**Project: Bank Marketing (Campaign)**

**Week 8: Deliverables**

Name: Supin Hooda

Email: [hoodasupin@gmail.com](mailto:hoodasupin@gmail.com)

Country: Canada

Batch Code: LISUM25

Specialization: Data Science

Submission Date: 24th Oct 2023

Submitted to: Data Glacier (Individual project)

**Table of Contents**

1. Problem Description

2. Data understanding (Type of data, problems and approaches to solve the problems)

3. Github Repo link

**Problem Description**

Background:

ABC Bank is planning to launch a new term deposit product and is looking to create a predictive model to determine whether a customer is likely to subscribe to this product based on their interactions with the bank and other financial institutions. This predictive model aims to assist the bank in optimizing its marketing efforts and improving the effectiveness of its campaigns.

Objective:

The primary objective of this project is to develop a predictive model that can accurately classify customers into two groups: those who are likely to subscribe to the term deposit ("yes") and those who are not likely to subscribe ("no").

Data Source:

The dataset provided for this project contains various customer attributes and information related to the bank's marketing campaigns. These attributes will be used to build and train the predictive model.

**Data Understanding**

Types of Data:

The dataset comprises both numerical and categorical variables. Here is a summary of the variables:

1. Numerical Variables:

- `age`: Customer's age (numeric)

- `duration`: Last contact duration, in seconds (numeric)

- `campaign`: Number of contacts performed during this campaign (numeric)

- `pdays`: Number of days passed after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)

- `previous`: Number of contacts performed before this campaign and for this client (numeric)

- `emp.var.rate`: Employment variation rate - quarterly indicator (numeric)

- `cons.price.idx`: Consumer price index - monthly indicator (numeric)

- `cons.conf.idx`: Consumer confidence index - monthly indicator (numeric)

- `euribor3m`: Euribor 3-month rate - daily indicator (numeric)

- `nr.employed`: Number of employees - quarterly indicator (numeric)

2. Categorical Variables:

- `job`: Type of job (categorical)

- `marital`: Marital status (categorical)

- `education`: Education level (categorical)

- `default`: Has credit in default? (categorical)

- `housing`: Has a housing loan? (categorical)

- `loan`: Has a personal loan? (categorical)

- `contact`: Contact communication type (categorical)

- `month`: Last contact month of the year (categorical)

- `day\_of\_week`: Last contact day of the week (categorical)

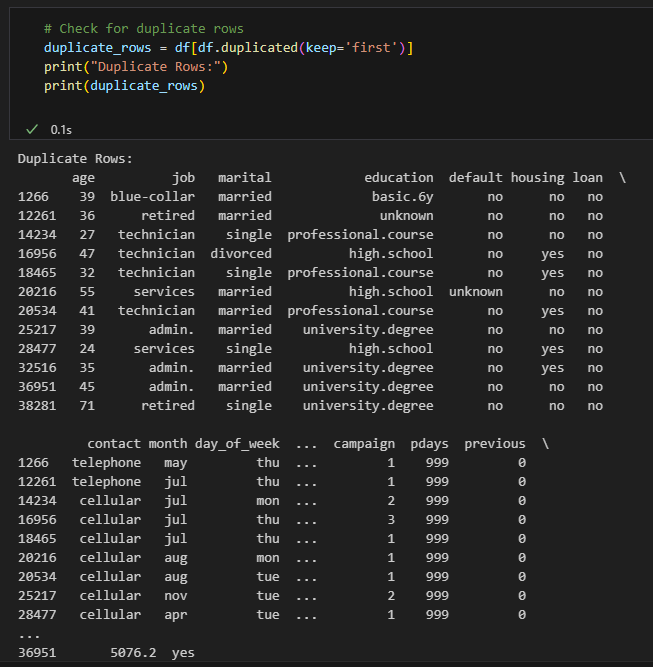
- `poutcome`: Outcome of the previous marketing campaign (categorical)

- `y`: The target variable, whether the client subscribed to a term deposit (binary: 'yes' or 'no')

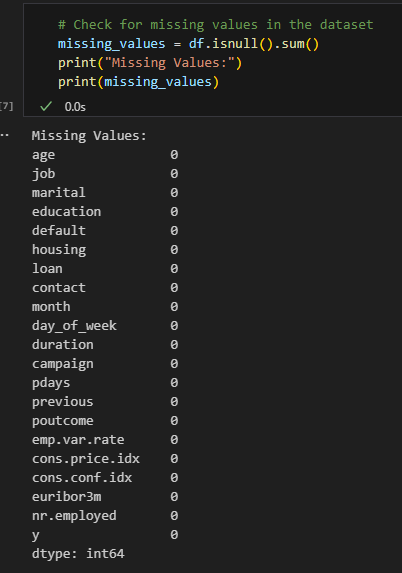


**Duplicate values**

The dataset contains a total of 41188 rows, out of which there were a total of 12 duplicate rows. The pandas drop\_duplicates method is used to drop the duplicate rows.



**Missing Values**: There are no missing values in the dataset.



**Outlier detection:** Outliers are detected using box plots

Outliers can significantly affect the performance of a predictive model. In this dataset, outliers were detected using box plots. The following features were found to have outliers:

i. Age:

- The maximum value of age is 98, which appears to be realistic. Therefore, no action was taken, and these data points were not dropped.

ii. Pdays:

- Pdays represents the number of days that passed after the client was last contacted from a previous campaign. The maximum value of pdays is 999, which, according to the dataset description, means the client was not previously contacted. Moreover, around 96% of rows contain this value. Therefore, dropping rows containing 999 seems unrealistic, and these data points were retained.

iii. Campaign:

- Campaign represents the number of contacts performed during this campaign and for this client. The statistics show that the maximum value is given as 56, which is clearly an outlier. Numbers for 'campaign' above 20 accounts for around 0.38% of the data. To address this, the suggestion is to impute those rows with the average of campaign values, which helps mitigate the impact of extreme outliers.

iv. Previous:

- Previous holds the number of contacts performed before this campaign and for this client. The maximum value of 7 does not seem to be a noise, so no action was taken to remove or modify these data points.

In summary, the approach to dealing with outliers is to consider the nature of the feature, its statistical characteristics, and its potential impact on the model. In some cases, it may be appropriate to retain outliers, while in others, it's advisable to impute or modify the extreme values to prevent them from adversely affecting the model's performance. The specific action taken depends on the context and domain knowledge.