Data Intake Report

Project name: Bank Marketing (Campaign) -- Group Project

Report date: 30th August 2022

Internship Batch: LISUM10

Version:1.0
Data intake by: Mohini Kalbandhe , Kashish Joshipura, Amir Shahcheraghian, Mohammed Maqsood

Data intake reviewer:

Data storage location: https://github.com/amohini099/Banco-de-portugal-marketing/tree/main/Week%2013

Tabular data details:

Total number of observations	45211
Total number of files	1
Total number of features	16
Base format of the file	.csv
Size of the data	3.80 MB

Name	bank-names
Total number of observations	-
Total number of files	1
Total number of features	-
Base format of the file	.txt
Size of the data	4 KB

Name	bank
Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	461KB

Name	bank-additional-full
Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv

Size of the data	5.8 MB

Proposed Approach:

1. The data looks pretty clean. These approaches are basically where were the pain points and what we are trying to solve , The columns which has two values('yes' and 'no') and slightly imbalanced such as default, loan, y, has been converted to (1,0) numerical values. rest are continuous variable were binned so that outliers value are converted into count values.
2. While approaching data cleaning method dropped the outliers and ambiguous values, such as "others" and "unknown".
3. Skewness doesn't provides much insights in data, as values of columns are nearly zero apart from 'previous'. data seems symmetrical.
4. Data has been cleaned with flooring and clapping using interquintile range(IQR) Outliers are removed by dropping values that is below 25% and 75% percentile.
5. In Feature Engineering used undersampling method to delete the samples in majority class as There are huge difference between y=1 and y=0, So the ML algorithm will omit the smaller value, which may affect the performance of algorithm.

Assumptions during data quality analysis:

1. Imbalanced dataset is a common problem in data science; however some approaches have been used on the dataset such as over and undersampling methods as well as boosting algorithm (for traditional machine learning approach) so we have used XGBOOST and Catboost algorithms
2. We can also use booting trees like adaboost and random forests that are more robust to imbalanced datasets.
3. Also a popular method for dealing with imbalanced datasets for machine learning models and deep learning models within the preprocessing phase is data augmentation.
4. We used logistic Regression model as it can provide better accuracy after providing model pipeline like min_max normalization.
5. Heterogeneous ensembling methods which combines several base model XGBoosting, Gradient Boosting, Logistic Regression and Catboost to produce final optimum solution by calculating their cross entropy loss.