Data Science Intern at Data Glacier

Project: Bank Marketing Campaign (Data Science)

Week 8: Deliverables

Name: Anthony Sanogo

University: Bishop's University

Email: anthonysanogo@hotmail.com

Country: Canada

Specialization: Data Science

Batch Code: LISUM17

Date: 12 April 2023

Submitted to: Data Glacier

Table of Contents:

1. Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product, they want to develop a model that can help them understand whether a particular customer will buy their product or not based on the customer's past interaction with the bank or other financial institutions. The bank wants to use an ML model to shortlist customers whose chances of buying the product are more so that their marketing channels (tele marketing, SMS/email marketing, etc.) can focus only on those customers whose chances of buying the product are more. This will save resources and time, which is directly involved in the cost (resource billing).

2. Business Understanding

The bank wants to understand the customers' behavior and identify potential buyers to optimize their marketing efforts and increase the chances of selling their term deposit product. By developing an ML model, they want to shortlist the customers whose chances of buying the product are higher, so that they can focus their marketing efforts on those customers and save resources and time. The model will help the bank to predict the probability of a customer buying the term deposit product based on their past interactions with the bank or other financial institutions. This will allow the bank to optimize their marketing strategy and increase the chances of selling their product.

3. Project Plan

Weeks	Date	plan
Weeks 07	April 11, 2023	Problem Statement, Data Collection, Data Report
Weeks 08	April 18, 2023	Data Preprocessing
Weeks 09	April 25, 2023	Feature Extraction
Weeks 10	May 8, 2023	Building the Model
Weeks 11	May 14, 2023	Model Result Evaluation
Weeks 12	May 21, 2023	Flask Development + Heroku
Weeks 13	May 30, 2023	Final Submission (Report + Code + Presentation)

4. Data Understanding

The dataset contains information about the bank clients and their interactions with the bank. The dataset includes 21 columns, where 20 columns represent input features and the last column (column 21) is the output variable, which is the target variable that we want to predict.

The input features include the client's demographic information such as age, job, marital status, and education, as well as their financial information such as having credit in default, housing loan, and personal loan. The dataset also includes information about the client's interactions with the bank, such as the last contact duration, the number of contacts performed during the campaign, the number of days since the client was last contacted, and the outcome of the previous marketing campaign. Moreover, the dataset contains social and economic context attributes such as employment variation rate, consumer price index, consumer confidence index,

Euribor 3-month rate, and the number of employees.

The output variable is a binary variable that indicates whether the client has subscribed to a term deposit or not.

5. Problems in the Data

There are no missing values in the dataset. However, there are a few potential issues to consider:

Outliers: There are extreme values in some columns. For example, the maximum value in the duration column is 4918, while the 75th percentile is 319. This indicates the presence of outliers in the data.

Skewed data: The difference between the mean and median in some columns suggests that the data may be skewed. For instance, the mean age is 40.02, but the median is 38. Similarly, the mean duration is 258.29, but the median is 180.

Categorical variables: There are 11 categorical variables in the dataset, which need to be properly encoded before using them in the model.

To handle outliers and skewed data, one approach is to use data normalization techniques, such as z-score normalization or Min-Max scaling, which can rescale the data to a common range, so that the impact of the extreme values is minimized. Another approach could be to remove the extreme values if they are deemed to be noise or errors in the data.

To encode the categorical variables, we can use techniques such as one-hot encoding, label encoding or target encoding, depending on the nature of the variables and their importance in the model. One-hot encoding creates a binary variable for each level of the categorical variable, label encoding assigns an integer value to each level, and target encoding replaces each level with its average target value.