Virtual Internship
Data Science
Data Intake Report

Group Name: LISUM01: Data science Group 1
Members:
1. William Ogweli Okomba, willokomba@gmail.com, Kenya
2. Ece Kurnaz, eceeee.kurnaz@gmail.com, Turkey
3. Collin Mburugu, colinmburugu@gmail.com, Kenya
4. Udbhav Balaji, udbhavbalaji@gmail.com, India

Name: Bank Marketing(Campaign)
Report date: 02/08/2021
Internship Batch: LISUM01
Version:1.0
Data intake by: William Ogweli Okomba
Data intake reviewer: Intern who viewed the report
Data storage location: https://github.com/williamokomba/DataGlacier_Internship-Data-science/tree/main/week_8_Data_science_group%20Project

Problem statement

ABC bank (a Portuguese banking institution) has a term deposit product that is desired to be sold to clients. We will focus on customer's past interactions with the bank or other financial institutions to have a better understanding on whether these particular clients will buy this product or not. Developing a model with using machine learning for this aim is reasonable. With performing this project, our aim is to save resources and time for ABC bank.

Data Understanding

1. We were provided with 2 files(datasets) namely bank_additional_full.csv, and bank_additional.csv
2. The second dataset is a sample of the first dataset
3. The data cover a period from May 2008 to November 2010.
4. The 1st dataset (bank_additional_full.csv) contains 41188 records & 21 columns, whereas the 2nd dataset (bank_additional.csv) has 4119 records and 21 columns.
5. Columns are not uniformed named for example "day_of_week", and "emp.var.rate". This need to be modified for make it easier to work with.

6. Some of the columns names were changed for easy understanding.
7. All variables in both datasets have the right datatypes, and all variables unique values are fine.
8. The mean age of the respondents is 40 years, minimum age is 17years and maximum age is 98years
9. The mean duration of the time taken talking to the respondent is 258.3 seconds, minimum time is 0 seconds. This will be discarded as it does not add value to our analysis
10. The minimum number of contacts during this campaign was 1 contact and the maximum was 56 contacts
11. pdays variable value 999 means the lapsed days before the person was conducted thus we will make this value to be zero.
12. Some variables are right skewed (e.g “age” variable) while others are left skewed (e.g “nr.employed”)variable.
13. Our aim is to predict on whether the customer will buy “y” product or note based on other variable details.

What type of data you have got for analysis

14. The dataset contains numeric (5 int64, & 5 float 64), and 11 categorical data types.

The Files Description:

15. bank_additional_full.csv, our dataframe: bank_df dataset
16. Bank_additional.csv, data frame: bank dataframe.

What are the problems in data?

17. There are 12 duplicates values in the original dataset, we dropped them since they are minimal.
18. The target variable is imbalanced class, "no" class has more observation than "yes" class in both dataset. We will impute it to balance during modeling.
19. Missing values were recorded as Unknown, we changed them to NaN values.
20. The variable “job”, “marital”, “education”, “dafualt”, “housing”, and “loan” has missing values.

    The missing values were imputed as follows

    a. “job”, marital, default variables was imputed by the mode of the column is it is a categorical column.
    b. “education” variable was imputed with “N/A”(not applicable), this means we assumed the respondents didn’t have any formal education thus not applicable.
    c. “housing”, and “loan” variable missing values were dropped since they were minimal and will not affect out analysis.
21. “age”, “duration”, “campaign”, “pdays”, “previous” and “cons.conf.idx” have outliers. We removed outlier in age variable since it was only one at the extremed end with was altering the variable mean. Outliers on other variables we decided to not to remove since they looks genuine.
22. We dropped “contact”, and “duration” columns, they are not going to add value to our analysis.

23. Our columns has different formats, we changed them to standard form.