Data Wrangling Report

Capstone Project 1: Bank Marketing

Data collection :

The data used for this project is available at the UCI Repository.

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

The data was acquired directly from here. There are two data sets available here. One with 21 attributes and the other with 17 attributes. The dataset with 17 attributes is used for this project.

The dataset has 45,211 instances with a total of 17 attributes (it contains one attribute for the output variable 'y').

Data Reading :

The data was loaded in the Jupyter Notebook using Pandas library, (pd.read_csv()). Using this, The data is loaded as a Pandas DataFrame. The delimiter was not the ',' as usual. Hence, the delimiter was specified as ';'.

Data wrangling :

- Initially, **.head()** function is used to get a clear picture of what the data looks like.
- **.describe()** - To compute various statistics on the DataFrame columns. It computed count, mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile and maximum for the numerical data columns.
- **.info()** - gives the basic information of each of the columns. It returns the column name along with the count of the non-null rows of the data along with the data type of the column.
- **Missing values check: isnull()** function of the panda library was used to check if there are any missing values in the DataFrame. A logical function, that checks if there are any missing values was passed for each row inside the DataFrame. The output was applied to each column by specifying (axis =1). The count of isnull() function returned no missing values.
- **Knowing the categorical variables- .unique()** was used to display the various values of the categorical columns. It was computed for 10 categorical columns.

- **Knowing the numerical variables- boxplots** were plotted for the numerical columns. The boxplots revealed **outliers** for the numerical data. But the outliers are not removed for now for further analysis.

Data cleaning :

The data is clean and it does not require any pivoting or melting of the DataFrame. It does not require any further modifications.