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: Mufunwa Nemushungwa

Email: mufunwanemushungwa@gmail.com

Country: South Africa

College/Company: University of the Witwatersrand

Specialization: Data Science

Group Member 4

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 Understanding

The data to be used in the project contains 21 columns and 41188 rows. The data is enclosed in a csv file delimited by semicolon. Description of each column is given below.

Column	**Description**
Age	Age of the customer
Job	Type of job taken by the customer

Martial	Martial status of the customer
Education	Educational qualification of the customer
Default	Does the customer have a defaulted credit
Housing	Does the customer have a housing loan
Loan	Does the customer have a personal loan
Contact	Communication type for the customer
Month	Last contact month of the customer
Day_of_week	Last contact day of the week
Duration	Last contact duration of the customer
Campaign	Number of times the customer is contacted
Pdays	Number of days passed by after client was contacted
Previous	Number of contacts made to client before campaign
Poutcome	Outcome of the previous campaign for the client
Emp.var.rate	Employment Variation Rate - Quarterly
Cons.price.idx	Consumer price index - Monthly
Cons.conf.idx	Consumer Confidence index - Monthly
Euribor3m	Euribor three-month rate - Daily
Nr.employed	Number of employees - Quarterly
Y	Target Variable – If client subscribed to the plan

Job, martial, education, default, housing, loan, contact, month, and day_of_week and poutcome are categorical variables and the rest of the columns are numeric.

Problems

1. **NA Values:** The data implicitly does not contain any None or NaN values. However, some columns contain 'unknown' in some of the rows. We can consider 'unknown' synonymous to NaN since both convey the same meaning. The following images describes the number of unknowns in each column.

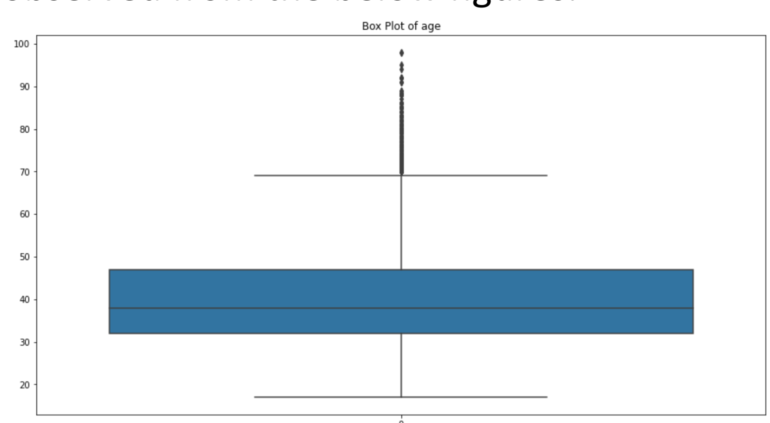
```
age                0
job              330
marital           80
education       1731
default         8597
housing          990
loan             990
contact            0
month              0
day_of_week        0
duration           0
```

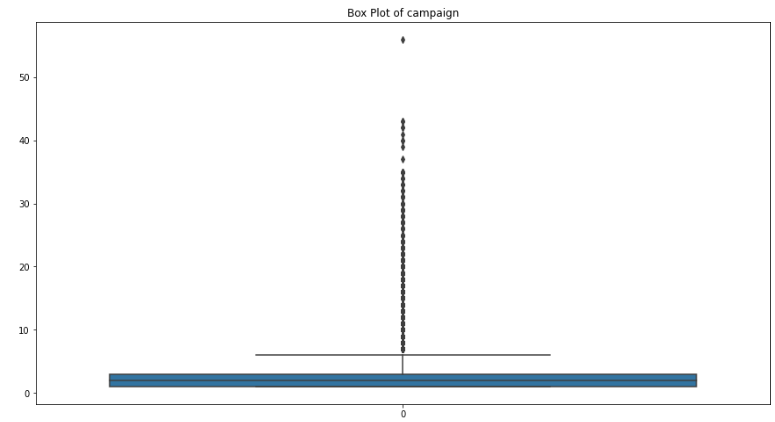
```
campaign           0
pdays              0
previous           0
poutcome           0
emp.var.rate       0
cons.price.idx     0
cons.conf.idx      0
euribor3m          0
nr.employed        0
y                  0
```

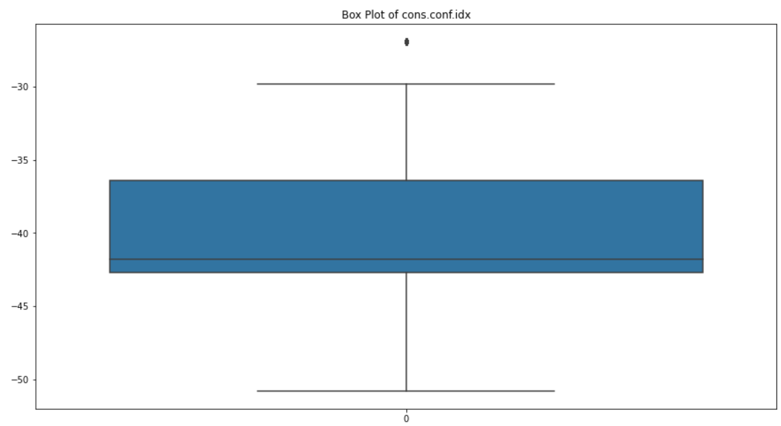
2. **Categorical Variables:** The data contains categorical columns such as job, martial, education, etc. These variables needs to encoded to pass it through a machine learning model. The categorical variables are given below.

```
job            object
marital        object
education      object
default        object
housing        object
loan           object
contact        object
month          object
day_of_week    object
poutcome       object
y              object
```

3. **Outliers:** We have created box plots to identify any outliers present in the numeric columns. It is observed that no other columns except age, campaign and cons.conf.idx contain outliers. We can interpret from the below figures that age above 70, campaign above 5 days and consumer confidence index above -30 are all outliers as observed from the below figures.





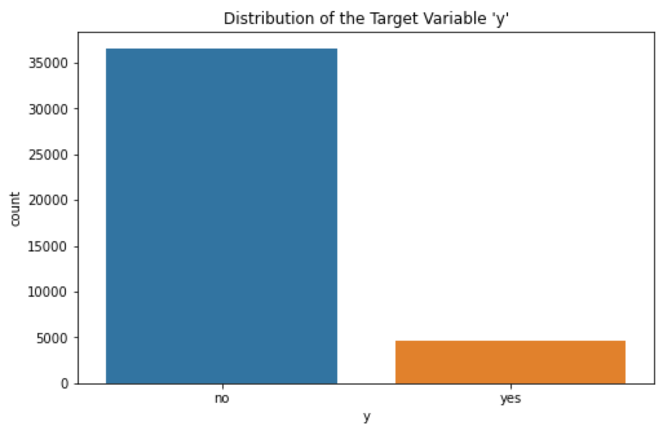


4. **Skewness:** We have observed skewness present in our dataset. The amount of skewness present in each numeric column is given below.

age	0.784697
campaign	4.762507
pdays	-4.922190
previous	3.832042
emp.var.rate	-0.724096
cons.price.idx	-0.230888
cons.conf.idx	0.303180
euribor3m	-0.709188
nr.employed	-1.044262

5. **Duplicates:** The data contains some duplicate values which might affect the model prediction and must be handled effectively.

6. **Target Variable Distribution:** In our analysis, we found that the more ‘No’ than ‘Yes’ in the distribution of the target variable y. This can lead to bias and would negatively impact the performance of the model.



Approaches

- For NA or unknown values, we can impute the unknowns with mean, median or mode value of the column. In this way, we will make the unknown value as close as possible to its true value. It is also possible to replace the unknown with a random sample taken from the data. One more technique is to use a model-based approach to fill the unknowns by considering information from other columns to predict the unknown value. In this project we shall evaluate the results from all the above-mentioned techniques and choose the method that produces the best results.
- Duplicates contribute to inconsistencies in the prediction. Duplicates also lead to overfitting of the model. In our analysis, we find that there are only 12 duplicate records in

the data and thus it is better to drop them before applying any machine learning model.

- Outliers can be handled by truncating them with some upper or lower threshold values. We can also delete the values or apply any mathematical transformations like log, square root, etc., to reduce the impact. We aim to replace the outliers with upper threshold, verify the results and then proceed to other methods based on the results. In this way we can reduce the complexity without inducing more inconsistencies in the prediction.
- Categorical variables need to be encoded to numeric values to pass it to the model. We will be using label encoder or one hot encoder to achieve this. This will create new columns after converting the categorical values to numeric which can then be interpreted by the model.
- For handling the skewness in the data, we aim to normalize the data manually or apply transformations such as logarithmic or box cox. Normalization will convert the data into a normal distribution to have a constant mean and standard deviation. This will ensure that the prediction is not biased by the skewness present in the data. Logarithmic or box cox transformations will compress the extreme values to smooth the data and introduce normality in the distribution.
- For the Imbalanced target variable distribution, we need to apply under sampling or oversampling techniques to balance the class distribution. Oversampling will increase

the number of instances of the minority class and under sampling will reduce the instances of the majority class. Since the dataset is not extremely large, will shall apply oversampling, verify the results and also try out under sampling if required.