# Brief of Exploratory Data Analysis (EDA)

This section presents some characteristics of Portuguese bank dataset. As we can see in Figure 1 and as we mentioned earlier that this dataset is imbalanced as it has only 11.6% of the class “Yes” in the response variable “y”.



The data characteristics composed of two kinds: categorical and numeral attributes, as shown in Table 4. This table shows that three kinds of attributes , which are qualitative/discrete like (Age, Balance, Day, Duration, Campaign, Pdays, Previous), qualitative/nominal/binary are all the attributes that represented as yes or no in their classes; for example, the attributes (Default, Housing, Loan, y (output variable)), and qualitative/nominal which are (Job, Marital, Education, Contact, Month, Poutcome).

The column under name “Unique values” shows the number classes for each attribute.

For example, we can notice that the age attribute has range of numbers between 18 and 95., the second attribute “job” shows the kinds of jobs of the bank customers as (admin, unknown,

unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, and services). The attribute Marital can be illustrated in classes as (married,

divorced, and single) where the class divorce means divorced or widowed.

The Education classes are divided into unknown, secondary, primary, and tertiary; however, in attributes Default, Housing, Loan, and the y attribute has only two classes (yes, and no). The contact communication classes in the Contact attribute are unknown, telephone, and cellular.

In addition to that, in the attribute Month the classes are month’s names Jan, Feb, Mar etc. The attribute Poutcome presents the outcome of the previous marketing campaign like unknown, other, failure, and success. It is notable that the value “unknown” in the attribute, is determined as the greatest one with the percentage 82% (36,959 unknown records). The attribute Balance has a range of numbers between -8,019 and 1,362 which reflects the average yearly balance, in euros. Day attribute shows a range between 1 and 31 which is last contact day of the month for the current campaign. The Duration attribute which reflects the last contact duration in seconds and its range between 0 to 4,918 seconds.

attribute Campaign shows in its domain the number of contacts performed during this campaign, and for this, client (includes last contact) is in the interval from 1 to 63; however, the domain ranged between -1 to 871 is representing the number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted) in the attribute Pdays.

Finally, the attribute Previous presents the number of contacts performed before this campaign and for this client, its domain from 0 to 275.

Data discription of the Portuguese bank dataset

Table 5 displays more information about these variables such as (mean, median, maximum, minimum, and missing values checking results) as bellow.

Data discription of the Portuguese bank dataset

Another exploratory analysis was conducted to see the distribution of each numeric attribute as Figure 1 below. It is important to mention that non of the parametric attributes has a normal distribution.

Distribution of numeric attributes

The following figure presents the correlation between the number variables (Figure 3), the correlation between the categorical variables (Figure 4).

Correlation between the numeric variables

 Correlation between the categorical variables