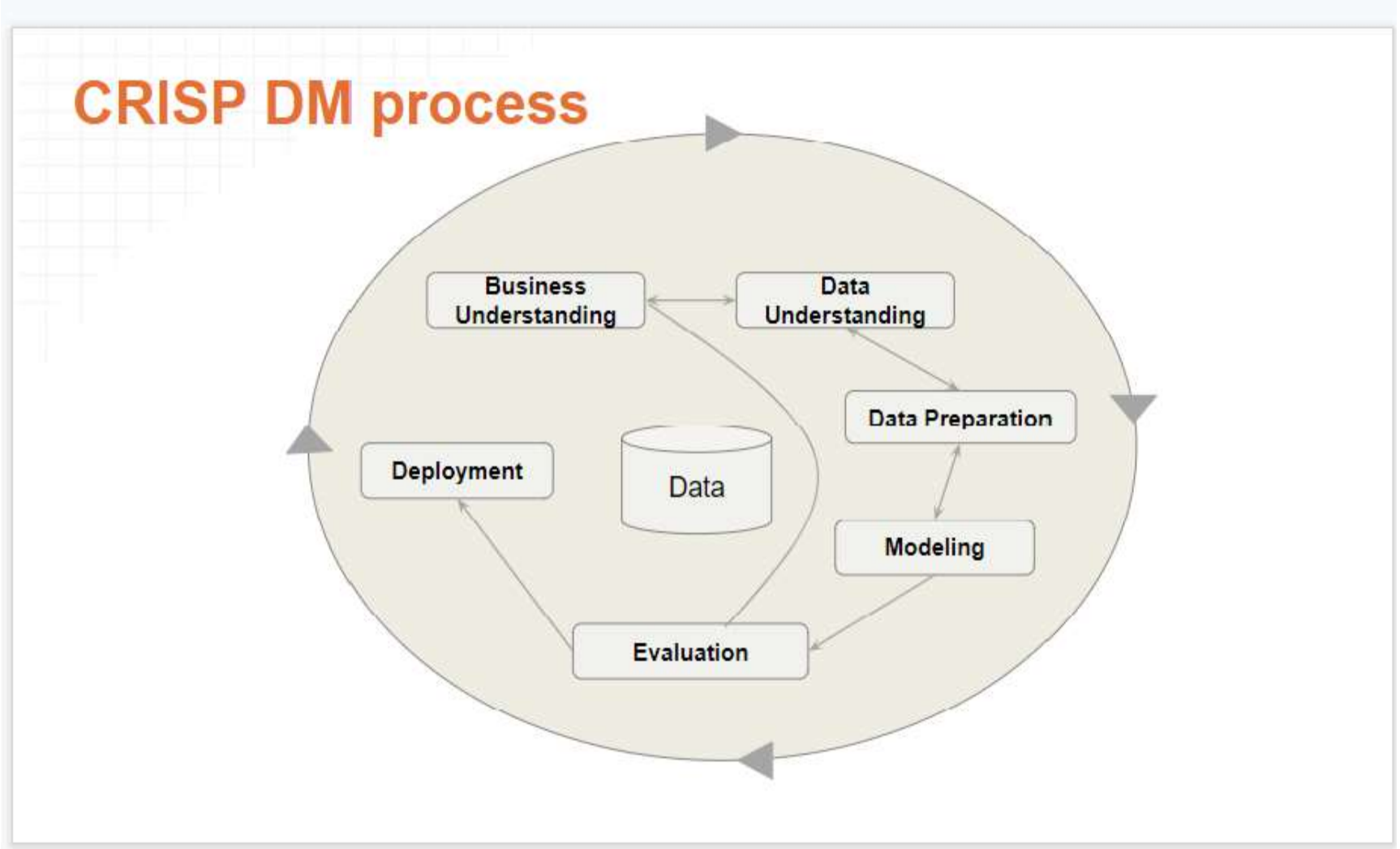
Chapter 2

Data Science Process

CRISP Data Mining Framework



CRISP Data Mining Framework

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six sequential phases:

Business understanding – What does the business need?

Data understanding – What data do we have / need? Is it clean?

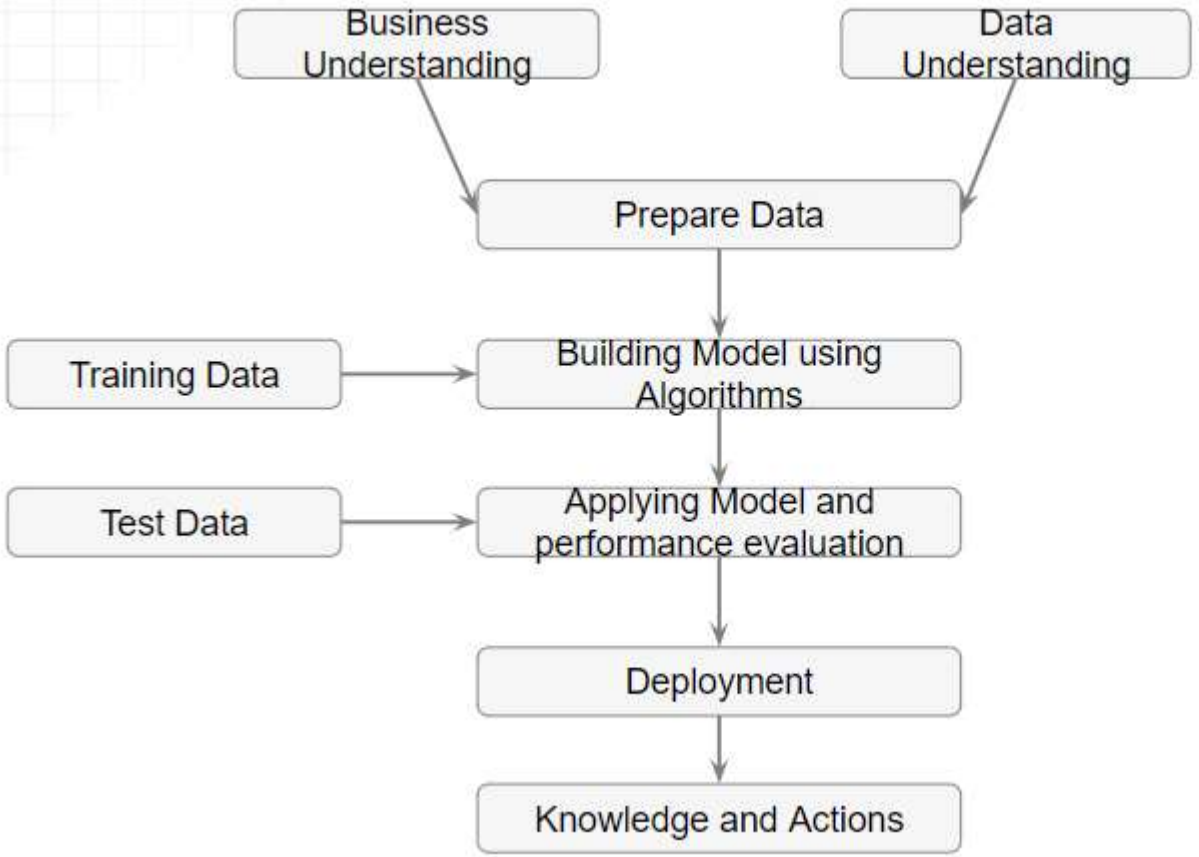
Data preparation – How do we organize the data for modeling?

Modeling – What modeling techniques should we apply?

Evaluation – Which model best meets the business objectives?

Deployment – How do stakeholders access the results?

Data Science Process



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

Prior Knowledge

Gaining information on:

- Objective of the problem
- Subject area of the problem
- Data

Objective of the problem

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. 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.

Example : 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?

Subject area of the problem

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

Example: The lending business is one of the oldest, most prevalent, and complex of al the businesses. If the objective is to predict the lending interest rate, then it is important to know how the lending business works,

Data

Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process. There are quite a range of factors to consider: quality of the data, quantity of data. The objective of this step is to come up with a dataset to answer the business question through the data science process. For the following example, a sample dataset of ten data points with three attributes has been put together: identifier, credit score, and interest rate.

Data Types

- Two types of data: Labelled Data & Unlabelled Data
- **Labelled data**
- Specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen. Data of this kind is called labelled.

Outlook=sunny	Temp=79	Humidity=88	Windy=false	Class=?

Outlook	Temp (°F)	Humidity (%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

Data Types

- **Unlabelled data**
  - Data that does not have any specially designated attribute is called unlabelled.
  - Here the aim is simply to extract the most information we can from the data available.

Age	Gender	Income	Profession	Tenure	City
35	M	60,000	IT	12	KRK
23	F	90,000	Sales	3	WAW
18	M	12,000	Student	1	KRK
42	F	128,000	Doctor	13	KRK
34	M	63,000	Manager	8	WAW
56	M	82,000	Teacher	30	WAW

Learning Methods

- **Supervised Learning**
  - Data mining using labelled data is known as supervised learning.

- **Classification**
    - If the designated attribute is categorical, the task is called classification.
    - Classification is one form of prediction, where the value to be predicted is a label.
    - a hospital may want to classify medical patients into those who are at high, medium or low risk
    - we may wish to classify a student project as distinction, merit, pass or fail
    - Nearest Neighbour Matching, Classification Rules, Classification Tree, …

Learning Methods

- **Numerical Prediction (Regression)**
  - If the designated attribute is numerical, the task is called numerical prediction (regression).
  - Numerical prediction (often called regression) is another. In this case we wish to predict a numerical value, such as a company's profits or a share price.
  - A very popular way of doing this is to use a Neural Network

Learning Methods

- **Unsupervised Learning**
  - Data mining using unlabelled data is known as unsupervised learning.
  - **Association Rules**
    - Sometimes we wish to use a training set to find any relationship that exists amongst the values of variables, generally in the form of rules known as association rules.
    - APRIORI
    - Market Basket Analysis
  - **Clustering**
    - Clustering algorithms examine data to find groups of items that are similar.
    - K-Means Clustering, Agglomerative Hierarchical Clustering

- 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 Data Set

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%

- 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.

Table 2.2 New Data With Unknown Interest Rate

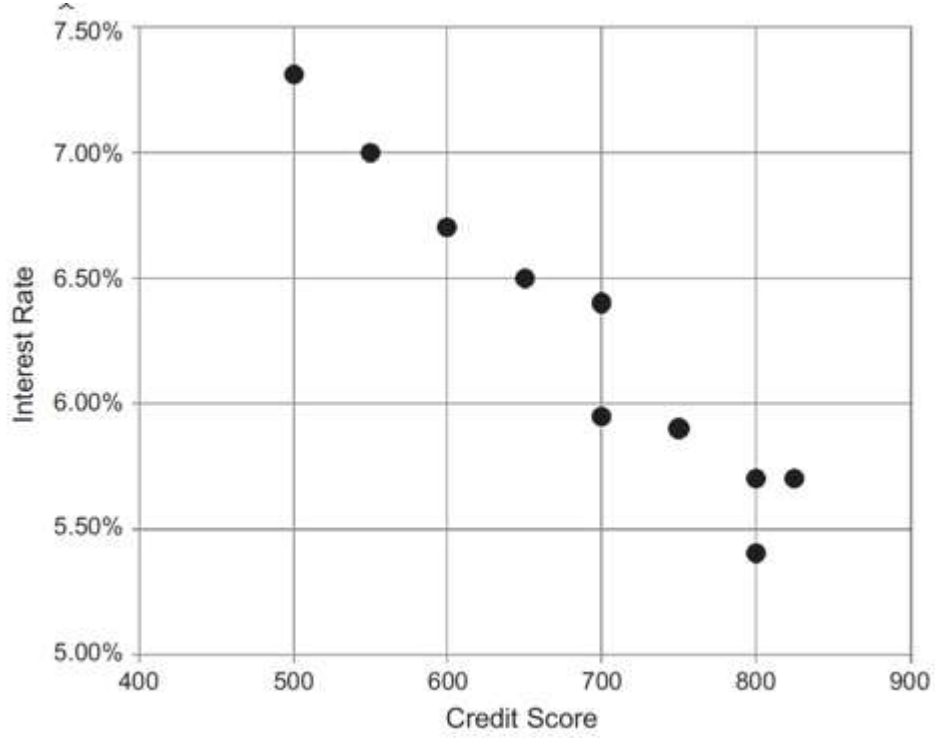
Borrower ID	Credit Score	Interest Rate
11	625	?

Data Preparation

- Data Exploration
- Data quality
- Missing values
- Noisy values
- Invalid values
- Data types and Conversion
- Transformation
- Outliers
- Feature selection
- Sampling

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 statis



Data Quality

Data quality is the measure of how well suited a data set is to serve its specific purpose. Measures of data quality are based Data Correctness, Data Freshness and Data Completeness.

Data Correctness: How accurately the data value describes real-world facts. For example, a B2B sales rep wishes to look at a prospect company's number of employees. If they accidentally grab the wrong company from their database, because its name and location are similar to another organization, they will report a wrong number, be misinformed, and potentially lose an opportunity to sell to a qualified prospect. In this case, the rep used incorrect data. The data correctness metric is usually measured with classification metrics such as precision – a compound of correct data points compared to incorrect data points. There are many potential root causes of collection issues such as collection noise, faulty data transformations, outdated data, or incorrect schema description.

Data Freshness：This refers to how relevant the data is to describe the current state of an entity, and takes into consideration the timeliness of the data and how frequently it is updated. This is a tricky measurement as “freshness” ranges from data that is updated in real-time to data that is updated annually. Each business use case will differ in its data freshness thresholds and requirements. For example, data that doesn’t change frequently, like a person or institution name, would not require the same freshness as stock market or Twitter trends. In any case, data must be up-to-date, if it is not it could mislead a decision. This metric is typically measured with time

Data Completeness： A measure which describes how whole and complete a data asset is. Completeness is especially important when you want to attach new attributes to your existing data. In cases where you have low coverage, you would get limited support for the different attributes that you enrich, and the data becomes less useful. Coverage is also important if you want to extract insights from your dataset.

Missing Values

In many real-world datasets data values are not recorded for all attributes. This can happen simply because there are some attributes that are not applicable for some instances, a malfunction of the equipment used to record the data, a data collection form to which additional fields were added after some data had been collected, information that could not be obtained, e.g. about a hospital patient

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

Methods to Handle Missing Values

- **Discard Instances**
- **Replace by Most Frequent/Average Value**

Discard Instances

- This is the simplest strategy: delete all instances where there is at least one missing value and use the remainder.
- It has the advantage of avoiding introducing any data errors. Its disadvantage is that discarding data may damage the reliability of the results derived from the data

Replace by Most Frequent/Average Value

- A less cautious strategy is to estimate each of the missing values using the values that are present in the dataset.
- A straightforward but effective way of doing this for a categorical attribute is to use its most frequently occurring (non-missing) value
- In the case of continuous attributes, it is likely that no specific numerical value will occur more than a small number of times. In this case the estimate used is generally the average value.

Noisy Values

- A noisy value to mean one that is valid for the dataset, but is incorrectly recorded
- The number 69.72 may accidentally be entered as 6.972, or a categorical attribute value such as brown may accidentally be recorded as another of the possible values, such as blue.

Invalid Values

- 69.7X for 6.972 or bbrown for brown
- An invalid value can easily be detected and either corrected or rejected

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.

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. To overcome this problem ,we generally normalize the values of continuous attributes.

The idea is to make the values of each attribute run from 0 to 1. In general, if the lowest value of attribute A is min and the highest value is max, we convert each value of A, say a, to (a − min)/(max − min).

Mileage (miles)	Number of doors	Age (years)	Number of owners
18,457	2	12	8
26,292	4	3	1

Outliers

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. Detecting outliers may be the primary purpose of some data science applications, like fake email detection, fraud or intrusion detection.

Feature Selection

Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection.

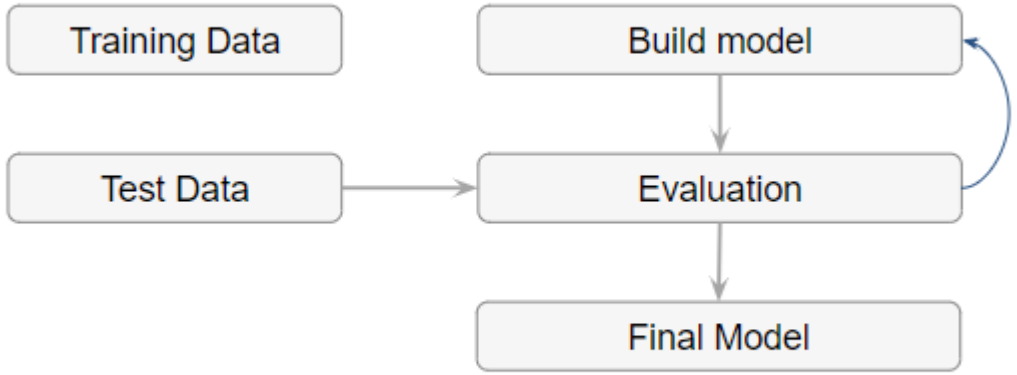
many data science problems involve a dataset with hundreds to thousands of attributes. In text mining applications, every distinct word in a document forms a distinct attribute in the dataset. Not all the attributes are equally important or useful in predicting the target. The presence of some attributes might be counter productive. 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

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.

MODELING

A model is the abstract representation of the data and the relationships in a given dataset.



MODELING

Splitting Training and Test data sets: 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

Training Set and Test Set

Table 2.3 Training Data Set

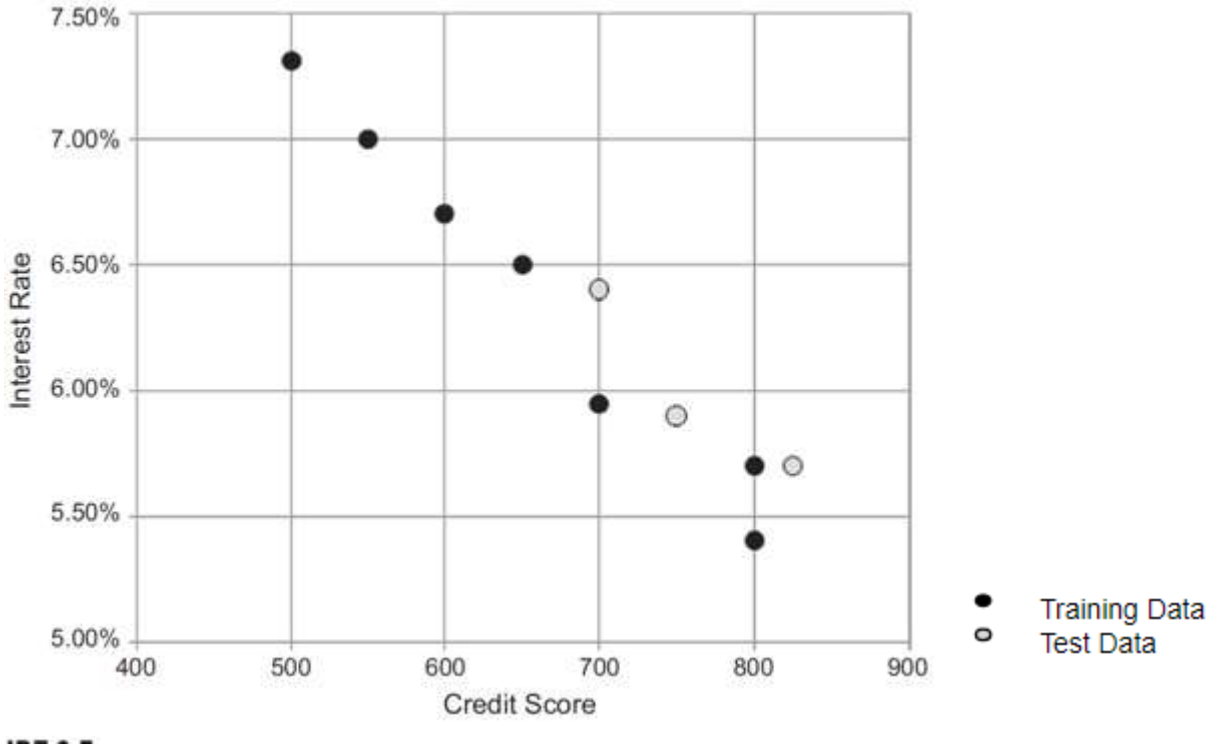
Borrower	Credit Score (X)	Interest Rate (Y)
01	500	7.31%
02	600	6.70%
03	700	5.95%
05	800	5.40%
06	800	5.70%
08	550	7.00%
09	650	6.50%

Table 2.4 Test Data Set

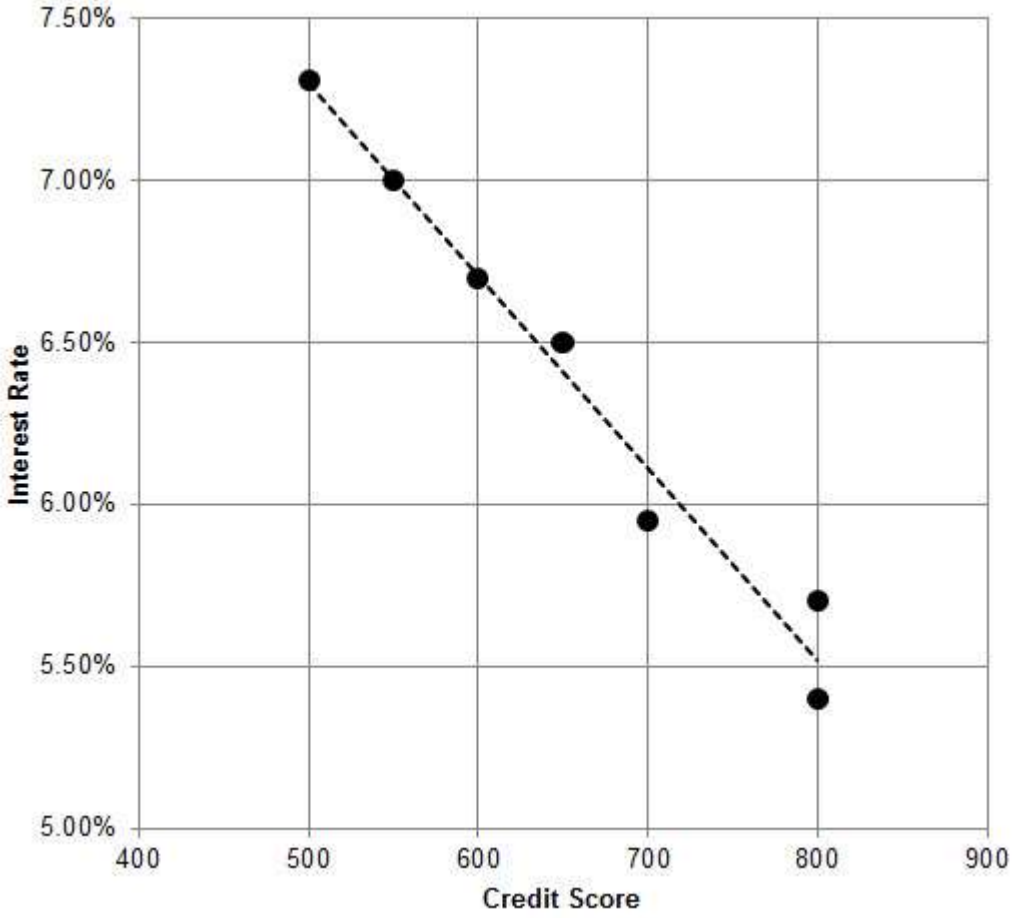
Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40%
07	750	5.90%
10	825	5.70%

MODELING

Splitting Training and Test data sets



MODELING


7.50%
7.00%
6.50%
6.00%
5.50%
5.00%
Interest Rate
400
500
600
700
800
900
Credit Score

MODELING

Evaluation of test dataset

Table 2.5 Evaluation of Test Data Set

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%

Application

Product readiness
Technical integration
Model response time
Remodeling
Assimilation

Knowledge

Posterior knowledge