Data Intake Report

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

Report date:9th 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/week10

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

Assumptions:

- After performing EDA on dataset, we have realized the importance of feature engineering and repeated testing in order to find best parameters; We have understood how hard it is to get an increase of 1% in scores. We learnt how to do feature engineering and efficiently find the best hyperparameters for classifiers. As problem is classification problem we recommend using sklearn.accuracy as the evolution metric. We recommend to use the area under the curve (AUC) and F1 score (average of precision and recall)as a metric.