**DATA WRANGLING**

The Car Insurance Cold Calls Dataset is obtained from Kaggle as a CSV file. In order to work with the dataset, the python libraries such as Pandas, Numpy, Matplotlib and Seaborn which support data analysis has to be imported. Next the dataset is loaded into pandas dataframe by using the pandas. read\_csv function and using the document name and Id field as the Column index.

The dataset is now loaded into the pandas dataframe and ready for analysis. First to have a sneak peek at the data, the head function is used which lists the top 5 rows in the dataset. This shows the organization of our fields and whether they have been loaded into the dataframe without any errors. The tail function just works opposite to the head function, it displays the last 5 rows of the dataframe.

**EXPLORATORY DATA ANALYSIS**

The dimensionality of the dataframe is checked using the shape attribute within the dataframe and there are 18 columns to deal with. The columns built in attribute within the dataframe displays the name of all the columns associated with the dataset. If there are any error in the name it can be corrected here. The dataframe we are working with has no issues with the column names and hence we are good to go. The describe function displays the statistics of numerical data attributes in the dataframe. The mean, std, min and max helps to figure out how the data is distributed throughout. Here by looking at the descriptive statistics the fields Default, HHInsurance, CarLoan, CarInsurance have binary values. The dtypes attribute is used to figure out the datatypes for each field in the dataframe. The numerical columns are integer and categorical are object datatypes in the dataset. The descriptive statistics of the categorical data is also viewed like count, freq, top and unique values.

**OUTLIER ANALYSIS**

The Balance field in our data is the only continuous field apart from age and univariate analysis of the Balance field is done by plotting a boxplot using the seaborn library. Well, looks interesting there are quite a few outliers after the maximum. But the Balance field is not evenly distributed and the outliers are also of our interest since a large quantity lies outside the whisker of the boxplot. One value here is very high when compared to the mean and 75% value is 1619 only. Since the last outlier can be dropped from the dataset. To do this let’s figure out the maximum value in the Balance field. Turns out it is 98417 and now displaying the column of interest using the maximum value to capture the index. Drop function is used to drop the outlier in the Balance field.

**HANDLING MISSING VALUES**

The missing values in the dataframe can be found using the isnull function and using sum to count all the null values. The major missing values are present in Outcome and Communication fields followed by Education and Age. The missing values in these fields need to be addressed since it helps with further analysis and in model building.First we look at the count of the fields using value\_counts function and use fillna function and frontfill option to fill the Age column. Similar method is performed for education by using frontfill and fillna option. The missing values in Age and Education has been imputed.

Further Communication NaN is imputed using None , the sam fillna function is used here also and since outcome is also a categorical field None can be used here as well to change the NaN missing values to None.