Data Glacier Intern Project Report

Project: Bank Marketing (Campaign)

Group: Model Maestros

Group Member 1

Name: Nrusimha Saraswati Sai Teja Jampani

Email: njampani@buffalo.edu

Country: United States

College: State University of New York at Buffalo

Specialization: Data Science

Group Member 2

Name - Purvesh Mehta

Email - mpurvesh007@gmail.com

Country - United Kingdom

University - University of Sussex

Specialization - Data Science

Group Member 3

Name: Aysha Abdul Azeez

Email: ayshaabdulazeez41@gmail.com

Country: United Kingdom

College/Company: University of Central Lancashire

Specialization: Data Science

Problem Description

ABC bank aims to launch a new term deposit scheme and wants to sell this product to customers. Prior to the launch, the bank plans to start a marketing campaign for the product through various marketing channels like Telephone, SMS, Emails, etc. To save time and to minimize the costs associated with this process, the bank wants to shortlist all the potential customers who have a greater possibility of buying the term deposit product.

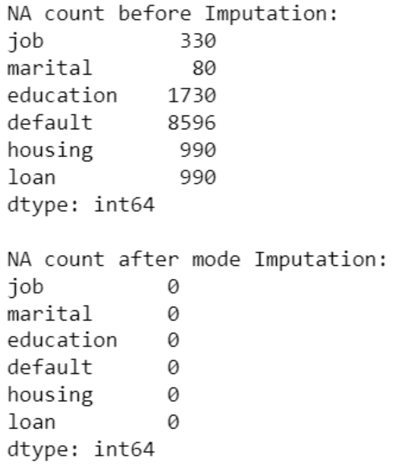
This will help the marketing team to start a campaign on a set lot of customers without wasting their resources on any unlikely buyers. To achieve this outcome, we will need to develop a classification model with high accuracy to determine if a customer will subscribe to the term deposit or not based on the available marketing data.

Data Cleaning

Duplicates: The data contains 12 duplicates and are removed.

Handling NA values: We have used three imputation methods to handle the NA values.

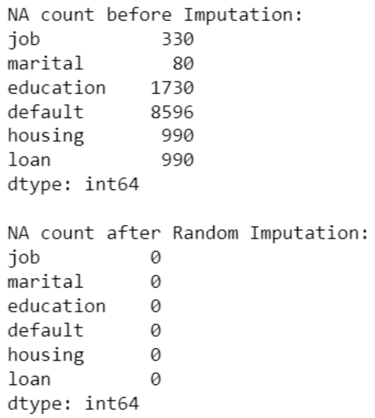
In method 1, we have replaced the NA values with the most frequent value in the column i.e., mode.



```
NA count before Imputation:
job            330
marital         80
education     1730
default       8596
housing        990
loan           990
dtype: int64

NA count after mode Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

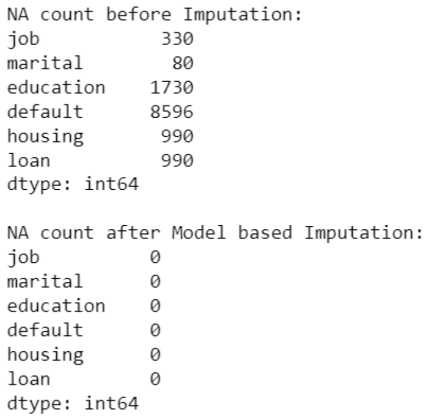
In method 2, we have imputed the NA values with random values taken from the column.



```
NA count before Imputation:
job           330
marital        80
education    1730
default      8596
housing       990
loan          990
dtype: int64

NA count after Random Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

In method 3, we have imputed the NA values by values predicted by a random forest classifier. For this, we have considered the column with NA values as target and the rest of the columns as feature variables.



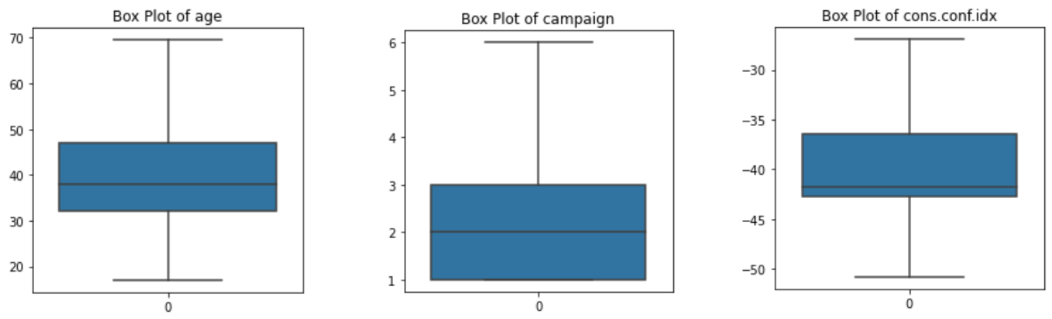
```
NA count before Imputation:
job           330
marital        80
education    1730
default      8596
housing       990
loan          990
dtype: int64

NA count after Model based Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

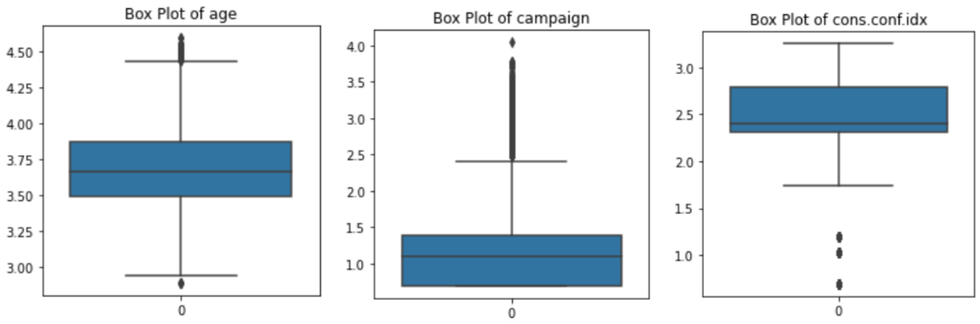
Outlier and Skewness

Previously, we have identified that columns ‘age’, ‘campaign’ and ‘cons.conf.idx’ contain outliers.

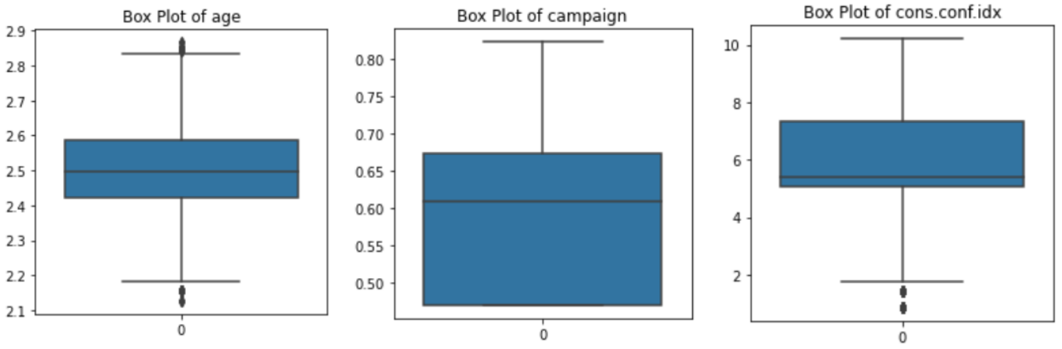
In method 1, we have trimmed the values to lie between upper and lower quartile. In this way we have eliminated almost all the outliers from our data. This method is easy to implement but does not handle skewness in the data.



In method 2, we have applied log transformation to our columns containing outliers. The log transformation does not fully eliminate outliers but eliminates skewness to some extent.



In method 3, we have applied boxcox transformation to our data. Box cox handled most of the outliers while making the data more symmetric and eliminating skewness.


Box Plot of age
Box Plot of campaign
Box Plot of cons.conf.idx