Group Name: intwentytwelve

Name: Wen Li

Email: wen3@ualberta.ca

Country: Canada

College: University of Alberta

Specialization: Data Science

Problem description:

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which can help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with the bank or other Financial Institutions).

Business understanding:

ABC Bank wants to use ML models to shortlist customers whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only on those customers.

Applying ML technique will save resource and time which are directly involved in the cost (resource billing).

Project lifecycle and deadlines:

Initiation – Understanding of the problem statement and the business insights – Before 19 Nov

Planning – Choice of the ML models – Before 26 Nov

Execution – EDA and modelling – Before 9 Dec

Closure – Dashboard and presentation – Before 16 Dec

Data Intake Report

Name: Bank Marketing (Campaign)

Report date:

Internship Batch: LISUM14

Version: 1.0

Data intake by: Wen Li

Data intake reviewer:

Data storage location:

Tabular data details:

Bank:

Total number of observations	4521
Total number of files	
Total number of features	17
Base format of the file	.csv
Size of the data	367KB

Bank-full:

Total number of observations	45211
Total number of files	
Total number of features	17
Base format of the file	.csv
Size of the data	3664KB

Bank-additional:

Total number of observations	4119
Total number of files	
Total number of features	21
Base format of the file	.csv
Size of the data	482KB

Bank-additional-full:

Total number of observations	41188

Total number of files	
Total number of features	21
Base format of the file	.csv
Size of the data	4814KB

Proposed Approach:

- Mention approach of dedup validation (identification)

  I will detect if there are missing values in the dataframe and remove outliers which can make a detrimental effect on model performance.

  Some features that are highly correlated to the target variable should be removed. Because A high correlation coefficient would lead to a chance that the performance of the model will be impacted by the multicollinearity.

- Mention your assumptions (if you assume any other thing for data quality analysis)

  One of the simplest methods for detecting outliers is box plots. A box plot is a graphical display for describing the distributions of the data. Box plots use the median and the lower and upper quartiles.