Data Glacier

Your Deep Learning Partner

Virtual Internship
Data Science
Data Cleaning

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: 03/08/2021
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_9_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 cleaning and transformation done on the data

1. Columns are not uniformly named for example "day_of_week", and "emp.var.rate". The columns were modified with underscores between the spaces.
2. Some of the columns names were changed for easy understanding.

3. There are 12 duplicates values in the original dataset, we dropped them since they are minimal.
4. 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.
5. Missing values were recorded as Unknown, we changed them to NaN values.
6. 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.
7. “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.
8. We dropped “contact”, and “duration” columns, they are not going to add value to our analysis.
9. Our columns has different formats, we changed them to standard form.