Data Glacier Intern Project Report

Project: Bank Marketing (Campaign)

Group: __Data Maestros_______

Group Member 1

Name: Nrusimha Saraswati Sai Teja Jampani

Email: njampani@buffalo.edu

Country: United States

College: State University of New York at Buffalo

Specialization: Data Science

Group Member 2

Name - Purvesh Mehta

Email - mpurvesh007@gmail.com

Country - United Kingdom

University - University of Sussex

Specialization - Data Science

Group Member 3

Name: Mufunwa Nemushungwa

Email: mufunwanemushungwa@gmail.com

Country: South Africa

College/Company: University of the Witwatersrand

Specialization: Data Science

Group Member 4

Name: Aysha Abdul Azeez

Email: ayshaabdulazeez41@gmail.com

Country: United Kingdom

College/Company: University of Central Lancashire

Specialization: Data Science

Problem Description

ABC bank aims to launch a new term deposit scheme and wants to sell this product to customers. Prior to the launch, the bank plans to start a marketing campaign for the product through various marketing channels like Telephone, SMS, Emails, etc. To save time and to minimize the costs associated with this process, the bank wants to shortlist all the potential customers who have a greater possibility of buying the term deposit product.

This will help the marketing team to start a campaign on a set lot of customers without wasting their resources on any unlikely buyers. To achieve this outcome, we will need to develop a classification model with high accuracy to determine if a customer will subscribe to the term deposit or not based on the available marketing data.

Data Understanding

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y). The dataset comprises 41,188 rows and 21 columns, representing multivariate data. The variables in the dataset are separated by semicolons, and the data types include 5 float columns, 5 integer columns, and 11 object columns. The memory usage for this dataset is approximately 6.6 MB. The float columns likely contain continuous numerical data, while integer columns may represent discrete numerical values. Object columns often include categorical data or text. The memory usage indicates the amount of system memory required to store the dataset.

Checking for Problems in data

- Missing values

In the dataset with 41,188 rows and 21 columns, around 10,700 rows exhibit unknown values, affecting columns like 'job', 'marital', 'education', 'default', 'housing', and 'loan'. To address this, imputation techniques for categorical data are applied, ensuring the dataset's completeness and informativeness. Imputation, as opposed to deletion, is preferred to prevent significant data loss, preserving the integrity of the dataset for subsequent analyses and model development.

- Outliers

To identify outliers, different visualization techniques based on the data types of the variables were utilized. For numerical (integer and float) features, box plots were employed. Box plots provide a visual summary of the distribution of the data, making it easy to identify potential outliers based on their position outside the whiskers of the box. On the other hand, for categorical (object) variables, histograms were employed. Histograms help to visualize the frequency distribution of categorical data, and anomalies or irregularities in the distribution may indicate potential outliers or unexpected patterns in the data. These techniques enable a comprehensive exploration of the dataset, allowing for the detection and understanding of potential outliers across different variable types.

- Duplicating rows

Identifying and addressing duplicate rows is a crucial step in data cleaning. In your dataset, 12 duplicate rows were detected, and a decision was made to retain only the first occurrence of each duplicate, removing the subsequent duplicates. This strategy helps in preserving the uniqueness of each record while eliminating redundancy.

- Encoding Categorical Variables

In the preprocessing stage, categorical variables were transformed into a numerical format using a binary encoding approach. Specifically, for binary categorical variables where the options were 'yes' and 'no', a simple label encoding technique was applied. The label encoding replaced 'yes' with 1 and 'no' with 0, effectively converting these categorical values into corresponding numerical representations. his transformation is particularly useful in machine learning models, as it allows algorithms to interpret and process categorical information, making the data compatible with various statistical and machine learning techniques.