KCA University
Nairobi, Kenya

BSD 3101: PRINCIPLES OF DATA SCIENCE

Prepared By:

Linus Aloo

E-mail: linusaloo88@gmail.com

Phone: 0754188380

Recommended Reading

1. Sinan Ozdemir, "Principles of Data Science," Packt Publishing,1st Ed. 2016.
2. Vijay Kotu and Bala Deshpande, "Data Science: Concepts and Practice," Elsevier, Second Edition, 2019.
3. Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, "An Introduction to Statistical Learning (with Applications in R), Springer (1st Ed.), 2014.

Additional Reading

1. Joel Grus, "Data Science from Scratch," O'Reilly Media, 1st Ed., 2015.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

The five essential steps to perform data science are as follows:

1. **Asking an interesting question**: Start writing down questions regardless of whether or not you think the data to answer these questions even exists.

2. **Obtaining the data**: Once you have selected the question you want to focus on, it is time to scour the world for the data that might be able to answer that question.

3. **Exploring the data**: begin to break down the types of data that we are dealing with. Once this step is completed, the analyst generally has spent several hours learning about the domain, using code or other tools to manipulate and explore the data, and has a very good sense of what the data might be trying to tell them.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

The five essential steps to perform data science Cont":

4. **Modeling the data**: This step involves the use of statistical and machine learning models. In this step, we are not only fitting and choosing models; we are implanting mathematical validation metrics in order to quantify the models and their effectiveness.

5. **Communicating and visualizing the results**: While it might seem obvious and simple, the ability to conclude your results in a digestible format is much more difficult than it seems. Fig. 2.1 shows the basic processes in data science.

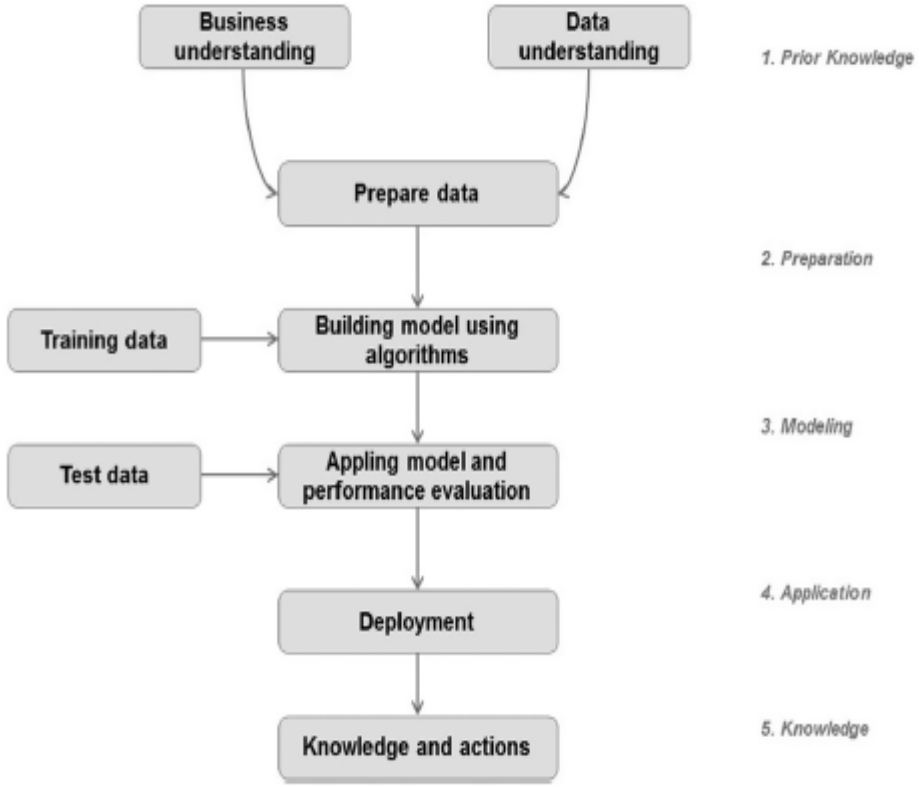


Fig. 2.1 The basic processes in data science.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

1. Prior Knowledge

- 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.
- A *dataset* (example set) is a collection of data with a defined structure. Table 2.1 shows a dataset.
- A *data point* (record, object or example) is a single instance in the dataset. Each row in Table 2.1 is a data point.
- An *attribute* (feature, input, dimension, variable, or predictor) is a single property of the dataset. Each column in Table 2.1 is an attribute.
- 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.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

1. **Prior Knowledge**

- 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.

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

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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.
- **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.

2.1. Data Science Process

- **2.2.1 Data Exploration**

- Fig. 2.2 shows the scatterplot of credit score vs. loan interest rate and it can be observed that as credit score increases, interest rate decreases.

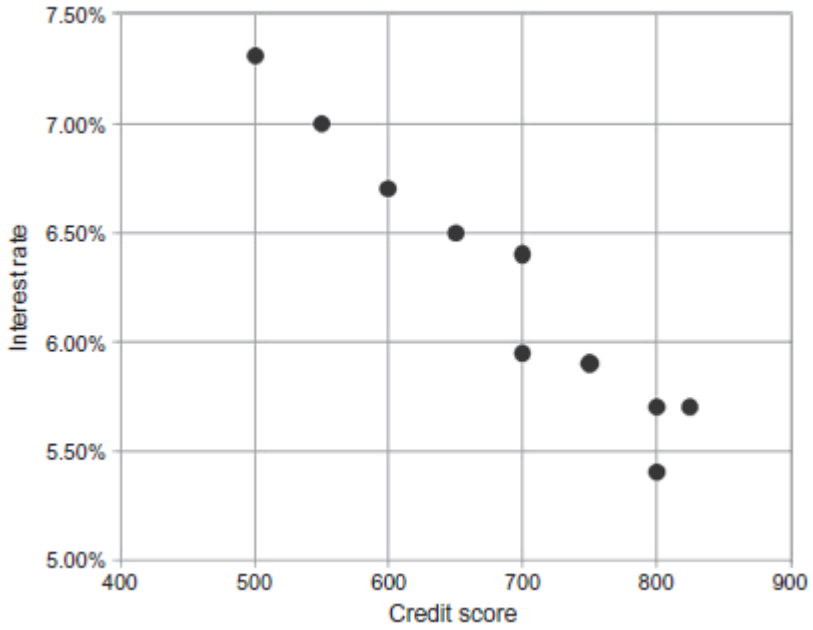


FIGURE 2.2 Scatterplot for interest rate dataset

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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.

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.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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.
- Mitigation methods include tracking the data lineage (provenance) of the data source, the missing value can be substituted with a range of artificial data.
- Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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. 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.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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.

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).

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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.
- 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.

CHAPTER TWO DATA SCIENCE PROCESS

2.1. Data Science Process

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

2.9. Acquiring and Storing Data

2.9.1. Different Sources of Data for Data Analysis

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.
- The data, which is collected, is known as raw data, which is not useful now, but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as "knowledge". The main goal of data collection is to collect information-rich data.

2.9. Acquiring and Storing Data

2.9.1. Different Sources of Data for Data Analysis

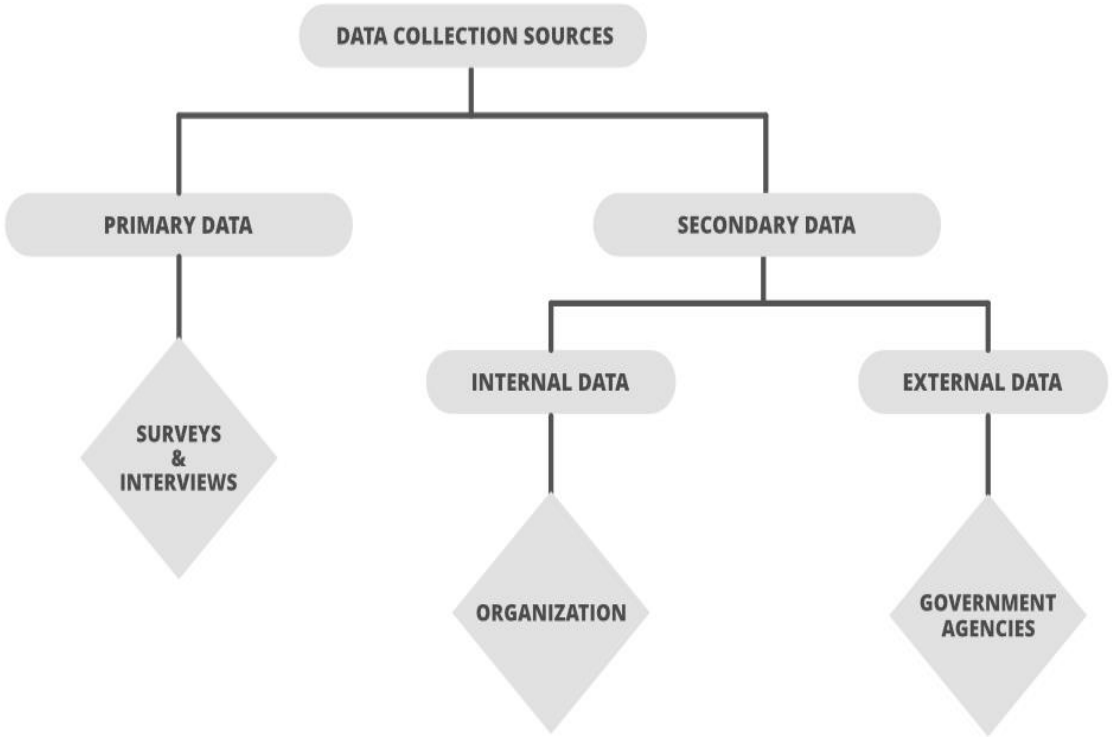
- Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection.
- Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

2.9. Acquiring and Storing Data

2.9.1. Different Sources of Data for Data Analysis

- The actual data is then further divided mainly into two types known as:

1. Primary data
2. Secondary data



2.9. Acquiring and Storing Data

Few methods of collecting primary data:

1. Interview method:

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.
- Can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

- Is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.
- The survey method can be obtained in both online and offline mode like through website forms and email.

CHAPTER TWO DATA SCIENCE PROCESS

2.9. Acquiring and Storing Data

Few methods of collecting primary data:

3. Observation method:

- The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats.

4. Experimental method:

- The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

i. **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication.

CHAPTER TWO DATA SCIENCE PROCESS

2.9. Acquiring and Storing Data

4. Experimental method:

ii. **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA).

iii. **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains equal no. of rows and columns (N×N squares) which contain letters that occurs only once in a row.

iv. **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

2.9. Acquiring and Storing Data

2. Secondary data:

- Secondary data is the data, which has already been collected and reused again for some valid purpose.
- This type of data is previously recorded from primary data and it has two types of sources:

a)Internal source:

- These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc.
- The cost and time consumption is less in obtaining internal sources.

b) External source:

- Is the data which can't be found at internal organizations and can be gained through external third party resources.
- The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, etc.

2.9. Acquiring and Storing Data

Other sources:

- **Sensors data:** e.g. sensors that come with IoT devices collect data, which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras.
- **Web traffic:** Through the internet, many formats of data, uploaded by users on different platforms, can be predicted and collected with their permission for data analysis.

  The search engines also provide their data through keywords and queries searched mostly.

2.9. Acquiring and Storing Data

2.9.2. Data Storage Types

- Data storage essentially means that ***files and documents are recorded digitally and saved in a storage system for future use***.
- Data storage can occur on ***physical hard drives, disk drives, USB drives or virtually in the cloud***.
- There are two broad types of data storage: **direct attached storage** and **network attached storage**.

a) Direct Attached Storage (DAS)

- Direct attached storage (DAS) includes types of data storage that are physically connected to your computer.
- Generally accessible to only a single machine. Examples include: Hard Drives, Solid-State Drives (SSD) CD/DVD Drives, Flash Drives,etc.
- DAS solutions are great for creating local backups and can be more affordable than NAS solutions, but sharing data between machines can be cumbersome.

2.9. Acquiring and Storing Data

2.9.2. Data Storage Types

b) **Network Attached Storage (NAS)**

- Network attached storage (NAS) allows for multiple machines to share storage over a network.
- This is accomplished with multiple hard drives or other storage devices in a RAID configuration.
- One of the key benefits of NAS is the ability to centralize data and improve collaboration.
- Data can be easily shared among connected machines, and permission levels can be set to control access.
- While NAS solutions tend to be more costly than DAS solutions, they are still very affordable as storage technology has advanced significantly.
- Exercise: Discuss the following types of storage devices: SSD Flash Drive Arrays, Hybrid Flash Arrays, Hybrid Cloud Storage, Backup Software , Backup Appliances and Cloud Storage.

2.9. Acquiring and Storing Data

2.9.2. Data Storage Types

Benefits of Efficient Data Storage

- Reliable data preservation
- Data continuity and accessibility
- Quicker and easier data recovery
- Flexible price points and capacity options
- Effective security for protected files

2.9.3. Big Data storage methods

There are currently two well-established big data storage methods:

i. **Warehouse Storage** – Similar to a warehouse for storing physical goods, a data warehouse is a large building facility, which its primary function is to store, and process data on an enterprise level.

ii. **Cloud Storage** – With cloud storage, data and information are stored electronically online where it can be accessed from anywhere, negating the need for direct attached access to a hard drive or computer. With this approach, you can store virtually boundless amount of data online and access it from anywhere.