CHAPTER 2

Data Science Process

The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process. The standard data science process involves (1) understanding the problem, (2) preparing the data samples, (3) developing the model, (4) applying the model on a dataset to see how the model may work in the real world, and (5) deploying and maintaining the models. Over the years of evolution of data science practices, different frameworks for the process have been put forward by various academic and commercial bodies. The framework put forward in this chapter is synthesized from a few data science frameworks and is explained using a simple example dataset. This chapter serves as a high-level roadmap to building deployable data science models, and discusses the challenges faced in each step and the pitfalls to avoid.

One of the most popular data science process frameworks is Cross Industry Standard Process for Data Mining (CRISP-DM), which is an acronym for Cross Industry Standard Process for Data Mining. This framework was developed by a consortium of companies involved in data mining (Chapman et al., 2000). The CRISP-DM process is the most widely adopted framework for developing data science solutions. Fig. 2.1 provides a visual overview of the CRISP-DM framework. Other data science frameworks are SEMMA, an acronym for Sample, Explore, Modify, Model, and Assess, developed by the SAS Institute (SAS Institute, 2013); DMAIC, is an acronym for Define, Measure, Analyze, Improve, and Control, used in Six Sigma practice (Kubiak & Benbow, 2005); and the Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework used in the knowledge discovery in databases process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). All these frameworks exhibit common characteristics, and hence, a generic framework closely resembling the CRISP process will be used. As with any process framework, a data science process recommends the execution of a certain set of tasks to achieve optimal output. However, the process of extracting information and knowledge from the data is *iterative*. The steps within the data science process are not linear and have to undergo many

Data Science. DOI: https://doi.org/10.1016/B978-0-12-814761-0.00002-2

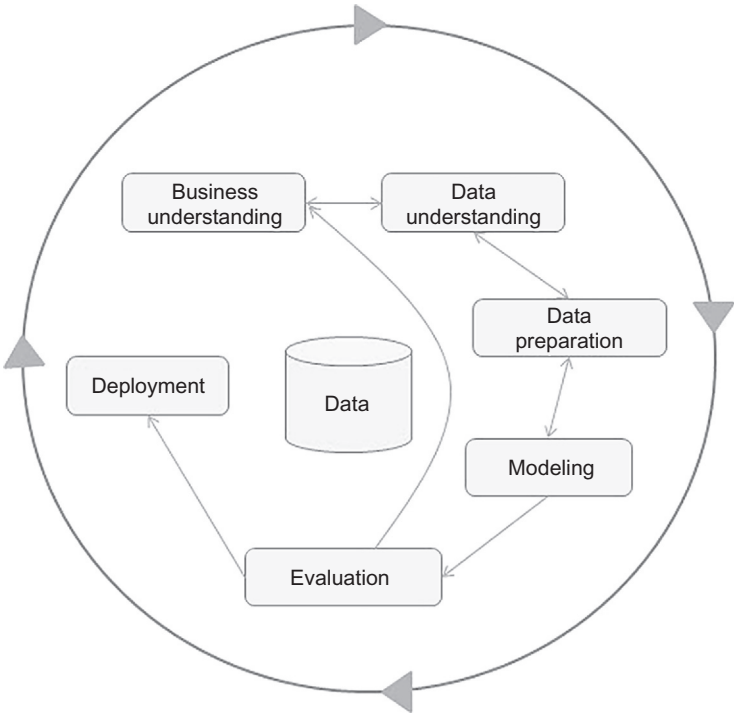

FIGURE 2.1
CRISP data mining framework.

loops, go back and forth between steps, and at times go back to the first step to redefine the data science problem statement.

The data science process presented in Fig. 2.2 is a generic set of steps that is problem, algorithm, and, data science tool agnostic. The fundamental objective of any process that involves data science is to address the analysis question. The problem at hand could be a segmentation of customers, a prediction of climate patterns, or a simple data exploration. The learning algorithm used to solve the business question could be a decision tree, an artificial neural network, or a scatterplot. The software tool to develop and implement the data science algorithm used could be custom coding, RapidMiner, R, Weka, SAS, Oracle Data Miner, Python, etc., (Piatetsky, 2018) to mention a few.

Data science, specifically in the context of big data, has gained importance in the last few years. Perhaps the most visible and discussed part of data science is the third step: *modeling*. It is the process of building representative models that can be inferred from the sample dataset which can be used for either predicting (*predictive modeling*) or describing the underlying pattern in the data (*descriptive or explanatory modeling*). Rightfully so, there is plenty of

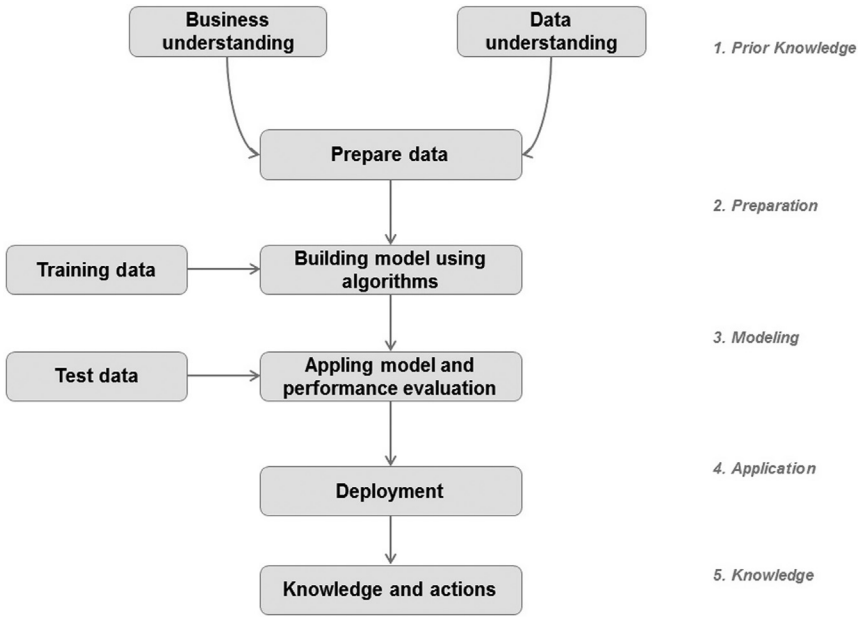

FIGURE 2.2
Data science process.

academic and business research in the modeling step. Most of this book has been dedicated to discussing various algorithms and the quantitative foundations that go with it. However, emphasis should be placed on considering data science as an end-to-end, multi-step, iterative process instead of just a model building step. Seasoned data science practitioners can attest to the fact that the most time-consuming part of the overall data science process is not the model building part, but the preparation of data, followed by data and business understanding. There are many data science tools, both open source and commercial, available on the market that can automate the model building. Asking the right business question, gaining in-depth business understanding, sourcing and preparing the data for the data science task, mitigating implementation considerations, integrating the model into the business process, and, most useful of all, gaining knowledge from the dataset, remain crucial to the success of the data science process. It's time to get started with Step 1: Framing the data science question and understanding the context.

2.1 PRIOR KNOWLEDGE

Prior knowledge refers to information that is already known about a subject. The data science problem doesn't emerge in isolation; it always develops on

top of existing subject matter and contextual information that is already known. The prior knowledge step in the data science process helps to define what problem is being solved, how it fits in the business context, and what data is needed in order to solve the problem.

2.1.1 Objective

The data science process starts with a need for analysis, a question, or a business objective. This is possibly the most important step in the data science process (Shearer, 2000). Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm. As an iterative process, it is common to go back to previous data science process steps, revise the assumptions, approach, and tactics. However, it is imperative to get the first step—the objective of the whole process—right.

The data science process is going to be explained using a hypothetical use case. Take the consumer loan business for example, where a loan is provisioned for individuals against the collateral of assets like a home or car, that is, a mortgage or an auto loan. As many homeowners know, an important component of the loan, for the borrower and the lender, is the interest rate at which the borrower repays the loan on top of the principal. The interest rate on a loan depends on a gamut of variables like the current federal funds rate as determined by the central bank, borrower's credit score, income level, home value, initial down payment amount, current assets and liabilities of the borrower, etc. The key factor here is whether the lender sees enough reward (interest on the loan) against the risk of losing the principal (borrower's default on the loan). In an individual case, the status of default of a loan is Boolean; either one defaults or not, during the period of the loan. But, in a group of tens of thousands of borrowers, the default rate can be found—a continuous numeric variable that indicates the percentage of borrowers who default on their loans. All the variables related to the borrower like credit score, income, current liabilities, etc., are used to assess the default risk in a related group; based on this, the interest rate is determined for a loan. The business objective of this hypothetical case is: *If the interest rate of past borrowers with a range of credit scores is known, can the interest rate for a new borrower be predicted?*

2.1.2 Subject Area

The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes. But the problem is that it uncovers a lot of patterns. The false or spurious signals are a major problem in the data science process. It is up to the practitioner to sift through the exposed patterns and accept the ones that are valid and relevant to the answer of the objective

question. Hence, it is essential to know the subject matter, the context, and the business process generating the data.

The lending business is one of the oldest, most prevalent, and complex of all the businesses. If the objective is to predict the lending interest rate, then it is important to know how the lending business works, why the prediction matters, what happens after the rate is predicted, what data points can be collected from borrowers, what data points cannot be collected because of the external regulations and the internal policies, what other external factors can affect the interest rate, how to verify the validity of the outcome, and so forth. Understanding current models and business practices lays the foundation and establishes known knowledge. Analysis and mining the data provides the new knowledge that can be built on top of the existing knowledge (Lidwell, Holden, & Butler, 2010).

2.1.3 Data

Similar to the prior knowledge in the subject area, prior knowledge in the data can also be gathered. Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process. This part of the step surveys all the data available to answer the business question and narrows down the new data that need to be sourced. There are quite a range of factors to consider: quality of the data, quantity of data, availability of data, gaps in data, does lack of data compel the practitioner to change the business question, etc. The objective of this step is to come up with a dataset to answer the business question through the data science process. It is critical to recognize that an inferred model is only as good as the data used to create it.

For the lending example, a sample dataset of ten data points with three attributes has been put together: identifier, credit score, and interest rate. First, some of the terminology used in the data science process are discussed.

- A *dataset* (*example set*) is a collection of data with a defined structure. Table 2.1 shows a dataset. It has a well-defined structure with 10 rows and 3 columns along with the column headers. This structure is also sometimes referred to as a "data frame".
- A *data point* (*record*, *object* or *example*) is a single instance in the dataset. Each row in Table 2.1 is a data point. Each instance contains the same structure as the dataset.
- An *attribute* (*feature*, *input*, *dimension*, *variable*, or *predictor*) is a single property of the dataset. Each column in Table 2.1 is an attribute. Attributes can be numeric, categorical, date-time, text, or Boolean *data types*. In this example, both the credit score and the interest rate are numeric attributes.

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

- A *label* (*class label*, *output*, *prediction*, *target*, or *response*) is the special attribute to be predicted based on all the input attributes. In Table 2.1, the interest rate is the output variable.
- *Identifiers* are special attributes that are used for locating or providing context to individual records. For example, common attributes like names, account numbers, and employee ID numbers are identifier attributes. Identifiers are often used as lookup keys to join multiple datasets. They bear no information that is suitable for building data science models and should, thus, be excluded for the actual modeling step. In Table 2.1, the attribute ID is the identifier.

2.1.4 Causation Versus Correlation

Suppose the business question is inverted: *Based on the data in* Table 2.1, *can the credit score of the borrower be predicted based on interest rate?* The answer is yes—but it does not make business sense. From the existing domain expertise, it is known that credit score *influences* the loan interest rate. Predicting credit score based on interest rate inverses the direction of the causal relationship. This question also exposes one of the key aspects of model building. The correlation between the input and output attributes doesn't guarantee causation. Hence, it is important to frame the data science question correctly using the existing domain and data knowledge. In this data science example, the interest rate of the new borrower with an unknown interest rate will be predicted (Table 2.2) based on the pattern learned from known data in Table 2.1.

2.2 DATA PREPARATION

Preparing the dataset to suit a data science task is the most time-consuming part of the process. It is extremely rare that datasets are available in the form required by the data science algorithms. Most of the data science algorithms would require data to be structured in a tabular format with records in the rows and attributes in the columns. If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.

2.2.1 Data Exploration

Data preparation starts with an in-depth exploration of the data and gaining a better understanding of the dataset. Data exploration, also known as *exploratory data analysis*, provides a set of simple tools to achieve basic understanding of the data. Data exploration approaches involve computing descriptive statistics and visualization of data. They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset. Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data. On the other hand, a visual plot of data points provides an instant grasp of all the data points condensed into one chart. Fig. 2.3 shows the scatterplot of credit score vs. loan interest rate and it can be observed that as credit score increases, interest rate decreases.

2.2.2 Data Quality

Data quality is an ongoing concern wherever data is collected, processed, and stored. In the interest rate dataset (Table 2.1), how does one know if the credit score and interest rate data are accurate? What if a credit score has a recorded value of 900 (beyond the theoretical limit) or if there was a data entry error? Errors in data will impact the representativeness of the model. Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called *data warehouses*. Data sourced from well-maintained data warehouses have higher quality, as there are proper controls in place to ensure a level of data accuracy for new and existing data. The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc. Regardless, it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models.

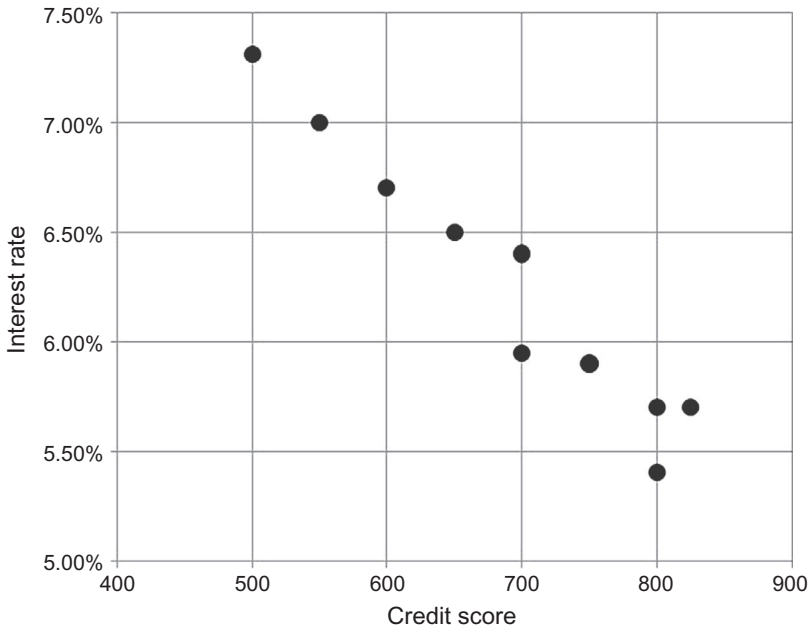

FIGURE 2.3
Scatterplot for interest rate dataset.

2.2.3 Missing Values

One of the most common data quality issues is that some records have missing attribute values. For example, a credit score may be missing in one of the records. There are several different mitigation methods to deal with this problem, but each method has pros and cons. The first step of managing missing values is to understand the reason behind why the values are missing. Tracking the data lineage (provenance) of the data source can lead to the identification of systemic issues during data capture or errors in data transformation. Knowing the source of a missing value will often guide which mitigation methodology to use. The missing value can be substituted with a range of artificial data so that the issue can be managed with marginal impact on the later steps in the data science process. Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute). This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare. Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset. Some data science algorithms are good at handling records with missing values, while others expect the data preparation step to handle it before the model is

inferred. For example, k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values. Neural network models for classification tasks do not perform well with missing attributes, and thus, the data preparation step is essential for developing neural network models.

2.2.4 Data Types and Conversion

The attributes in a dataset can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical. For example, the credit score can be expressed as categorical values (poor, good, excellent) or numeric score. Different data science algorithms impose different restrictions on the attribute data types. In case of linear regression models, the input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute. A specific numeric score can be encoded for each category value, such as poor = 400, good = 600, excellent = 700, etc. Similarly, numeric values can be converted to categorical data types by a technique called *binning*, where a range of values are specified for each category, for example, a score between 400 and 500 can be encoded as "low" and so on.

2.2.5 Transformation

In some data science algorithms like k-NN, the input attributes are expected to be numeric and *normalized*, because the algorithm compares the values of different attributes and calculates distance between the data points. Normalization prevents one attribute dominating the distance results because of large values. For example, consider income (expressed in USD, in thousands) and credit score (in hundreds). The distance calculation will always be dominated by slight variations in income. One solution is to convert the range of income and credit score to a more uniform scale from 0 to 1 by normalization. This way, a consistent comparison can be made between the two different attributes with different units.

2.2.6 Outliers

Outliers are anomalies in a given dataset. Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m). Regardless, the presence of outliers needs to be understood and will require special treatments. The purpose of creating a representative model is to generalize a pattern or a relationship within a dataset and the presence of outliers skews the representativeness of the inferred model. Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

2.2.7 Feature Selection

The example dataset shown in Table 2.1 has one *attribute* or *feature*—the credit score—and one *label*—the interest rate. In practice, many data science problems involve a dataset with hundreds to thousands of attributes. In text mining applications, every distinct word in a document forms a distinct attribute in the dataset. Not all the attributes are equally important or useful in predicting the target. The presence of some attributes might be counterproductive. Some of the attributes may be highly correlated with each other, like annual income and taxes paid. A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model due to the *curse of dimensionality*. In general, the presence of more detailed information is desired in data science because discovering nuggets of a pattern in the data is one of the attractions of using data science techniques. But, as the number of dimensions in the data increase, data becomes sparse in high-dimensional space. This condition degrades the reliability of the models, especially in the case of clustering and classification (Tan, Steinbach, & Kumar, 2005).

Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection. It leads to a more simplified model and helps to synthesize a more effective explanation of the model.

2.2.8 Data Sampling

Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis or modeling. The sample data serve as a representative of the original dataset with similar properties, such as a similar mean. Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling. In most cases, to gain insights, extract the information, and to build representative predictive models it is sufficient to work with samples. Theoretically, the error introduced by sampling impacts the relevancy of the model, but their benefits far outweigh the risks.

In the build process for data science applications, it is necessary to segment the datasets into training and test samples. The training dataset is sampled from the original dataset using simple sampling or class label specific sampling. Consider the example cases for predicting anomalies in a dataset (e.g., predicting fraudulent credit card transactions). The objective of anomaly detection is to classify the outliers in the data. These are rare events and often the dataset does not have enough examples of the outlier class. *Stratified sampling* is a process of sampling where each class is equally represented in the sample; this allows the model to focus on the difference between the patterns of each class that is, normal and outlier records. In classification applications,

sampling is used create multiple base models, each developed using a different set of sampled training datasets. These base models are used to build one meta model, called the *ensemble model*, where the error rate is improved when compared to that of the base models.

2.3 MODELING

A model is the abstract representation of the data and the relationships in a given dataset. A simple rule of thumb like "*mortgage interest rate reduces with increase in credit score*" is a model; although there is not enough quantitative information to use in a production scenario, it provides directional information by abstracting the relationship between credit score and interest rate.

There are a few hundred data science algorithms in use today, derived from statistics, machine learning, pattern recognition, and the body of knowledge related to computer science. Fortunately, there are many viable commercial and open source data science tools on the market to automate the execution of these learning algorithms. As a data science practitioner, it is sufficient to an overview of the learning algorithm, how it works, and determining what parameters need to be configured based on the understanding of the business and data. Classification and regression tasks are predictive techniques because they predict an outcome variable based on one or more input variables. Predictive algorithms require a prior known dataset to learn the model. Fig. 2.4 shows the steps in the modeling phase of predictive data science. Association analysis and clustering are descriptive data science techniques where there is no target variable to predict; hence, there is no test dataset. However, both predictive and descriptive models have an evaluation step.

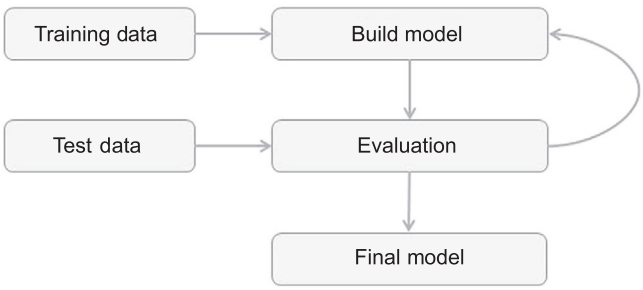

FIGURE 2.4
Modeling steps.

Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

2.3.1 Training and Testing Datasets

The modeling step creates a representative model inferred from the data. The dataset used to create the model, with known attributes and target, is called the *training dataset*. The validity of the created model will also need to be checked with another known dataset called the *test dataset* or *validation dataset*. To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset. A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test dataset. Tables 2.3 and 2.4 show the random split of training and test data, based on the example dataset shown in Table 2.1. Fig. 2.5 shows the scatterplot of the entire example dataset with the training and test datasets marked.

2.3.2 Learning Algorithms

The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used. The practitioner determines the appropriate data science algorithm within the chosen category. For example, within a classification task many algorithms can be chosen from: decision trees, rule induction, neural networks, Bayesian models, k-NN, etc. Likewise, within decision tree techniques, there are quite a number of variations of learning algorithms like classification and regression tree (CART), CHi-squared Automatic Interaction Detector (CHAID) etc.

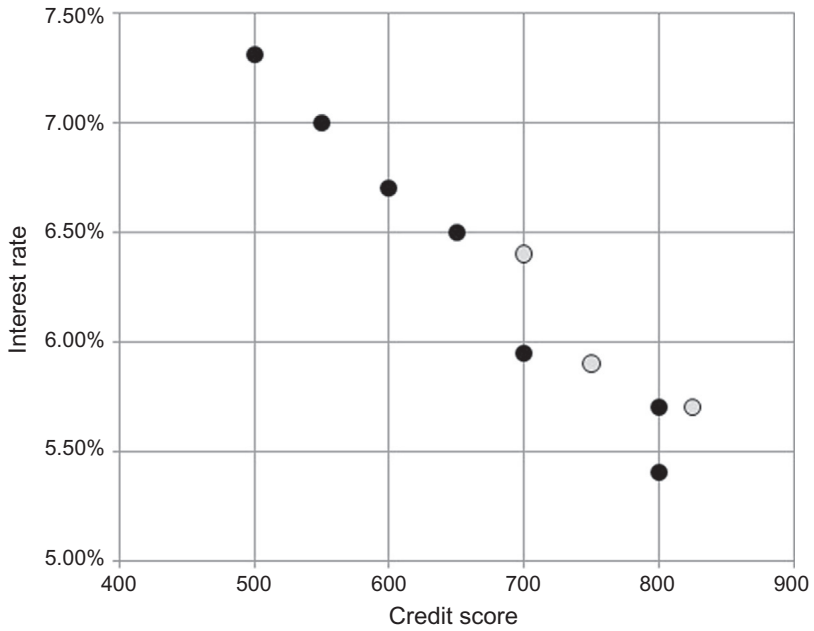

FIGURE 2.5
Scatterplot of training and test data.

These algorithms will be reviewed in detail in later chapters. It is not uncommon to use multiple data science tasks and algorithms to solve a business question.

Interest rate prediction is a regression problem. A simple linear regression technique will be used to model and generalize the relationship between credit score and interest rate. The training set of seven records is used to create the model and the test set of three records is used to evaluate the validity of the model.

The objective of simple linear regression can be visualized as fitting a straight line through the data points in a scatterplot (Fig. 2.6). The line has to be built in such a way that the sum of the squared distance from the data points to the line is minimal. The line can be expressed as:

$$y = a * x + b \tag{2.1}$$

where y is the output or dependent variable, x is the input or independent variable, b is the y-intercept, and a is the coefficient of x. The values of a and b can be found in such a way so as to minimize the sum of the squared residuals of the line.

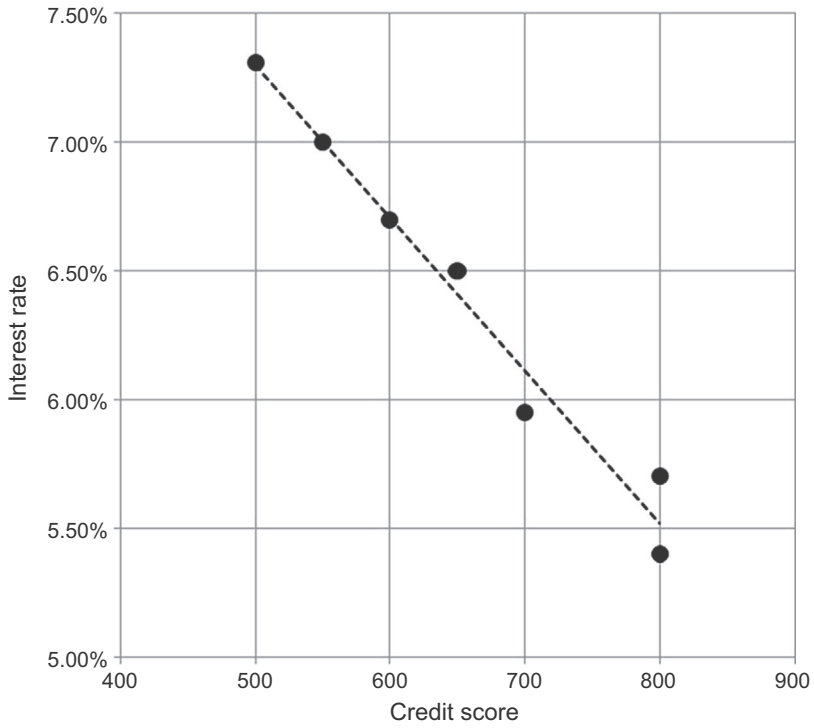

FIGURE 2.6
Regression model.

The line shown in Eq. (2.1) serves as a model to predict the outcome of new unlabeled datasets. For the interest rate dataset, the simple linear regression for the interest rate (y) has been calculated as (details in Chapter 5: Regression):

$$y = 0.1 + \frac{6}{100,000}x$$

$$\text{Interest rate} = 10 - \frac{6 \times \text{credit score}}{1000}$$

Using this model, the interest rate for a new borrower with a specific credit score can be calculated.

2.3.3 Evaluation of the Model

The model generated in the form of an equation is generalized and synthesized from seven training records. The credit score in the equation can be substituted to see if the model estimates the interest rate for each of the

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)	Model Predicted (*Y*) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

seven training records. The estimation may not be exactly the same as the values in the training records. A model should not memorize and output the same values that are in the training records. The phenomenon of a model memorizing the training data is called *overfitting*. An overfitted model just memorizes the training records and will underperform on real unlabeled new data. The model should generalize or *learn* the relationship between credit score and interest rate. To evaluate this relationship, the validation or test dataset, which was not previously used in building the model, is used for evaluation, as shown in Table 2.5.

Table 2.5 provides the three testing records where the value of the interest rate is known; these records were not used to build the model. The actual value of the interest rate can be compared against the predicted value using the model, and thus, the *prediction error* can be calculated. As long as the error is acceptable, this model is ready for deployment. The error rate can be used to compare this model with other models developed using different algorithms like neural networks or Bayesian models, etc.

2.3.4 Ensemble Modeling

Ensemble modeling is a process where multiple diverse base models are used to predict an outcome. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and *independent*, the prediction error decreases when the ensemble approach is used. The approach seeks the wisdom of crowds in making a prediction. Even though the ensemble model has multiple base models within the model, it acts and performs as a single model. Most of the practical data science applications utilize ensemble modeling techniques.

At the end of the modeling stage of the data science process, one has (1) analyzed the business question; (2) sourced the data relevant to answer the question; (3) selected a data science technique to answer the question; (4) picked a data science algorithm and prepared the data to suit the algorithm; (5) split the data into training and test datasets; (6) built a generalized model from the training dataset; and (7) validated the model against the test

dataset. This model can now be used to predict the interest rate of new borrowers by integrating it in the actual loan approval process.

2.4 APPLICATION

Deployment is the stage at which the model becomes production ready or *live*. In business applications, the results of the data science process have to be assimilated into the business process—usually in software applications. The model deployment stage has to deal with: assessing model readiness, technical integration, response time, model maintenance, and assimilation.

2.4.1 Production Readiness

The production readiness part of the deployment determines the critical qualities required for the deployment objective. Consider two business use cases: determining whether a consumer qualifies for a loan and determining the groupings of customers for an enterprise by marketing function.

The consumer credit approval process is a real-time endeavor. Either through a consumer-facing website or through a specialized application for frontline agents, the credit decisions and terms need to be provided in real-time as soon as prospective customers provide the relevant information. It is optimal to provide a quick decision while also proving accurate. The decision-making model has to collect data from the customer, integrate third-party data like credit history, and make a decision on the loan approval and terms in a matter of seconds. The critical quality of this model deployment is real-time prediction.

Segmenting customers based on their relationship with the company is a thoughtful process where signals from various customer interactions are collected. Based on the patterns, similar customers are put in cohorts and campaign strategies are devised to best engage the customer. For this application, batch processed, time lagged data would suffice. The critical quality in this application is the ability to find unique patterns amongst customers, not the response time of the model. The business application informs the choices that need to be made in the data preparation and modeling steps.

2.4.2 Technical Integration

Currently, it is quite common to use data science automation tools or coding using R or Python to develop models. Data science tools save time as they do not require the writing of custom codes to execute the algorithm. This allows the analyst to focus on the data, business logic, and exploring patterns

from the data. The models created by data science tools can be ported to production applications by utilizing the Predictive Model Markup Language (PMML) (Guazzelli, Zeller, Lin, & Williams, 2009) or by invoking data science tools in the production application. PMML provides a portable and consistent format of model description which can be read by most data science tools. This allows the flexibility for practitioners to develop the model with one tool (e.g., RapidMiner) and deploy it in another tool or application. Some models such as simple regression, decision trees, and induction rules for predictive analytics can be incorporated directly into business applications and business intelligence systems easily. These models are represented by simple equations and the "if-then" rule, hence, they can be ported easily to most programming languages.

2.4.3 Response Time

Data science algorithms, like k-NN, are easy to build, but quite slow at predicting the unlabeled records. Algorithms such as the decision tree take time to build but are fast at prediction. There are trade-offs to be made between production responsiveness and modeling build time. The quality of prediction, accessibility of input data, and the response time of the prediction remain the critical quality factors in business application.

2.4.4 Model Refresh

The key criterion for the ongoing relevance of the model is the representativeness of the dataset it is processing. It is quite normal that the conditions in which the model is built change after the model is sent to deployment. For example, the relationship between the credit score and interest rate change frequently based on the prevailing macroeconomic conditions. Hence, the model will have to be refreshed frequently. The validity of the model can be routinely tested by using the new known test dataset and calculating the prediction error rate. If the error rate exceeds a particular threshold, then the model has to be refreshed and redeployed. Creating a maintenance schedule is a key part of a deployment plan that will sustain a relevant model.

2.4.5 Assimilation

In the descriptive data science applications, deploying a model to live systems may not be the end objective. The objective may be to assimilate the knowledge gained from the data science analysis to the organization. For example, the objective may be finding logical clusters in the customer database so that separate marketing approaches can be developed for each customer cluster. Then the next step may be a classification task for new customers to bucket them in one of known clusters. The association analysis

provides a solution for the market basket problem, where the task is to find which two products are purchased together most often. The challenge for the data science practitioner is to articulate these findings, establish relevance to the original business question, quantify the risks in the model, and quantify the business impact. This is indeed a challenging task for data science practitioners. The business user community is an amalgamation of different points of view, different quantitative mindsets, and skill sets. Not everyone is aware about the process of data science and what it can and cannot do. Some aspects of this challenge can be addressed by focusing on the end result, the impact of knowing the discovered information, and the follow-up actions, instead of the technical process of extracting the information through data science.

2.5 KNOWLEDGE

The data science process provides a framework to extract nontrivial information from data. With the advent of massive storage, increased data collection, and advanced computing paradigms, the available datasets to be utilized are only increasing. To extract knowledge from these massive data assets, advanced approaches need to be employed, like data science algorithms, in addition to standard business intelligence reporting or statistical analysis. Though many of these algorithms can provide valuable knowledge, it is up to the practitioner to skillfully transform a business problem to a data problem and apply the right algorithm. Data science, like any other technology, provides various options in terms of algorithms and parameters within the algorithms. Using these options to extract the right information from data is a bit of an art and can be developed with practice.

The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained. As with any quantitative technique, the data science process can bring up spurious irrelevant patterns from the dataset. Not all discovered patterns lead to incremental knowledge. Again, it is up to the practitioner to invalidate the irrelevant patterns and identify the meaningful information. The impact of the information gained through data science can be measured in an application. It is the difference between gaining the information through the data science process and the insights from basic data analysis. Finally, the whole data science process is a framework to invoke the right questions (Chapman et al., 2000) and provide guidance, through the right approaches, to solve a problem. It is not meant to be used as a set of rigid rules, but as a set of iterative, distinct steps that aid in knowledge discovery.

In the upcoming chapters, the details of key data science concepts along with their implementation will be explored. Exploring data using basic statistical and visual techniques are an important first step in preparing the data for data science. The next chapter on data exploration provides a practical tool kit to explore and understand data. The techniques of data preparation are explained in the context of individual data science algorithms in the chapters on classification, association analysis, clustering, text mining, time series, and anomaly detection.

References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. Retrieved from <ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf>.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54.

Guazzelli, A., Zeller, M., Lin, W., & Williams, G. (2009). PMML: An open standard for sharing models. *The R Journal*, *1*(1), 60–65.

Kubiak, T., & Benbow, D. W. (2005). *The certified six sigma black belt handbook*. Milwaukee, WI: ASQQuality Press.

Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Beverly, MA: Rockport Publishers.

Piatetsky, G. (2018). *Top software for analytics, data science, machine learning in 2018: Trends and analysis*. Retrieved from <https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html> Accessed 07.07.18.

SAS Institute. (2013). *Getting started with SAS enterprise miner 12.3*.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. *Journal of School Psychology*, *19*, 51–56. Available from https://doi.org/10.1016/0022-4405(81)90007-8.

Weisstein, E. W. (2013). Retrieved from <http://mathworld.wolfram.com/LeastSquaresFitting.html> *Least squares fitting*. Champaign, Illinois: MathWorld—Wolfram Research, Inc.