Project: Bank Marketing (Campaign)

Week 9: Deliverables

Name: Supin Hooda
Email: hoodasupin@gmail.com
Country: Canada
Batch Code: LISUM25
Specialization: Data Science
Submission Date: 6th Nov 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')

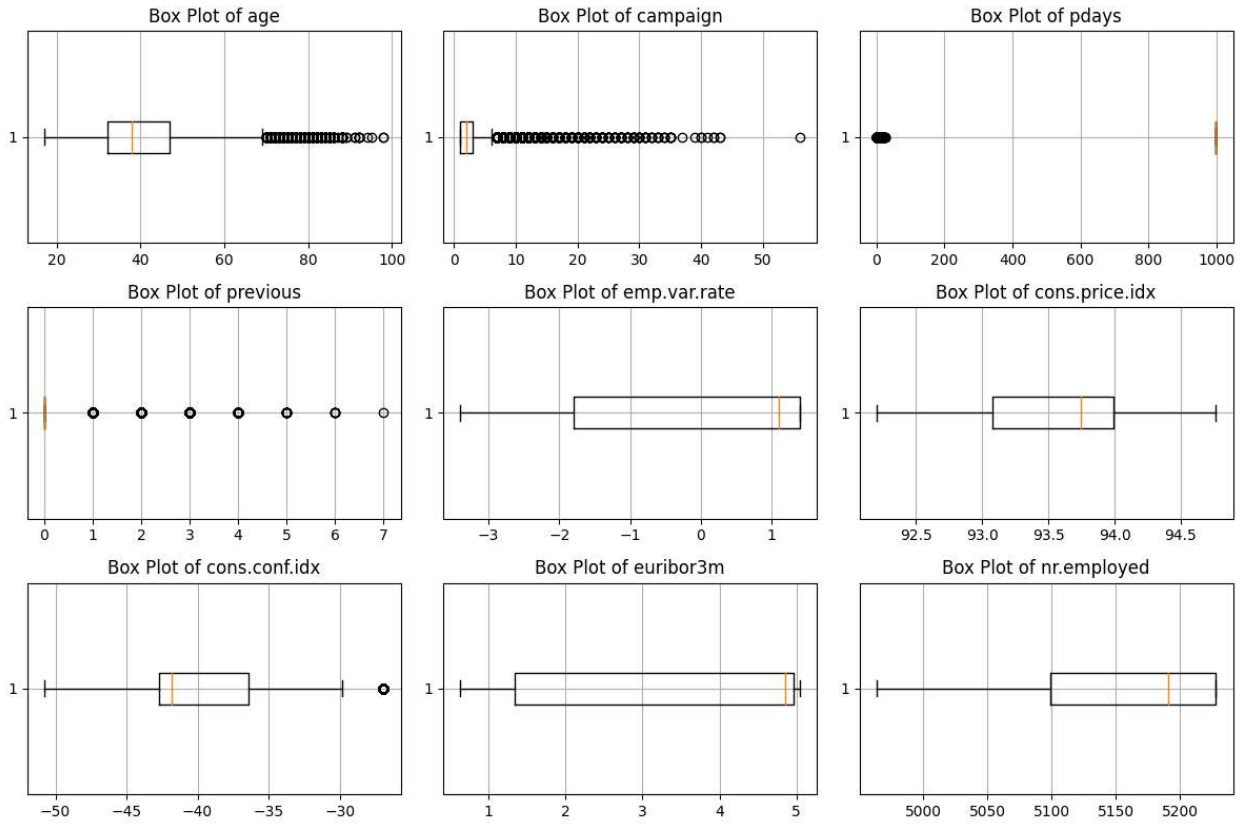
Problems in the Data

The dataset contains 41,188 rows, with 12 duplicate rows identified and removed using the `drop_duplicates` method. Fortunately, there are no missing values in the dataset.

Outlier Detection

Outliers in the data are detected using box plots. Outliers are data points that deviate significantly from the central distribution of data, impacting key statistics such as mean and mode. Managing outliers during data cleaning is essential to ensure that model performance is not compromised.

Outliers are observed in the following features: "age," "campaign," "pdays," and "previous." These outliers are identified as data points located outside the box plot whiskers.



Approaches to Overcome the Problems

- **Age**: The maximum age value of 98 appears to be realistic and is retained in the dataset.

- **Pdays**: The maximum value of 999 indicates that the client was not previously contacted, and it's present in around 96% of rows. Thus, it's considered unrealistic to drop rows with this value, and they are retained.

- **Campaign**: The "campaign" feature, which denotes the number of contacts made during the campaign, contains a maximum value of 56, which is considered noise. The portion of records with "campaign" values exceeding 20 is approximately 0.38%. It is proposed to impute these rows with the average of campaign values.

- **Previous**: The "previous" feature represents the number of contacts made before this campaign. The maximum value of 7 does not appear to be an outlier, and it is retained.

Treating Outliers

Following outlier detection, two techniques were applied for handling outliers: median and mean imputation. Both approaches resulted in similar statistics for the dataset, with minimal variation.

Github Repo Link

The project repository can be accessed at [GitHub Repo](
https://github.com/supinhooda/assignment/tree/main/Week9).