sas

Introduction to

Statistical and Machine Learning Methods for Data Science

Carlos Andre Reis Pinheiro
Mike Patetta

The correct bibliographic citation for this manual is as follows: Pinheiro, Carlos Andre Reis and Mike Patetta. 2021. *Introduction to Statistical and Machine Learning Methods for Data Science*. Cary, NC: SAS Institute Inc.

Introduction to Statistical and Machine Learning Methods for Data Science

ISBN 978-1-953329-64-6 (Hardcover)
ISBN 978-1-953329-60-8 (Paperback)
ISBN 978-1-953329-61-5 (Web PDF)
ISBN 978-1-953329-62-2 (EPUB)
ISBN 978-1-953329-63-9 (Kindle)

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

August 2021

Contents

About This Book

What Does This Book Cover?

This book gives an overview of the statistical and machine learning methods used in data science projects, with an emphasis on the applicability to business problem solving. No software is shown, and the mathematical details are kept to a minimum. The book describes the tasks associated with all stages of the analytical life cycle, including data preparation and data exploration, feature engineering and selection, analytical modeling considering supervised and unsupervised techniques, and model assessment and deployment. It describes the techniques and provides real-world case studies to exemplify the techniques. Readers will learn the most important techniques and methods related to data science and when to apply them for different business problems. The book provides a comprehensive overview about the statistical and machine learning techniques associated with data science initiatives and guides readers through the necessary steps to successfully deploy data science projects.

This book covers the most important data science skills, the types of different data science applications, the phases in the data science lifecycle, the techniques assigned to the data preparation steps for data science, some of the most common techniques associated to supervised machine learning models (linear and logistic regression, decision tree, forest, gradient boosting, neural networks, support vector machines, and factorization machines), advanced supervised modeling methods like ensemble models and two-stage models, the most important techniques associated to unsupervised machine learning models (clustering, association rules, sequence analysis, link analysis, path analysis, network analysis, and network optimization), the method and fits statistics to assess model results, different approaches to deploy analytical models in production, and the main topics related to the model operationalization process.

This book does not cover the techniques for data engineering in depth. It also does not provide any programming code for the supervised and unsupervised models, nor does it show in practice how to deploy models in production.

Is This Book for You?

The audience of this book is data scientists, data analysts, data engineers, business analysts, market analysts, or computer scientists. However, anyone who wants to learn more about data science skills could benefit from reading this book.

What Are the Prerequisites for This Book?

There are no prerequisites for this book.

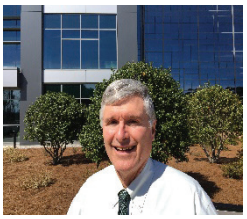
We Want to Hear from You

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 about how to become a SAS Press author
- Provide feedback on a book

About These Authors

Dr. Carlos Pinheiro is a Principal Data Scientist at SAS and a Visiting Professor at Data ScienceTech Institute in France. He has been working in analytics since 1996 for some of the largest telecommunications providers in Brazil in multiple roles from technical to executive. He worked as a Senior Data Scientist for EMC in Brazil on network analytics, optimization, and text analytics projects, and as a Lead Data Scientist for Teradata on machine learning projects. Dr. Pinheiro has a BSc in Applied Mathematics and Computer Science, an MSc in Computing, and a DSc in Engineering from the Federal University of Rio de Janeiro. Carlos has completed a series of postdoctoral research terms in different fields, including Dynamic Systems at IMPA, Brazil; Social Network Analysis at Dublin City University, Ireland; Transportation Systems at Université de Savoie, France; Dynamic Social Networks and Human Mobility at Katholieke Universiteit Leuven, Belgium; and Urban Mobility and Multi-modal Traffic at Fundação Getúlio Vargas, Brazil. He has published several papers in international journals and conferences, and he is author of Social Network Analysis in Telecommunications and Heuristics in Analytics: A Practical Perspective of What Influence Our Analytical World, both published by John Wiley Sons, Inc.



Michael Patetta has been a statistical instructor for SAS since 1994. He teaches a variety of courses including Supervised Machine Learning Procedures Using SAS® Viya® in SAS® Studio, Predictive Modeling Using Logistic Regression, Introduction to Data Science Statistical Methods, and Regression Methods Using SAS Viya. Before coming to SAS, Michael worked in the North Carolina State Health Department for 10 years as a health statistician and program manager. He has authored or co-authored 10 published papers since 1983. Michael has a BA from the University of Notre Dame and a MA from the University of North Carolina at Chapel Hill. In his spare time, he loves to hike in National Parks.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:

http://support.sas.com/pinheiro
http://support.sas.com/patetta

Chapter 1: Introduction to Data Science

Chapter Overview

This chapter introduces the main concepts of data science, a scientific field involving computer science, mathematics, statistics, domain knowledge, and communication. These are the main goals of this chapter:

- Explain the most important tasks and roles associated with the data science field.
- Explain how to apply data science to solve problems and improve business operations.
- Describe data science as a combination of different disciplines, such as computer science, mathematics and statistics, domain knowledge, and communications.
- Describe the skills related to mathematics and statistics and the role that they play in solving business problems through analytical modeling.
- Describe the skills related to computer science and the role that they play in solving business problems by supporting analytical models to be trained and deployed effectively.
- Describe the skills related to domain knowledge and the role that they play in solving business problems by adding value to data, models, and tactical and operational actions.
- Describe the skills related to communication and visualization and the role that they play in solving business problems by describing and explaining the solutions, model outcomes, and possible operational actions.

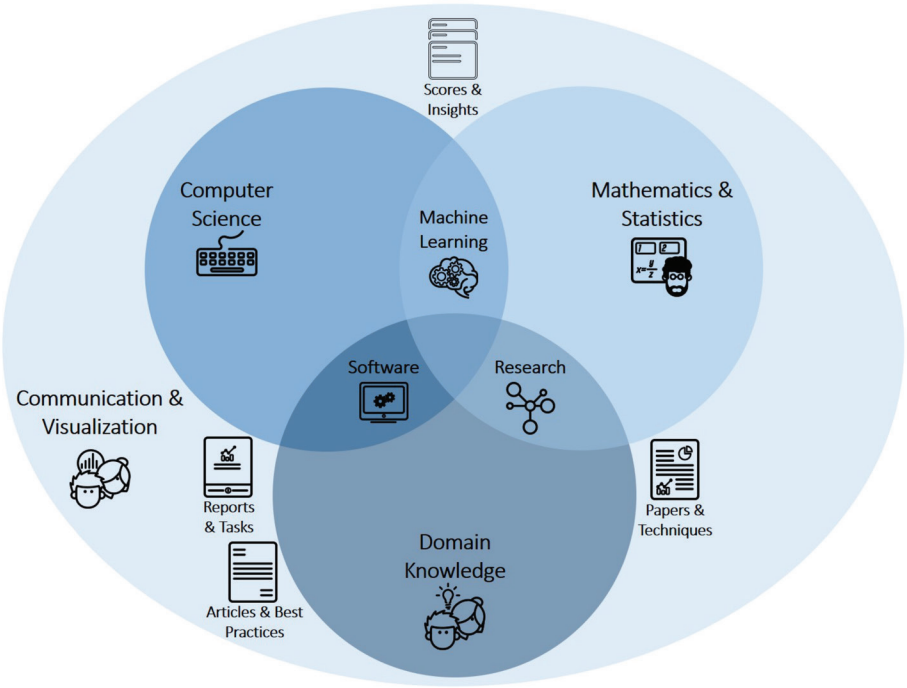
Data Science

Data science is not a single discipline. It comprises a series of different fields of expertise and skills, combining them to solve problems and to improve and optimize processes. Among several skills needed, the most important are mathematics and statistics, computer science, and domain knowledge.

Data scientists need mathematics and statistics to understand the data generated in the business scenario, to model this data to gain insights, or to classify or estimate future events. Mathematics and statistics are also needed to evaluate the models developed and assess how they fit to the problem and how they can be used to solve or improve a specific process.

The infographic in Figure 1.1 refers to three main areas: mathematics and statistics, computer science, and domain knowledge. In the next sections, we will discuss these areas more in depth as well as other important areas such as communication, visualization, and hard and soft skills.

Figure 1.1: Expertise Areas in Data Science



The intersection of each of the three main areas is also very important. *Machine learning* is the field intersecting mathematics, statistics, and computer science. Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, recognize behaviors, and make decisions with minimal human intervention. It is a method of data analysis that automates data preparation, feature engineering, model training, and eventually model deployment. Machine learning allows data scientists to implement very complex models, such as neural networks or support vector machines, and an ensemble of simple models like decision trees, gradient boosting, and random forests. These complex models can capture very unusual relationships between the inputs (independent variables) and the target (dependent variable).

The intersection of mathematics, statistics, and domain knowledge is the research field. Research skills enable data scientists to apply new techniques in model building. This combination allows the development of very complex models that are more accurate and less dependent on the functional form. Research skills can speed up the development process especially when fewer assumptions are needed about the distribution of the target and the relationship of independent and dependent variables.

Software skills in data science usually refer to the intersection of computer science and domain knowledge. Software skills such as familiarity of open-source languages and other world-class software languages help data scientists create new models. The combination of computer science skills, software skills, and domain knowledge can help data scientists solve the business problem or improve a specific business process.

Mathematics and Statistics

Data scientists need to have strong mathematics and statistics skills to understand the data available, prepare the data needed to train a model, deploy multiple approaches in training and validating the analytical model, assess the model's results, and finally explain and interpret the model's outcomes. For example, data scientists need to understand the problem, explain the variability of the target, and conduct controlled tests to evaluate the effect of the values of parameters on the variation of the target values.

Data scientists need mathematics and statistical skills to summarize data to describe past events (known as *descriptive statistics)*. These skills are needed to take the results of a sample and generalize them to a larger population (known as *inferential statistics)*. Data scientists also need these skills to fit models where the response variable is known, and based on that, train a model to classify, predict, or estimate future outcomes (known as *supervised modeling)*. These predictive modeling skills are some of the most widely used skills in data science.

Mathematics and statistics are needed when the business conditions don't require a specific event, and there is no past behavior to drive the training of a supervised model. The learning process is based on discovering previously unknown patterns in the data set (known as *unsupervised modeling)*. There is no target variable and the main goal is to raise some insights to help companies understand customers and business scenarios.

Data scientists need mathematics and statistics in the field of *optimization.* This refers to models aiming to find an optimal solution for a problem when constraints and resources exist. An objective function describes the possible solution, which involves the use of limited resources according to some constraints. Mathematics and statistics are also needed in the field of *forecasting* that is comprised of models to estimate future values in time series data. Based on past values over time, sales, or consumption, it is possible to estimate the future values according to the past behavior. Finally, mathematics and statistics are needed in the field of *econometrics* that applies statistical models to economic data, usually panel data or longitudinal data, to highlight empirical outcomes to economic relationships. These models are used to evaluate and develop econometric methods.

Mathematics and statistics are needed in the field of *text mining.* This is a very important field of analytics, particularly nowadays, because most of the data available is unstructured. Imagine all the social media applications, media content, books, articles, and news. There is a huge amount of information in unstructured, formatted data. Analyzing this type of data allows data scientists to infer correlations about topics, identify possible clusters of contents, search specific terms, and much more. Recognizing the sentiments of customers through text data on social media is a very hot topic called *sentiment analysis*.

Computer Science

The volume of available data today is unprecedented. And most important, the more information about a problem or scenario that is used, the more likely a good model is produced. Due to this data volume, data scientists do not develop models by hand. They need to have computer science skills to develop code, extract, prepare, transform, merge and store data, assess model results, and deploy models in

production. All these steps are performed in digital environments. For example, with a tremendous increase in popularity, cloud-based computing is often used to capture data, create models, and deploy them into production environments.

At some point, data scientists need to know how to create and deploy models into the cloud and use containers or other technologies to allow them to port models to places where they are needed. Think about image recognition models using traffic cameras. It is not possible to capture the data and stream it from the camera to a central repository, train a model, and send it back to the camera to score an image. There are thousands of images being captured every second, and this data transfer would make the solution infeasible. The solution is to train the model based on a sample of data and export the model to the device itself, the camera. As the camera captures the images, the model scores and recognizes the image in real time. All these technologies are important to solve the problem. It is much more than just the analytical models, but it involves a series of processes to capture and process data, train models, generalize solutions, and deploy the results where they need to be. Image recognition models show the usefulness of containers, which packages up software code and all its dependencies so that the application runs quickly and reliably from one computing environment to another.

With today's challenges, data scientists need to have strong computer science skills to deploy the model in different environments, by using distinct frameworks, languages, and storage. Sometimes it is necessary to create programs to capture data or even to expose outcomes. Programming and scripting languages are very important to accomplish these steps. There are several packages that enable data scientists to train supervised and unsupervised models, create forecasting and optimization models, or perform text analytics. New machine learning solutions to develop very complex models are created and released frequently, and to be up to date with new technologies, data scientists need to understand and use these all these new solutions.

Software to collect, prepare, and cleanse data are also very important. Any model is just a mapping of the input data and the event to be analyzed, classified, predicted, or estimated. If the data is poor quality or has a lot of inconsistences, the model will map this, and the outcomes will be inaccurate. Huge amounts of data should be stored in efficient repositories. Databases are needed to accomplish that, and data scientists are required to understand how the databases work. To process massive amounts of data, distributed environments are required. Often, data scientists need to understand how these massive parallel processing engines work. In addition to databases, there are lots of new structures of repositories to perform analytics.

In summary, there are too many skills to be learned and mastered. Much more than one human being can handle. Therefore, in most of the cases, data scientists will need to partner with someone else to perform all the activities needed to create analytical models, such as data engineers, application developers, database administrators, infrastructure engineers, and domain knowledge experts.

Domain Knowledge

One of the most critical skills data scientists need is domain knowledge. Yes, human beings are still important. It is important to understand the problem and evaluate what is necessary to solve it. It is important to understand how to link the model results to a practical action. The analytical model

gives a direction. How to use the results to really solve the problem, based on a series of information, policies, regulations, impacts, and so on, is the key factor to success. Another key factor in creating analytical models is awareness of the business problem and how it affects companies, communities, people, and governments. Knowing the business scenario helps data scientists create new input variables, transform or combine original ones, and select and discard important or useless information. This process called *feature engineering* includes creating new variables based on domain knowledge or based on machine learning. This is discussed in further chapters.

A good example of domain knowledge is in the telecommunications field. Data scientists need to know what type of data are available, how transaction systems are implemented, what billing system is used, and how data from the call centers are collected.

In addition to the domain knowledge, which allows data scientists to improve model development by incorporating business concepts, data scientists must be curious and try out multiple approaches to solve business problems. They need to be proactive in anticipating business issues and propose analytical solutions. There are situations where business issues are not clear, but improvements can be made. Data scientists need to be innovative to design and implement different approaches to solve problems. Creativity and innovation include combining distinct analytical models together to raise insights and to enrich the data analyses.

Collaboration is another key factor in data science. There are many fields of expertise, and it is almost impossible for a single person to master all of them. Collaboration allows data scientists to work alongside different professionals when seeking the best solution for a business problem. For example, if data scientists are developing models to predict loan defaults, then they probably need to partner with someone from the finance department to help them understand how the company charges the customers, what time frame is used to define a default, and how the company deals with defaults. All these policies should be considered when developing data analyses and training analytical models. The data associated with this problem most likely sit in different transactional systems with distinct infrastructures. Data scientists need to work with data engineers, software developers, and informational technology operation personnel to gather all the data sources in an effective way. Once the model is done and deployment is required, data scientists need to work with application developers to make the model's outcomes available to the organization in the way it is expected. The combination of all these skills creates an effective analytical framework to approach business problems in terms of data analysis, model development, and model deployment.

Communication and Visualization

One more key skill is essential to analyze and disseminate the results achieved by data science. At the end of the process, data scientists need to communicate the results. This communication can involve visualizations to explain and interpret the models. A picture is worth a thousand words. Results can be used to create marketing campaigns, offer insights into customer behavior, lead to business decisions and actions, improve processes, avoid fraud, and reduce risk, among many others.

Once the model's results are created, data scientists communicate how the results can be used to improve the operational process with the business side of the company. It is important to provide insights to the decision makers so that they can better address the business problems for which the

model was developed. Every piece of the model's results needs to be assigned to a possible business action. Business departments must understand possible solutions in terms of the model's outcomes and data scientists can fill that gap.

Data scientists use visual presentation expertise and story-telling capabilities to create an exciting and appealing story about how the model's results can be applied to business problems. Data analysis and data visualization sometime suffice. Analyzing the data can help data scientists to understand the problem and the possible solutions but also help to drive straightforward solutions with dashboards and advanced reports. In telecommunications, for example, a drop in services consumption can be associated with an engineering problem rather than a churn behavior. In this case, a deep data analysis can drive the solution rather than a model development to predict churn. This could be a very isolated problem that does not demand a model but instead a very specific business action.

Hard and Soft Skills

Hard skills include mathematics, statistics, computer science, data analysis, programming, etc. On the other side, there are a lot of soft skills essential to performing data science tasks such as problem-solving, communication, curiosity, innovation, storytelling, and so on. It is very hard to find people with both skill sets. Many job search sites point out that there is a reasonable increase in demand for data scientists every year. With substantial inexpensive data storage and increasingly stronger computational power, data scientists have more capacity to fit models that influence business decisions and change the course of tactical and strategical actions. As companies become more data driven, data scientists become more valuable. There is a clear trend that every piece of the business is becoming driven by data analysis and analytical models.

To be effective and valuable in this new evolving scenario, data scientists must have both hard and soft skills. Again, it is quite difficult to find professionals with both hard and soft skills, so collaboration as a team is a very tangible solution. It is critical that data scientists partner with business departments to combine hard and soft skills in seeking the best analytical solution possible.

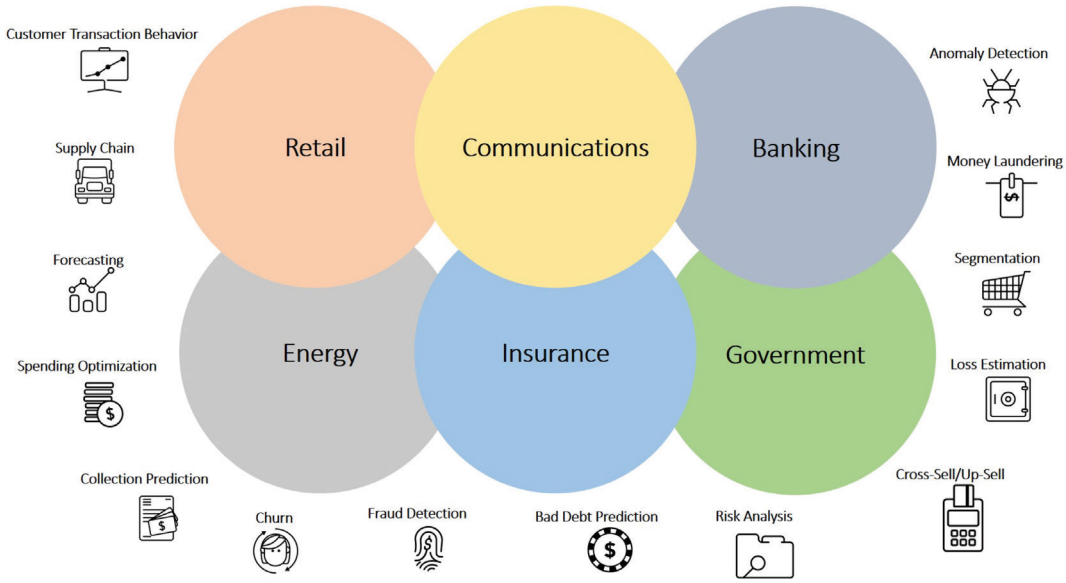
For example, in fraud detection, it is almost mandatory that data scientists collaborate with the fraud analysts and investigators to get their perspective and knowledge in business scenarios where fraud is most prevalent. In this way, they can derive analytical solutions that address feasible solutions in production, usually in a transactional and near real-time perspective.

Data Science Applications

It is difficult to imagine a company or even a segment of industry that cannot benefit from data science and advanced analytics. Today's market demands that all companies, private or public, be more efficient and accurate in their tactical and operational actions. Analytics can help organizations drive business actions based on data facts, instead of guesses, as shown in Figure 1.2.

Another key factor in data science is that all the techniques data scientists use in one industry can be easily applied to another. Business problems, even in different industries, can be remarkably similar. Insolvency is a problem in many industries: telecommunications, banking, and retail, for example. The way banks handle

Figure 1.2: Industries That Can Benefit from Data Science Projects



insolvency can help retail and telecom companies improve their process and be more effective in their business. How telecommunications companies handle transactional fraud can help banks improve their process to detect and react to fraudulent transactions in credit card accounts. Data analysis and analytical insights need to replace guessing and assumptions about market, customers, and business scenarios.

Collaboration among data scientists working in different industries is valuable and increases the spectrum of viable solutions. Industries can learn from each other and improve their process to identify possible analytical solutions and deploy practical business actions. This knowledge transferring from different domain areas, even between distinct industries, is beneficial for all sides involved. Several current business issues include fraud detection, churn analysis, bad debt, loss estimation, cross-sell/up-sell, risk analysis, segmentation, collecting, optimization, forecasting, supply chain, and anomaly detection, among many others.

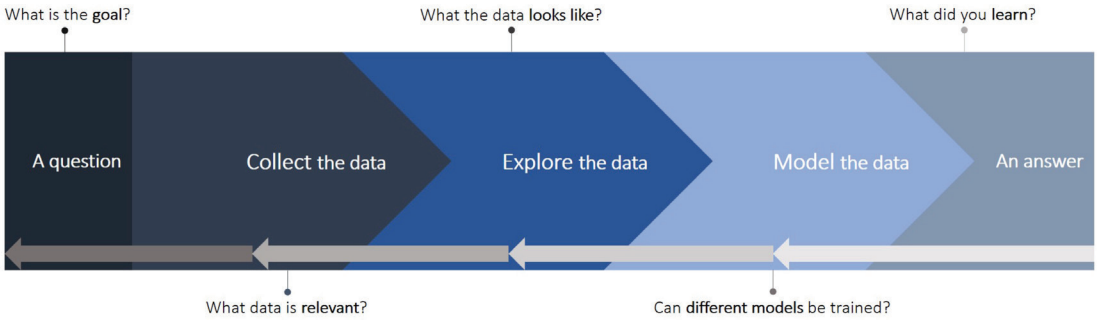
Data Science Lifecycle and the Maturity Framework

The analytical lifecycle, or the data science process flow, comprises a few steps. However, it is important to spend some time on them and make sure all of them are performed properly.

Understand the Question

The first step in the modeling process as shown in Figure 1.3 is to understand the question. Data scientists need to understand what they are trying to solve with the model that they are about to develop. To perform this step, work closely with the business department to verify if a model is

Figure 1.3: Data Science Projects Lifecycle



appropriate, if it is feasible, if there is enough data to use, and what practical actions are planned to be deployed based on the model's outcomes.

Some of the questions to ask during this phase are:

- What is the goal of the project? Do we want to predict some future event or classify, estimate, or forecast? Do we want to optimize a specific process or generate insights about customer behavior? Do we want to create groups or segments or produce a path analysis, sequence analysis, network analysis, and so on?
- What is the specific objective of the model? Is it a supervised or unsupervised model? Is the data structured or unstructured?
- Is there enough data to address this problem?
- What actions are planned based on the model's outcomes?

For example, in a churn model, you need to produce an actionable definition of churn. Is the model a classification (yes or no), and can a campaign be triggered based on the likelihood of churn? Multiple approaches or customer offerings can be assigned to the probability of customer churn. The response can be captured to produce feedback to the model and to improve the performance in subsequent actions.

Collect the Data

The second step in the data science process flow is to collect the data. Most likely, this phase requires multiple people and different skills. Database management, repositories, programming, data quality packages, data integration, and many other technologies might be required to accomplish this step properly.

Some of the pertinent questions during this phase are:

- What data is relevant?
- How many data sources are involved?
- Where do the data sources reside?
- Is the access to the data readily available?

- Are there any privacy issues?
- Are the data available when the model is deployed in production?

For example, some variables, such as gender and income, might not be available to use even though those predictors might be associated with the outcome. Furthermore, even if the data are available when you develop the model, are the data available when the model is in production, such as scoring a business transaction for fraud? Is it possible to access all the data used during the model training when the model is used for scoring? Are there any data privacy regulations? This is a problem because scored observations with missing values for the predictor variables in the scoring model have missing predicted outcomes.

Explore the Data

The third step is to explore the data and evaluate the quality and appropriateness of the information available. This step involves a lot of work with the data. Data analysis, cardinality analysis, data distribution, multivariate analysis, and some data quality analyses–all these tasks are important to verify if all the data needed to develop the model are available, and if they are available in the correct format. For example, in data warehouses, data marts, or data lakes, customer data is stored in multiple occurrences over time, which means that there are multiple records of the same customers in the data set. For analytical models, each customer must be a unique observation in the data set. Therefore, all historical information should be transposed from rows to columns in the analytical table.

Some of the questions for this phase are:

- What anomalies or patterns are noticeable in the data sets?
- Are there too many variables to create the model?
- Are there too few variables to create the model?
- Are data transformations required to adjust the input data for the model training, like imputation, replacement, transformation, and so on?
- Are tasks assigned to create new inputs?
- Are tasks assigned to reduce the number of inputs?

In some projects, data scientists might have thousands of input variables, which is far too many to model in an appropriate way. A variable selection approach should be used to select the relevant features. When there are too few variables to create the model, the data scientist needs to create model predictors from the original input set. Data scientists might also have several input variables with missing values that need to be replaced with reasonable values. Some models require this step, some do not. But even the models that do not require this step might benefit from an imputation process. Sometimes an important input is skewed, and the distribution needs to be adjusted. All these steps can affect the model's performance and accuracy at the end of the process.

Model the Data

The fourth step is the analytical model development itself. Some say that this is the most important part, or at least, where data scientists have more fun. Here they will use their creativity and innovation skills to try out multiple analytical approaches to solve the business problem. As stated before, data

science is a mix of science and art. This step is the time when data scientists apply both the science behind all algorithms and the art behind all analytical approaches.

Some questions to consider at this phase are:

- Which model had the highest predictive accuracy?
- Which model best generalizes to new data?
- Is it possible to validate the model? Is it possible to test the model? Is it possible to honestly test the models on new data?
- Which model is the most interpretable?
- Which model best explains the correlation between the input variables and the target? Which one best describes the effects of the predictors to the estimation?
- Which model best addresses the business goal?

This is the data scientists' playground, where they use different algorithms, techniques, and distinct analytical approaches! Yes, a lot of the modeling process involves simply trying new algorithms and evaluating the results. Data science differs from some exact sciences, like math and physics, where based on a robust equation and inputs, it is possible to predict the output. In data science, the set of inputs might be known, but the exact subset of predictors is still unknown until the end of the model training. The equation is created during the model training according to the input data. Then the results are revealed. Any change in the input data set implies a change in the output. Therefore, data science is very much tied to the statistical and mathematical algorithms. However, all the rest is art. Furthermore, many models are not robust as they should be. Some models or algorithms are very unstable, which means every training data set might represent a different result.

Maybe this is the fun part. In this phase, data scientists try to fit the model on a portion of the data and evaluate the model's performance on another part of the data. The first portion is the training set. The second one is the validation set. Sometimes there is a third portion called the test set. It should be noted that sometimes the best model, depending on the business goal, is the most interpretable and simplest model, rather than the one with the highest predictive accuracy. It depends on the business goal, the practical action, and if there is any regulation in the industry.

In summary, this fourth step includes the following tasks:

- Train different models (algorithms/techniques/analytical approaches).
- Validate all trained models based on a different data set.
- If possible, test all models trained on a different data set (different from the one used during the validation).
- Assess all models' outcomes and evaluate the results based on the business goals.
- Select the best model based on the business requirements.
- Deploy and score the best model to support the business action required.

Perhaps one of the most difficult phases of the analytical process is to analyze the results and evaluate how the model's outcomes can support the business action required. This step is somehow related to the previous one, when data scientists train multiple models and assess the results. At this phase, domain knowledge and communication skills play a key role.

Provide an Answer

The fifth and last step is to provide answers to the original questions, the ones raised and validated during the first step. Some pertinent questions are:

- What lessons were learned from the trained models?
- How do the trained models answer the original questions?
- How do the trained models tell a story and support business decisions?
- How can the trained models be used and deployed in production in the appropriate time frame to support the required business actions?

For example, can the trained model support a campaign that targets customers with the highest probabilities of churn and offer them incentives to keep them using and consuming the company's products and services? Can the trained model support a fraud detection process in real time to identify possible fraudulent business transactions? The model's results might be very accurate, but to benefit the organization, the model should be deployed in an appropriate time frame. For example, in cybersecurity, if the model does not generate real-time alerts in a way that fraud analysts can take immediate actions, then the model might be useless, since digital attacks must be identified within seconds, not weeks or months.

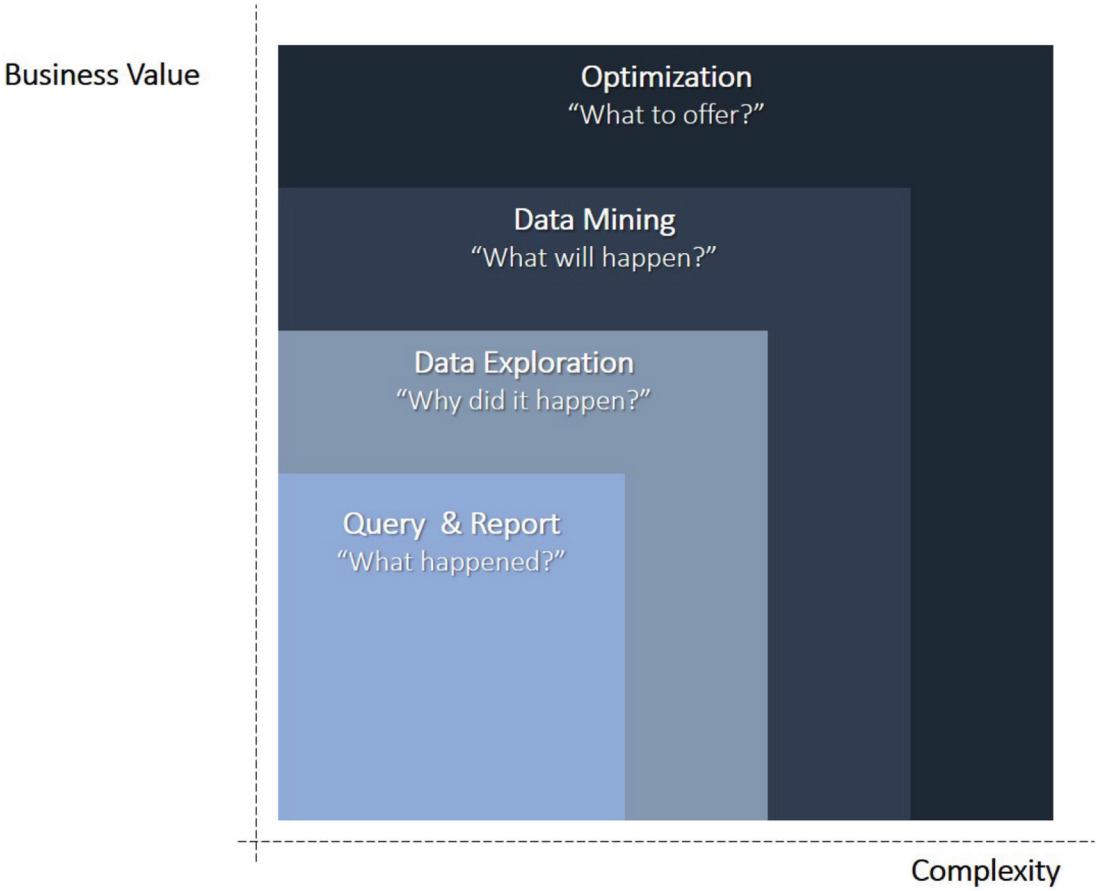
Once an answer is provided, it might generate more questions regarding the business problem. Therefore, the data science lifecycle is cyclical as the process is repeated until the business problem is solved.

The entire analytical process and the data science approach can be viewed as a dynamically evolving flow as shown in Figure 1.4. In data science, the more complex the analytical task, the more value added to the business. For example, a simple query report can add value to the business by simply illustrating the relationships in the data, showing what happened in the past. It is very much descriptive in the sense that nothing can be done to change that historical event. However, awareness is the first step to understand the business problem and aim for an analytical solution.

Data exploration analyses can add further value to the business with more complex queries to the data. Multi-dimensional queries can help business analysts to not only understand what happened, but why it happened in that way. Analyzing the historical data under multiple dimensions at the same time can answer many questions about the business, the market, and the scenarios. Data mining, analytics, or data science, regardless of the name, it is a further step to gain knowledge about the business. Some analytical models explain what is going on right now. Unsupervised models such as clustering, segmentation, association analysis, path analysis, and link analysis help business analysts understand what exactly happens in a very short time frame and allow companies to deploy business actions to take advantage of this knowledge. Furthermore, supervised models can learn from past events and predict and estimate future occurrences. Data science in this phase is basically trying to know what will happen in the future. This is very similar to econometric and forecast models trying to foresee what will happen soon with a business event.

The final stage in the evolving analytical process is optimization. Optimization algorithms add more business value by showing what specific offer to make each customer. Optimization models consider an objective function (what to solve), a limited set of resources (how to solve), and a set of constraints

Figure 1.4: Model's Evolution in Advanced Analytics



(solve it at what price). An organization might have several models in production to classify customers in multiple aspects, for example, the likelihood to make churn, the probability to not pay, etc. A combination of all these scores can be used to optimize campaigns and offerings. For example, the churn model predicts who is the customer most likely to make churn. However, not all customers have the same value to the business. Some might be insolvent. Some might not generate any profit. Some can be very valuable. The optimization process shows what incentives to offer to certain customers to maximize the profit in a specific campaign.

Advanced Analytics in Data Science

Data science and advanced analytics comprise more than simply statistical analysis and mathematical models. The field encompasses machine learning, forecasting, text analytics, and optimization. Data scientists must use all these techniques to solve business problems. In several business scenarios, a combination of some of these models are required to propose a feasible solution to a specific problem.

There are basically two types of machine learning models: *supervised learning* (when the response variable (also known as the *target*) is known and used in the model) and *unsupervised learning* (when the target is unknown or not used in the model). The *input variables* (also called features in the machine learning field or the independent attributes in the statistical field) contain information about the customers, who they are, how they consume the product or service, how they pay for it, for how long they are customers, from where they came from, where they went to, among many other descriptive information.

The target is the business event of interest, for example, when a customer makes churn, purchases a product, makes a payment, or simply uses a credit card or makes a phone call. This event is called a target because this is the event the model will try to predict, classify, or estimate. This is what the company wants to know. (A target is also called a *label* in the machine learning field or *dependent attribute* in the statistical field.) Unsupervised models do not require the target. These models are used to generate insights about the data, or market, or customers, to evaluate possible trends or to better understand some specific business scenarios. These models do not aim to classify, predict, or estimate a business event in the future.

As shown in Table 1.1, regression, decision tree, random forest, gradient boosting, neural network, and support vector machine are examples of supervised models. Clustering, association rules, sequence association rules, path analysis, and link analysis are examples of unsupervised models. There is a variation of these types of models called semi-supervised models. *Semi-supervised models* involve a small amount of data where the target is known and a large amount of data where the target is unknown. There are also models associated with reinforcement learning, where the algorithm is trained by using a system to reward the step when the model goes in the right direction and to punish the step when the model goes in the wrong direction. Semi-supervised models are becoming more prevalent and are often implemented in artificial intelligence applications. For example, reinforcement learning can be used to train a model to learn and take actions in self-driving cars. During the training, if the car drives safely in the road, the learning step is rewarded because it is going in the right direction. On the other hand, if the car drives off the road, the learning step is punished because the training is going in the wrong direction.

As statistical models try to approximate reality through mathematical formalized methods making predictions about future events, machine learning automates some of the most important steps in analytical models, the learning process. Machine learning models automatically improve the learning steps based on the input data and the objective function.

Table 1.1: Machine Learning Models

Supervised Models	Unsupervised Models
Regression	Clustering
Decision Tree	Association Rules
Random Forest	Sequence Association Rules
Gradient Boosting	Path Analysis
Neural Networks	Link Analysis
Support Vector Machine	

Data scientists should be able to build statistical and machine learning flows to access the available data, prepare the data for supervised and unsupervised modeling, fit different types of analytical models to solve the business problems, evaluate models according to the business requirements, and deploy the *champion model* (the chosen model based on some criteria) and *challenger models* (models that are trained differently such as using different algorithms) in production to support the planned business actions. All analytical models are trained based on data sets that, considering a specific time frame, describe the market scenario and the business problem. The models learn from the input variables and create a generalized map to the target, establishing relationships between the input variables and the target. This map describes the past behavior associated with the target in a specific point in time. As time goes by, customer behavior might change and thus the data that describes that pattern. When this happens, the current model in production drifts because it is based on a past customer behavior that no longer exists. That model needs to be retrained or a new model needs to be developed. This is a very important cycle in analytics, and the field of machine learning can contribute greatly with models that automatically retrain or learn from experience.

There is a great debate about when and how to use machine learning models and statistical models. Usually, machine learning models can be more accurate and perform better in production to support business actions. As a caveat, most of the machine learning models are not easy to interpret or explain. On the other hand, statistical models can generalize better estimations for future events. They are often simpler and easier to interpret and explain. In some industries, usually the strictest regulated ones, an interpretable model is mandatory. Statistical models also make the inputs and their effects on the prediction easier to explain, allowing the business departments to design campaigns and promotions more geared to the customers' behaviors.

Statistical analysis is the science of collecting, exploring, and presenting large amounts of data to discover underlying patterns, behaviors, and trends. Organizations use statistics and data analysis every day to make informed business decisions. As more data are collected every day and the infrastructure to store and process all this data gets cheaper, more data analyses are performed to drive business decisions. Statistical analysis includes *descriptive statistics*, where models summarize the available data to describe past events or previous scenarios. Another statistical analysis field is *inferential statistics*, where models take the results of a sample and generalize and extrapolate them to a larger population. Another field in statistical analysis is *predictive modeling*, in which models provide an estimate about the likelihood of a future outcome. This outcome can be a binary target, a multinomial target, or a continuous target. A binary target is assigned to a classification model, like yes or no. A multinomial target is assigned to a type of classification, but for multiple classes, like high, medium, and low. A continuous target is assigned to an estimation, where the event can be any continuous values, like the loss in a fraud event or the amount of purchase. For example, a credit score model can be used to determine whether a customer will make a future payment on time or not. The credit score model can also classify the range of the risk, such as high risk to default, medium risk to default or low risk to default. The credit score model can finally estimate the value associated to the default.

Finally, there is the field of *prescriptive statistics*, where models quantify the effect of future decisions. This area simulates and evaluates several possible courses of action and allows companies to assess different possible outcomes based on these actions. It is like a what-if type of analysis.

Forecasting describes an observed time series to understand the underlying causes of changes and to predict future values. It involves assumptions about the form of the data and decomposes time

series data into multiple components. *Auto-regressive integrated moving average (ARIMA) models* are forecasting models where the predictions are based on a linear combination of past values, past errors, and current and past values of other time series. Another type of forecasting model is the *causal model*, which forecasts time series data that are influenced by causal factors such as calendar events to describe possible seasonality. Finally, there are modern and complex forecasting models that incorporate time series data whose level, trend, or seasonal components vary with time. They might include hierarchical segments of time series and recurrent neural networks to account for stationary and nonstationary data.

Text analytics is a field associated with uncovering insights from text data, usually combining the power of natural language processing, machine learning, and linguistic rules. Data scientists can use text analytics to analyze unstructured text, extract relevant information, and transform it into useful business intelligence. For example, data scientists can use information retrieval to find topics of an unstructured document, such as using a search engine to find specific information. Sentiment analysis is another field in text analytics and very useful in business. *Sentiment analysis* determines levels of agreement from unstructured data associating the overall information as a positive, negative, or neutral sentiment. Data scientists also use text analytics for topics discovery and clustering, where topics or clusters are revealed from various text documents based on the similarity that they have between them.

Finally, *text categorization* is a technique where a text analytics model labels natural language texts with relevant categories from a predefined set. Domain experts and linguistics interact in this field to create and evaluate the categories.

Survival analysis is a class of statistical methods for which the outcome variable of interest is the time until an event occurs. Time is measured from when an individual first becomes a customer until the event occurs or until the end of the observation interval (the individual then becomes censored). In survival analysis, the basis of the analysis is tenure, or the time at risk for the event. Therefore, it is not just whether the event occurred, but when it occurred.

The goal in survival analysis is to model the distribution of time until an event. The job of the data scientist is to identify the variables that are associated with the time until an event and to predict the time until an event for new or existing customers. Survival analysis can be used in customer retention applications where the outcome is the time until the cancellation of all products and services. Another example is credit risk management applications where the outcome is the time until a loan defaults.

The last topic in the advanced analytics framework is *optimization*. Mathematical optimization is a major component in operations research, industrial engineering, and management science. An optimization model searches for an optimal solution, which considers a set of pre-determined constraints and a limited set of resources.

For example, in production planning, optimization models can determine the best mixes of products to be produced to achieve the highest profit. In pricing decisions, optimization models can determine the optimal price for products based on costs, demands and competitive price information. Finally, in promotional marketing, optimization models can determine the best combination of promotional offers, delivery channels, time for campaigns, and the best set of customers to be contacted to maximize the return of the marketing investment.

Another important area of optimization is network analysis and network optimization, which involve analysis of networks (nodes and links) and analysis of network flow. For example, data scientists can use network optimization to study traffic flows where the nodes are cities, and the thickness of the links indicate traffic flow. This information can be used to plan for road widening projects and to construct new roads.

Data Science Practical Examples

Data science and advanced analytics can help organizations in solving business problems, addressing business challenges, and monetizing information and knowledge. Areas like customer experience, revenue optimization, network analytics, and data monetization are just few examples.

Customer Experience

Customer experience and engagement management allow organizations to gain a deeper understanding of the customer experience and how customers respond to the company's stimulus. Some examples include:

- Enhanced Customer Experience – offer more personalized and relevant customer marketing promotions.
- Insolvency and Default Prediction – monitor expenses, bills, and usage over time within safe ranges allowing customers to be on time in their payment events.
- Churn Prediction and Prevention – forecast customer issues and prevent them from happening. For example, if a utility company forecasts an outage and sends a text to the customers affected by the outage, this might improve the customer experience.
- Next Best Offer – operationalize customer insights by using structured and unstructured data. For example, modeling the customer's purchasing history can predict what the next item is to offer the customer.
- Target Social Influencers – identify and track relationships between customers and target those with the most influence. These influencers within their networks can help companies to avoid churn and increase product adoption.

Revenue Optimization

Revenue optimization includes better forecasts to allow organizations to make better business decisions. Some examples include:

- Product Bundles and Marketing Campaigns – improve business decision-making processes related to product bundles and marketing campaigns.
- Revenue Leakage – identify situations leading to revenue leakage, whether due to billing and collections, network, or fraud issues. For example, a collection agency might be interested in who is more likely to pay their debts rather than who owes the most money.
- Personalize products and services – personalize packages, bundles, products, and services according to customers' usages over time and allow them to pay for what they consume.
- Rate Plans and Bundles – use advanced analytics to create new and more effective rate plans and bundles.

Network Analytics

Network analytics is very common in communication and utility industries, but it can be applied in almost any type of industry, in the sense that most complex problems can be viewed as a network problem. Network analytics aims to improve network performance. Network performance can refer to a communications network, energy network, computer systems network, political network, or supply chain network, among many others. Some examples include:

- Network Capacity Planning – use statistical forecasting and detailed network or supply chain data to accurately plan capacity. Use forecasting on ATM machines to make sure there is enough cash to meet customer demands or network analytics to make sure customers have a good mobile signal wherever they go.
- Service Assurance and Optimization – use network analytics to prevent network or supply chain problems before they happen, either in communications, utilities, or retail.
- Optimize supply chain – identify the best routes (cheapest, fastest, or shortest) to deliver goods across geographic locations to keep customers consuming products and services in a continuous fashion.
- Unstructured Data – use unstructured data for deeper customer and service performance insights. For example, optimizing call center staffing and identifying operational changes could lower the cost to serve the customers while improving service quality.

Data Monetization

Information is the new gold. Data science can be used for organizations to monetize the most important asset that they have, which is data. *Data monetization* refers to the act of generating measurable economic benefits from available data sources. Companies in different industries have very sensitive and important data about customers or people in general. Think about telecommunications companies that have precise information about where people go at any point in time. Where people are in a point in time can be easily monetized. Data scientists can use the results of mobile apps that track what customers do, when, and with who. They can also use the results of web search engines that know exactly what customers are looking for in a point in time. Some examples of data monetization include:

- Location-Based Marketing – develop specialized offers and promotions that are delivered to targeted customers via their mobile devices. For example, if you enter a specific area in a city, you might get a coupon for companies located in that area.
- Micro Segmentation – create highly detailed customer segments that can be used to send very specific campaigns, promotions, and offerings over time.
- Third-party Partnerships – partner with different companies to combine customer data to enrich the information used to create analytical models and data analyses about business actions.
- Real-Time Data Analysis – analyze real-time data streams from different types of transactions to keep customers consuming or using products and services with no outages or intermittent breaks.

In conclusion, data scientists can develop and deploy a set of techniques and algorithms to address business problems. Today, companies are dealing with information that comes in varieties and volumes never encountered before. As data scientists increase their skills in areas such as machine learning, statistical analysis, forecasting, text analytics, and optimization, their value to the company will increase over time.

Summary

This chapter introduced some of the most important concepts regarding analytical models for data science. It considers the aspects of hard skills, like computer science, mathematics and statistics, and soft skills, such as domain knowledge and communication. This chapter briefly presented tasks associated with the analytical lifecycle and how some of those data science analytical models can be applied to solve business problems. Different hard and soft skills are crucial in data science projects when trying to solve business problems. Each one of these skills contributes to various phases of the data science project, from preparing and exploring the data, training the analytical models, evaluating model results, and probably the most important phases, deploying the analytical models into production and communicating the model's results to the business departments. This communication is important to help the business areas to define the campaigns, the offerings, and all the strategies to approach customers.

Now that you have an overview of the data science field, the next chapter will discuss data preparation.

Additional Reading

1. Davenport, T.H. and Patil, D.J., "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, October 2012.
2. Flowers, A., "Data Scientist: A Hot Job that Pays Well," Indeed Hiring Lab, January 2019.