Bank Marketing Campaign Project

Group Name: Yomna's Group

Name: Yomna Eisa

Email: yomnaabelrahmaneisa@gmail.com

Country: Saudi Arabia

College/Company: N/A (fresh graduate)

Specialization: Data Science

GitHub Link:

1. **Problem description**

   ABC Bank plans to launch a term deposit product and seek to build a predictive model to identify potential customers likely to purchase it. By utilizing machine learning (ML) models, the bank aims to optimize its marketing efforts, targeting customers with a higher probability of buying the product. This strategy, implemented through telemarketing, SMS, and marketing channels, aims to save resources and reduce costs associated with resource billing.

2. **Data understanding**

   The dataset named 'bank-additional-full' is a CSV file that consists of 21 columns and 41188 rows. The file contains data from May 2008 to November 2010. The data covers information regarding the marketing campaign itself such as employment variation rate, number of employees, consumer confidence index, and Euribor 3-month rate. As well as all the basic client information, such as age, job, education, marital status…etc. Lastly, there's the variable 'y' which is an answer to the question

‘Has the client subscribed to a term deposit? ’ The answer is a binary yes or no.

3. Type of data for analysis

Column Name	Data Type	No. null/unknown values	No. of outliners
age	Integer	0	0
job	String	**330**	0
marital	String	**80**	0
education	String	**1731**	0
default	String	**8597**	0
housing	String	**990**	0
loan	String	**990**	0
contact	String	0	0
day	String	0	0
month	String	0	0
year	Integer	0	0
pdays	Integer	0	0
previous	Integer	0	0
poutcome	String	0	0
campaign	String	0	0
day_of_week	String	0	0

duration	Integer	0	**1446**
Emp.var.rate	Float	0	0
cons.price.idx	Float	0	0
Cons.conf.idx	Float	0	0
euribor3m	Float	0	0
nr.employed	Float	0	0
y	String	0	0

4. Problems in the data

4.1 Null values

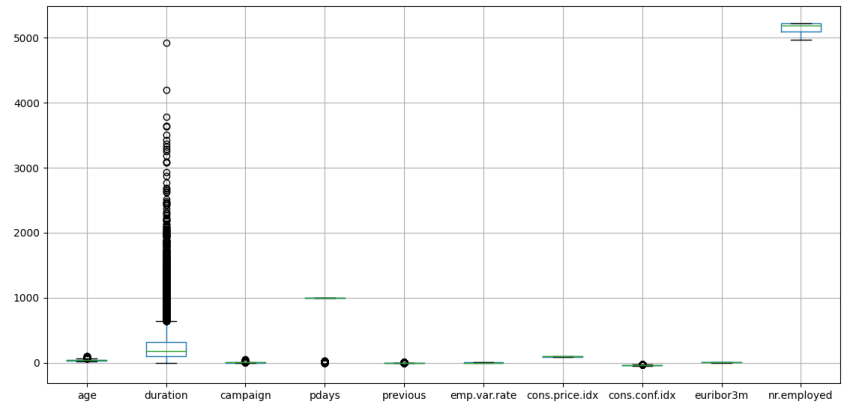
No null values were found in the dataset.

4.2 'Unknown' values

The following 'unknown' values were found in the dataset:
Column 'job' has **330** 'unknown' values.
Column 'marital' has **80** 'unknown' values.
Column 'education' has **1731** 'unknown' values.
Column 'default' has **8597** 'unknown' values.
Column 'housing' has **990** 'unknown' values.
Column 'loan' has **990** 'unknown' values.

4.3 Outliner values

The column 'duration' contains outliner data based on the fact that it has a mean of 258.28 while the max value is 4918. The graph below confirms the presence of outliner data in the column 'duration':



4.4 Unbalanced data

When calculating the proportion of each class in the target variable ‘y’, these were the results:

no *0.887346*

yes *0.112654*

We can conclude based on the above that the dataset is unbalanced given the fact that the class ‘no’ is significantly larger than the class ‘yes’

5. Approaches to solve problems in the data

5.1 Solution for unknown values

Depending on the column itself and the data it holds, either we will drop the row or replace the ‘unknown’ value with the mode. For the housing and loan columns, we replace the missing values with the mode. Moreover, for the columns job and marital, we will drop the missing values

5.2 Solution for Outliner Values

For removing the outliners, we’ll use the Z-Score Method and the IQR

5.3 Solution for unbalanced data

For the unbalanced data, we decided on undersampling: which involves reducing the number of instances in the majority class to balance it with the minority class.