze the growing volumes of structured and mostly unstructured data such as data from social media, emails, web servers, sensors, etc. Big data analytics helps companies to get accessibility to nontraditional variables or sources of information, which helps organizations to make quicker and smarter business decisions.

Big Data Analytics Challenges

Most of the organizations are experiencing effective benefits by using big data analytics, but there are some different obstacles that is making it difficult to achieve the benefits promised by big data analytics.[12] Some of the key challenges are listed below:

- Lack of internal skills: The most important challenge that organizations face in implementing big data initiatives is lack of internal skills, and there is a high cost of hiring data scientists and data miners for filling the gaps.
- Increasing growth of the data: Another important challenge of big data analytics is the growth of the data at a tremendous pace. It creates issues in managing the quality, security, and governance of the data.
- Unstructured Data: As most of the organizations are trying to leverage new and emerging data sources, it is leading to the more unstructured and semi-structured data. These new unstructured and semi-structured data sources are largely streaming data coming from social medial platforms like Twitter, Facebook, web server logs, Internet of Things (IOT), mobile applications, surveys, and many more. The data can be in the form of images, email messages, audio and video files, etc. Such unstructured data is not easy to analyze without having advanced big data analytical tools.
- Data Siloes: In organizations there are several types of applications for creating the data like customer relationship management (CRM), supply chain management (SCM), enterprise resource planning (ERP), and many more. Integrating the data from all these wide sources is not an easy task for the organization and is one of the biggest challenges faced by big data analytics.

Data Analytics and Big Data Tools

Data science and analytics tools are evolving and can be broadly classified into two classes: tools for those techies with high levels of expertise in programming and profound knowledge of statistics and computer science like R, SAS, SPSS, etc.; and tools for common audiences that can automate the general analysis and daily reports like Rapid Miner, DataRPM, Weka, etc. Figure 1-6 displays the currently prevalent languages, tools, and software that are used for various data analytics applications.

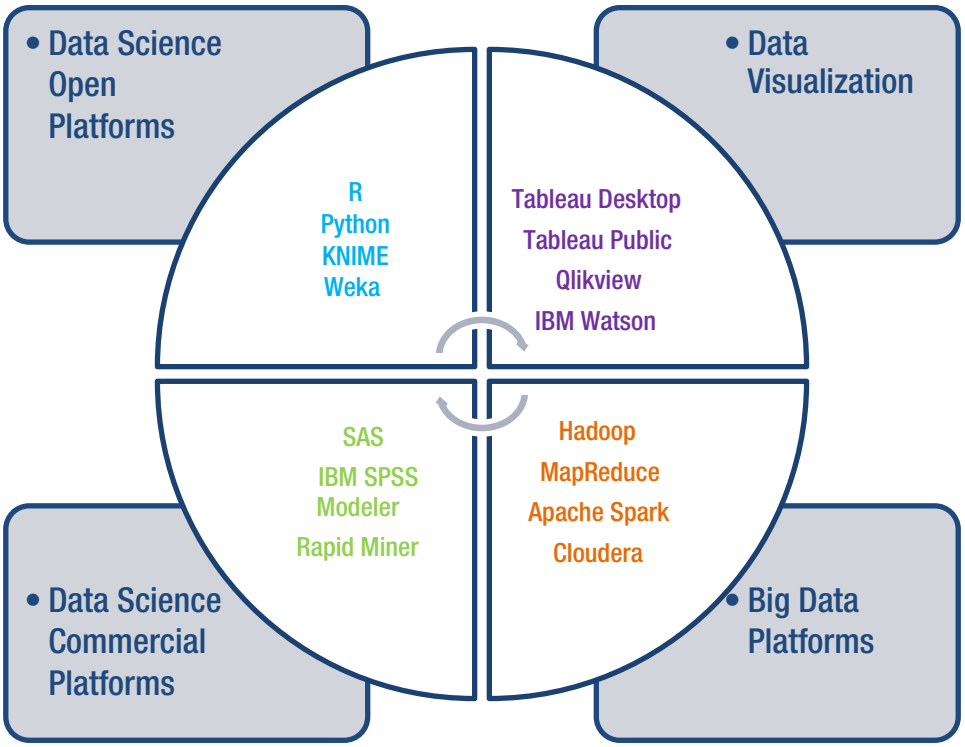

Figure 1-6. *Languages, Tools, and Software*

There is a long list of tools, and a few popular data science and analytical tools are discussed in the following section.

1. **R: The Most Popular Programming Language for statisticians and data scientists**

   R is an open source tool widely used by statisticians and data miners for conducting statistical analysis and modeling.[13] R has thousands of packages available easily that make the jobs of statisticians and data scientists easy for handling the tasks from text analytics to voice recognition, face recognition, and genomic science. The demand of R has increased dramatically across all the industries and is becoming popular because of its strong package ecosystem. R is used in industries for solving their big data issues and building statistical and predictive models for withdrawing the effective insights and hidden patterns from the data.

2. **SAS (Statistical Analysis System) Data Science and Predictive Analytics Software Suite**

   SAS is a software suite that is popular for handling large and unstructured data sets and is used in advance analytics, multivariate analysis, data mining, and predictive analytics, etc. The SAS software suite has more than 200 components like BASE SAS, SAS/ STAT, SAS/ETS, SAS/GRAPH, etc. BASE SAS software, SAS Enterprise Guide, and SAS Enterprise Miner are licensed tools and are used for commercial purposes by all the industries. SAS University Edition is free SAS software and is used for noncommercial uses like teaching and learning statistics and modeling in an SAS environment. It includes the SAS components BASE SAS, SAS/STAT, SAS/IML, SAS/ACCESS, and SAS Studio. SAS can be expensive but it is a very popular tool in industries; it has an effective and quick support system and more than 65,000 customers.

3. **IBM SPSS Statistics and SPSS Modeler: Data Mining and Text Analytics Software**

   SPSS Modeler and SPSS Statistics were acquired by IBM in 2009 and is considered as a data mining, statistical, and text analytics software. It is used to load, clean, prepare the data, and then build the predictive models and conduct other analytical and statistical tasks. It has the visual interface so users without good programming knowledge can easily build the predictive model and statistical analysis.[14] It has been widely used in industries for fraud detection, risk management, forecasting, etc. IBM SPSS modeler (version 17) is present in two separate bundles as:

   1. SPSS Modeler Professional: it is used for structured data such as databases, flat files, etc.
   2. SPSS Modeler Premium: it is a high-performance analytical tool that helps in gaining effective insights from the data. It includes all the features from SPSS Modeler Professional and in addition it is used for conducting Text Analytics,[15] Entity Analytics,[16] and Social Network Analytics.

4. **Python: High-Level Programming Language Software**

   Python is an object-oriented and high-level programming language.[17] Python is easy to learn and its syntax is designed to be readable and straightforward. Python is used for data science and machine learning. Robust libraries used for data science and machine learning are using the interface of Python, which is making the language more popular for data analytics and machine learning algorithms.[18] For example, there robust libraries for statistical modeling (Scipy and Numpy), data mining (Orange and Pattern), and supervised and unsupervised machine learning (Scikit-learn).[19]

5. **Rapid Miner: GUI Driven Data Science Software**

   Rapid Miner is open source data mining software. It was started in 2006 and was originally called Rapid-I. In 2013 the name was changed from Rapid-I to Rapid Miner. The older version of Rapid Miner is open source but the latest version is licensed. Rapid miner is widely used in industries for data preparation in visualization, predictive modeling, model evaluation, and deployment.[20] Rapid Miner has a user-friendly graphic user interface and a block diagram approach. Predefined blocks act as a plug and play system. Connecting the blocks accurately helps in building a wide variety of machine learning algorithms and statistical models without writing a single line of code. R and Python can also be used to program Rapid Miner.

Role of Analytics in Various Industries

The onset of the digital era has made vast amounts of data accessible, analyzable, and usable. This, coupled with a highly competitive landscape, is driving industries to adopt data analytics. Industries ranging from banking and telecommunication to health care and education, everyone is applying various predictive analytics algorithms in order to gain critical information from data and generate effective insights that drive business decisions.

There are vast numbers of applications within each industry where data analytics can be applied. Some applications are common across many industries. These include customer-centric applications like analyzing factors impacting customer churn, engagement, and customer satisfaction. Another big data analytics application is for predicting financial outcomes. These include forecasting of sales, revenues, operation costs, and profits. In addition to these, data analytics is also widely used for risk management and fraud detection and price optimization in various industries.

There are also large numbers of industry-specific applications of data analytics. To list a few: flight delay prediction in the aviation industry, prediction of cancer remission in health care, forecasting wheat production in agriculture.

An overview of some of the industries benefiting from predictive and big data analytics insights and, most importantly, how is discussed in this section.

1. **Insurance Industry:**

   The insurance industry has always relied on statistic to determine the insurance rates. Risk-based models form the basis for calculators that are used to calculate insurance premiums. Here is a case specific to the automotive insurance. In the United States, some of the variables in these risk-based models are reasonable but others are debatable. For example, gender is a variable that determines the insurance rate. An average American male driver pays more compared to a female driver with equivalent credentials. Today, people look upon these factors as discriminatory and demand a fairer method with higher weightage to variables that are in control of the actual drivers. The European Court of Justice has passed a ruling stating that gender cannot be used as a basis for deciding insurance premiums.[21] The current trend requires risk-based models to give consideration to individuals' statistics rather than generalized population statistics. This seems fair but does require handling significantly more data on a daily basis and new models to replace the traditional ones. Big data tools and advanced data analytics might pave the way for a fairer insurance industry of the future. Predictive analytics is also widely used by the insurance industries for fraud detection, claims analytics, and compliance & risk management.

2. **Travel & Tourism Industry:**

   The travel & tourism industry is also using big data analytics for enhancing customer experiences and offer customized recommendations. These firms use demographic statistics, average time spent by users on certain travel-related web pages, personal historic travel preferences, etc.