Data Science: Concepts and Practice

1. Introduction

- 1. What is data science and why do we need data science?
- 2. Data science classification and tasks.
- 3. Various methods to understand data.
- 4. Typical methods to visualize data.

Syllabus

- Introduction to data science - What is data science? Why data science?
- Data science classification
- Data science process - Prior knowledge, Data preparation, Modelling, Application
- Data exploration- Data sets, Descriptive statistics for univariate and multivariate data
- Data visualization – Histogram, Quartile plot, Distribution chart, Scatter plot, Bubble chart, Density chart

Data science

- Data science is a collection of techniques used to extract value from data.
- Data science techniques rely on finding useful patterns, connections, and relationships within data
- The use of the term science in data science indicates that the methods are evidence based, and are built on empirical knowledge, more specifically historical observations.

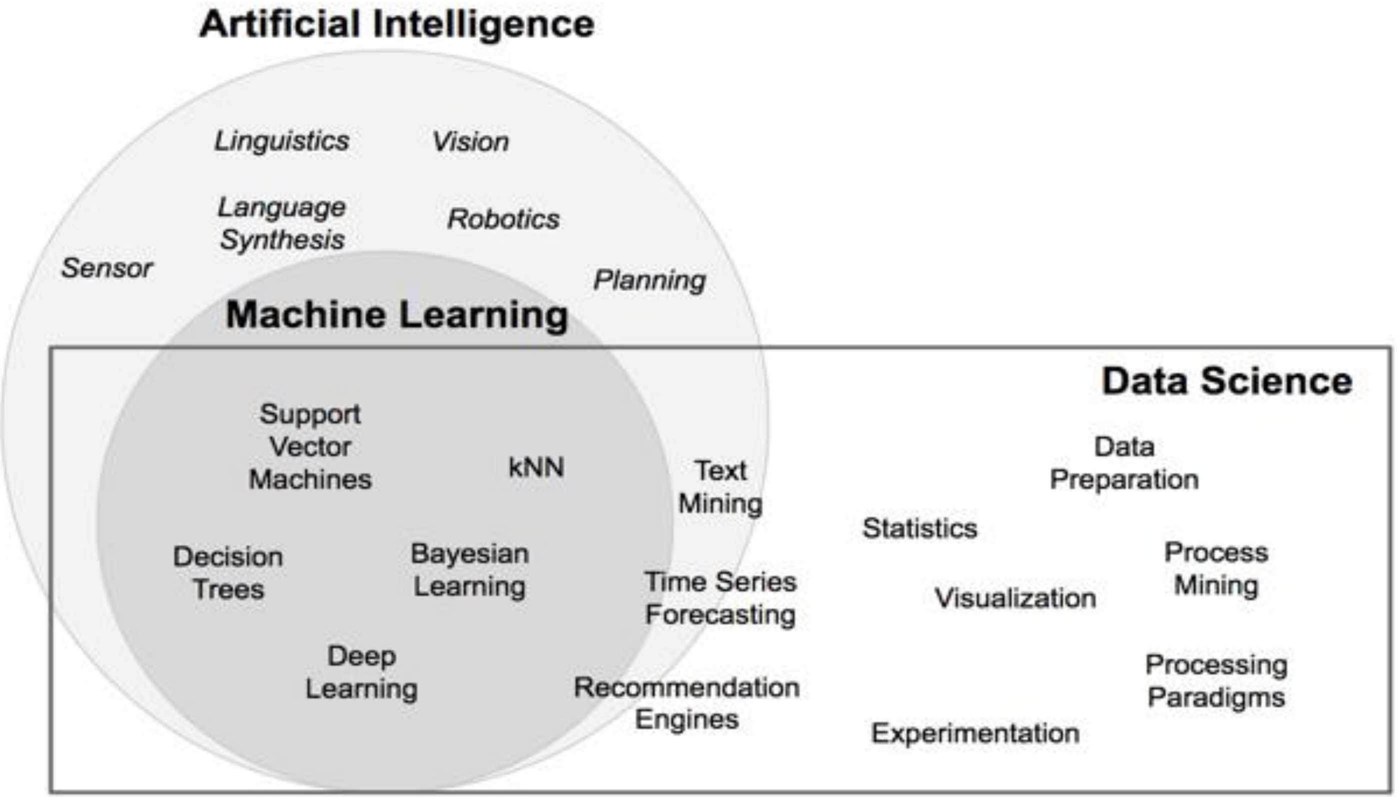
- Artificial intelligence is about giving machines the capability of mimicking human behaviour, particularly cognitive functions.

  eg: facial recognition, automated driving, sorting mail based on postal code

- Machine learning can either be considered a sub-field or one of the tools of artificial intelligence, is providing machines with the capability of learning

-

What is Data Science



What is Data Science

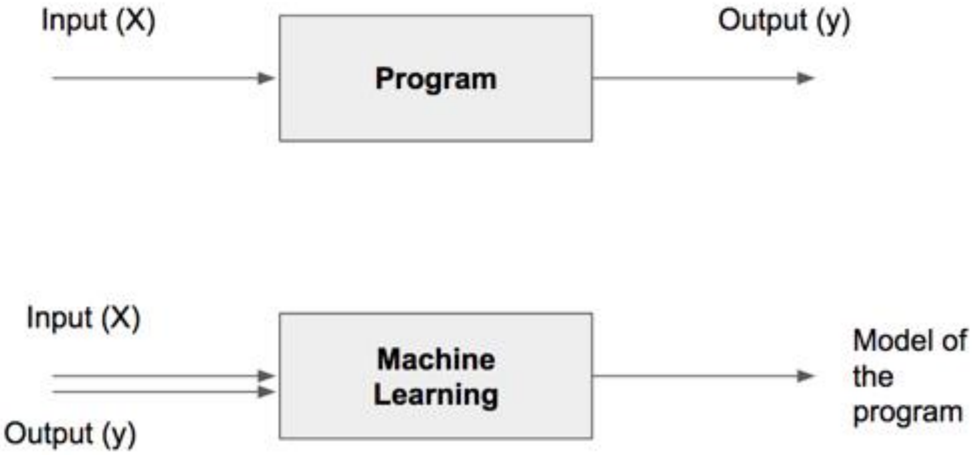
Data science starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset.

The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI)

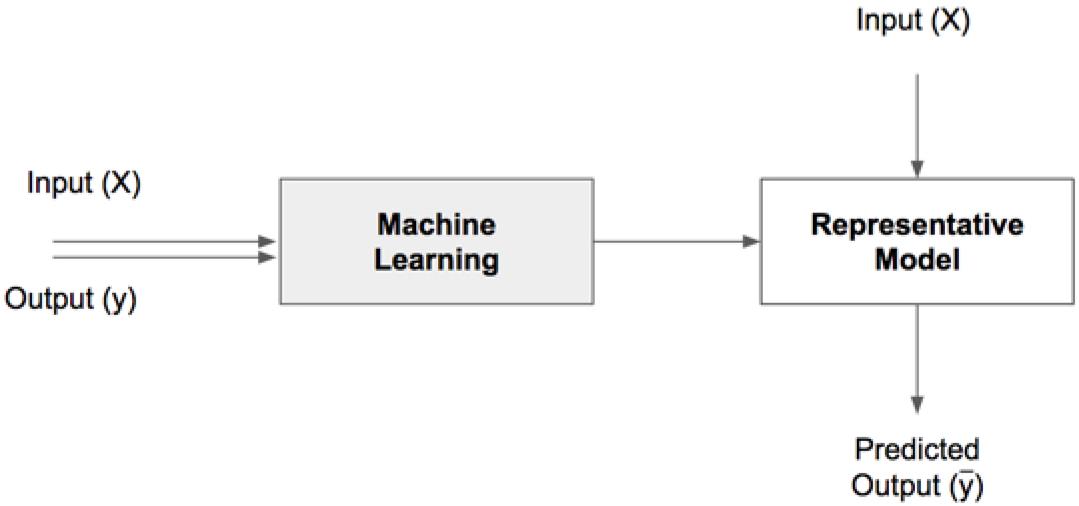
Data Science- features

- ➢ Extracting Meaningful Patterns
- ➢ Building Representative Models
- ➢ Combination of Statistics, Machine Learning, and Computing
- ➢ Learning Algorithms
- ➢ Associated Fields
  - ○ Descriptive statistics:
  - ○ Exploratory visualization
  - ○ Dimensional slicing
  - ○ Hypothesis
  - ○ Data engineering:
  - ○ Business intelligence

Models

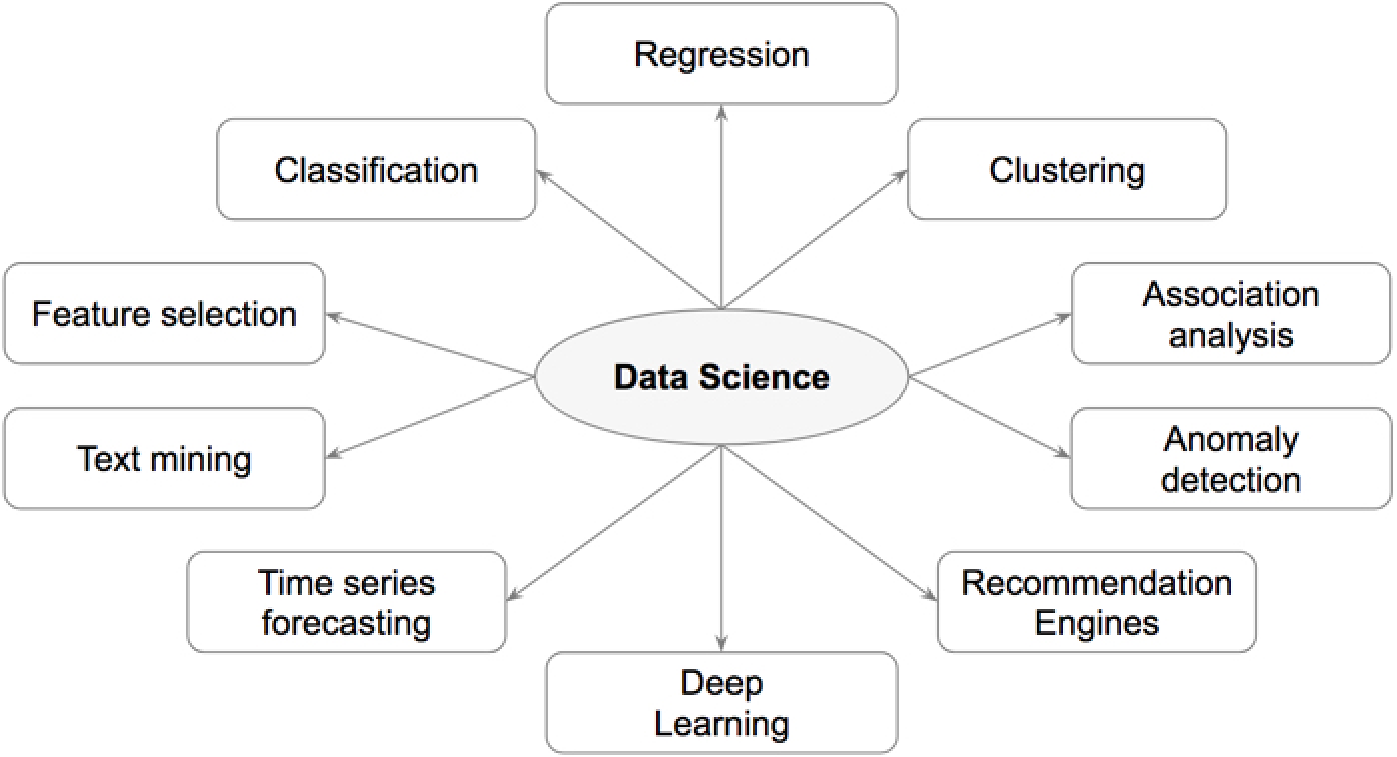


Traditional program and machine learning



Data science models

Types of Data Science



Data Science

Regression

Classification

Clustering

Feature selection

Association analysis

Text mining

Anomaly detection

Time series forecasting

Recommendation Engines

Deep Learning

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

2. Data science process

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

  (5) deploying and maintaining the models.

-

data science process

Over the years of evolution of data science practices, different frameworks for the process have been put forward by various academic and commercial bodies

One of the most popular data science process frameworks is Cross Industry Standard Process for Data Mining (CRISP-DM) (Developed in 2000)

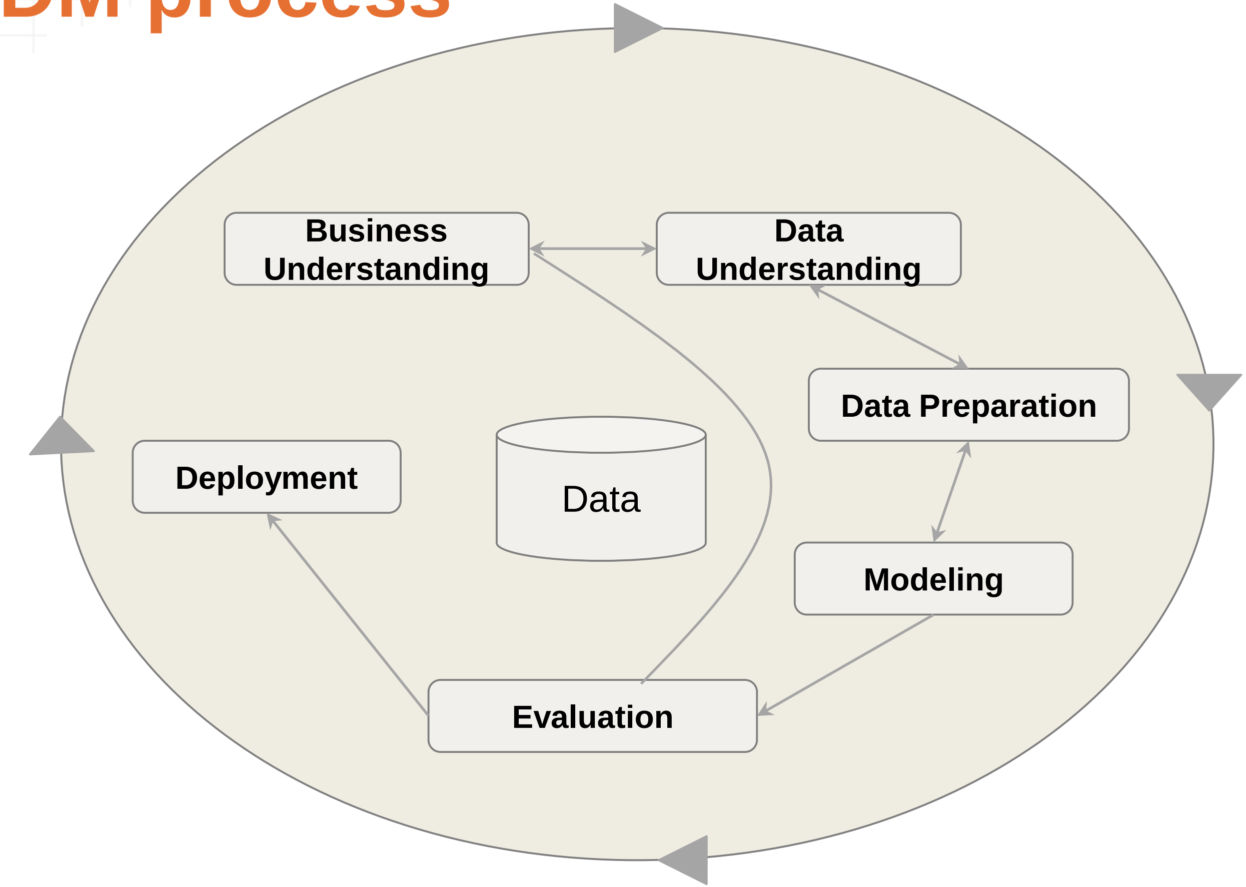
The CRISP-DM process is the most widely adopted framework for developing data science solutions

data science process

A data science process recommends the execution of a certain set of tasks to achieve optimal output.

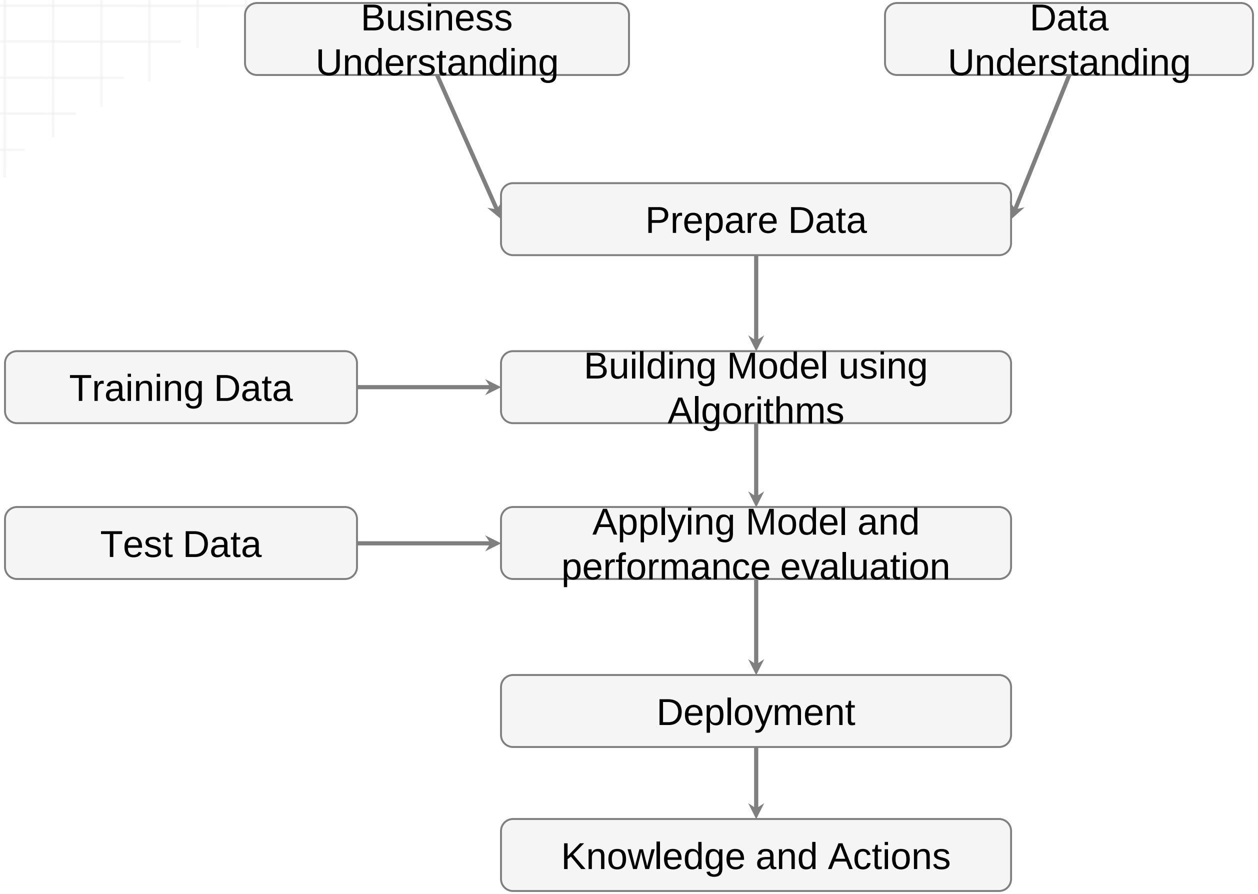
However, the process of extracting information and knowledge from the data is iterative.

CRISP DM process



Cross Industry Standard Process for Data Mining

Process



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

1. Prior Knowledge

Prior knowledge refers to information that is already known about a subject.

Gaining information on:

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

Eg.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

Data Preparation

- 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
- The objective of this step is to come up with a dataset to answer the business question through the data science process.

Data Preparation

- A dataset (example set) is a collection of data with a defined structure
- A data point (record, object or example) is a single instance in the dataset.
- An attribute (feature, input, dimension, variable, or predictor) is a single
- property of the dataset
- A label (class label, output, prediction, target, or response) is the special
- attribute to be predicted based on all the input attributes

2. Data Preparation

- ➢ Data Exploration
- ➢ Data quality
- ➢ Handling missing values
- ➢ Data type conversion
- ➢ Transformation
- ➢ Outliers
- ➢ Feature selection
- ➢ Sampling

Data Preparation

- Preparing the dataset to suit a data science task is the most time-consuming part of the process.
- Preparing the dataset to suit a data science task is the most time-consuming part of the process.

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.

Data Quality

- 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.

Missing Values

- 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.
- 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

Data Types and Conversion

- Different data science algorithms impose different restrictions on the attribute data types.
- Eg: in case of linear regression models,vthe input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute

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

Outliers

- 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.
- Regardless, the presence of outliers needs to be understood and will require special treatments.

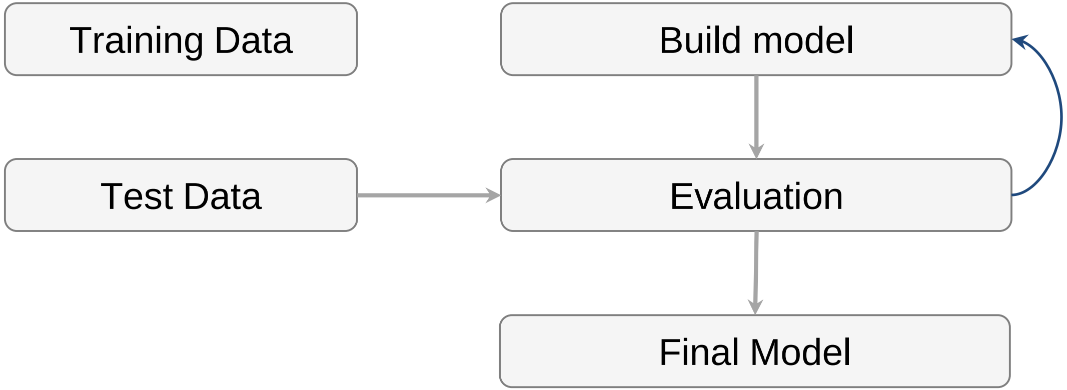
Feature Selection

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.

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.
- Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling

3. Modeling


Training Data
Build model
Test Data
Evaluation
Final Model

Modeling

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

Eg: “mortgage interest rate reduces with increase in credit score”

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.

Also, there are many viable commercial and open source data science tools on the market to automate the execution of these learning algorithms.

3. Modeling

Spliting training and test data sets

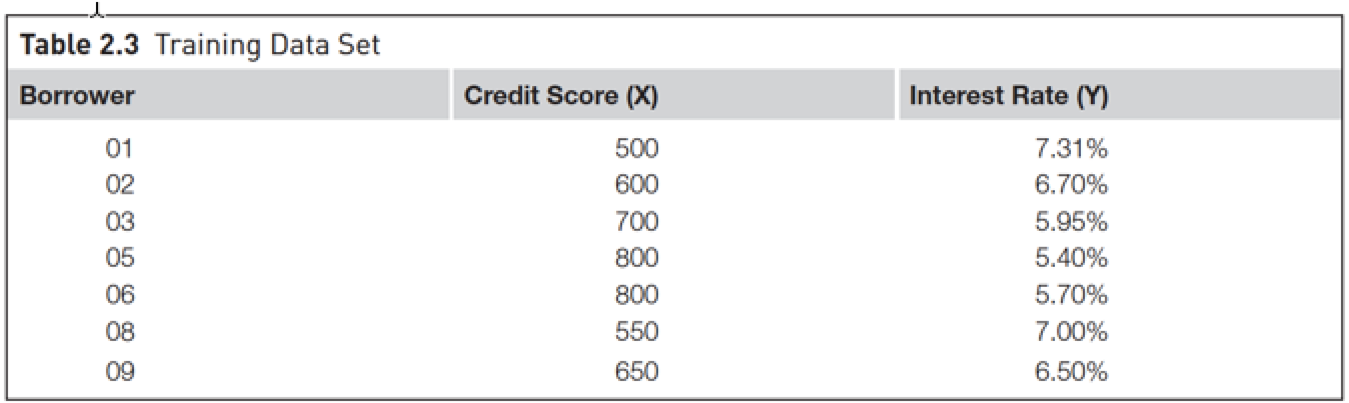


Table 2.3 Training Data Set

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%

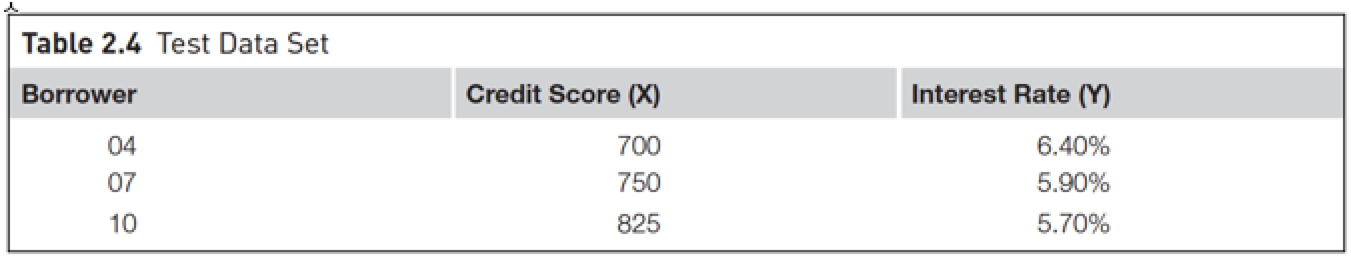
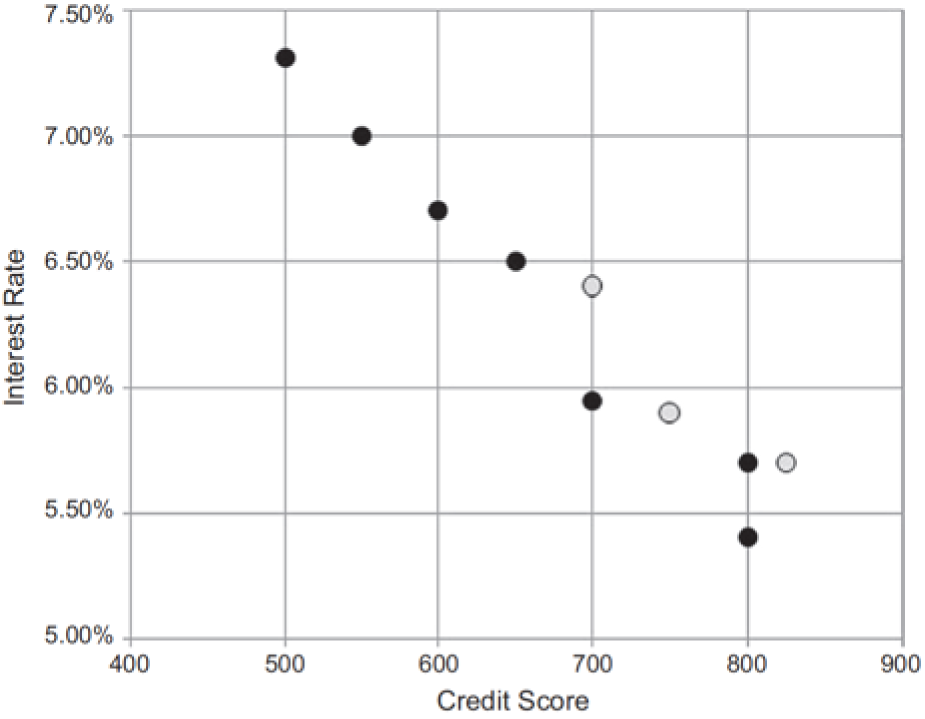


Table 2.4 Test Data Set

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

3. Modeling

Spliting training and test data sets



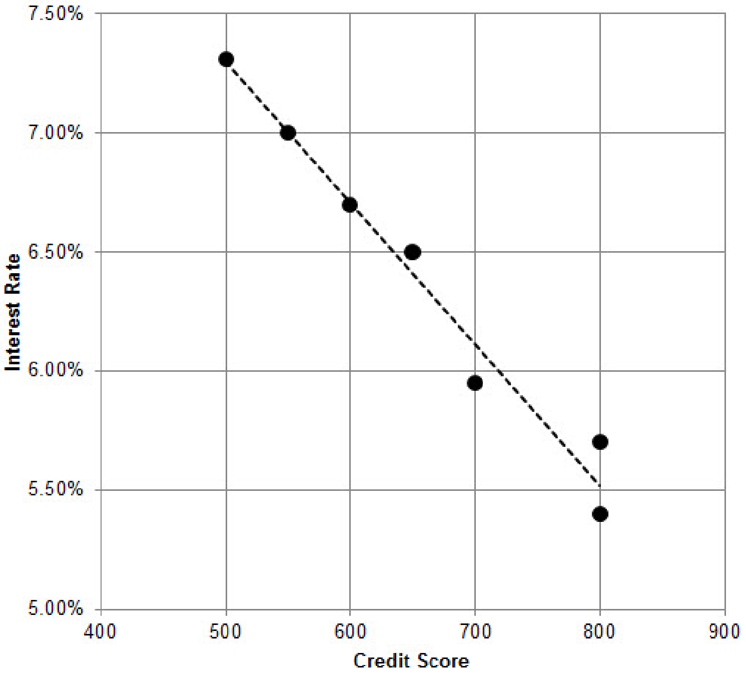
● Training Data

○ Test Data

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.

3. Modeling



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

- The model generated in the form of an equation is generalized and synthesized from seven training records. The estimation may not be exactly the same as the values in the training records
- 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
- 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

3. Modeling

Evaluation of test dataset

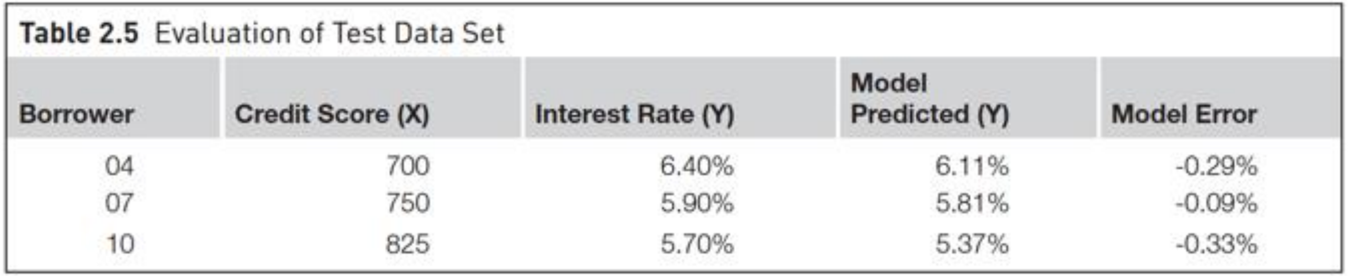


Table 2.5 Evaluation of Test Data Set

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%

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
  - By using different algorithms
  - By using different data sets

4. Application

Product readiness

Technical integration

Model response time

Remodeling

Assimilation

- Product readiness :determines the critical qualities required for the deployment objective.
  - Eg : The consumer credit approval process. 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.

- 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.
- 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

- Model 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

- Remodeling :

- 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

- 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 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

5. Knowledge

Posterior knowledge

- 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 echnology, 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

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.