Data Science Project: Bank Marketing (Campaign)
Week 8: Deliverables

Name: Trofymova Anastasiia
Email: anastasiia.trofymova@gmail.com
Country: United States
Specialization: Data Science
Batch Code: LISUM28
Date: 26 January 2024
Submitted to: Data Glacier

Problem Description:

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

Data Understanding:

The dataset selected for our analysis is named "bank-additional-full.csv." This dataset comprises 41,188 observations and encompasses 21 features. These features encompass various aspects of clients' basic information, including age, job, marital status, education, credit in default, housing, and loan.

Additionally, the dataset contains details about contact, such as communication type, last contact month, last contact day, last contact duration, and the number of contacts. Furthermore, it includes information about marketing campaigns, including outcome, employment variation rate, consumer price index, consumer confidence index, euribor 3-month rate, and the number of employees.

Among the dataset's variables, there is a target variable denoted as "y." This variable represents the answer to the yes-no question, "Has the client subscribed to a term deposit?" The target variable will be utilized in future predictions.

Types of the Data:

Feachure Name	Data Type	Categorical/Numerical
age	integer	numerical
job	string	categorical
marital	string	categorical
education	string	categorical
default	string	categorical
housing	string	categorical
loan	string	categorical
contact	string	categorical
month	string	categorical
day_of_week	string	categorical

duration	integer	numerical
campaign	integer	numerical
pdays	integer	numerical
previous	integer	numerical
poutcome	string	categorical
emp.var.rate	float	numerical
cons.price.idx	float	numerical
cons.conf.idx	float	numerical
euribor3m	float	numerical
nr.employed	float	numerical
y	string	categorical

Problems in the Data:

- The dataset reveals the presence of missing data in six categorical features: job, education, marital status, default, housing, and loan.
- Additionally, there is one numerical feature, "duration," exhibiting outlier data. Notably, the mean duration is approximately 258, but the maximum value is 4918, signaling the presence of outliers.
- Moreover, the dataset displays an imbalance, particularly in the target variable for the predictive classification model, where 90% of the cases are skewed towards the "N" category.

Approaches:

To address missing (NA) values, we will employ tailored techniques based on the severity of each column and its impact on the dataset.

- For features with fewer "unknown" data points, such as "marital" and "job," we will opt for dropping the missing data.
- For "housing" and "loan," we will replace missing data with the most frequent category.
- In the case of "default" and "education," missing values will be addressed using a machine learning classification model.

To address outlier numerical data, we will utilize an upper outer fence defined as three times the interquartile range (3IQR). This approach retains 97% of the original data.

Addressing the imbalance in the target variable, we will select an appropriate evaluation metric, likely utilizing the AUROC curve to identify models with optimal results for True Positive and False Negative predictions. Given the dataset's size, potential strategies include under-sampling from the majority case, ensuring rare cases are consistently retained during data splitting, or adjusting the ratio of rare to majority cases in the training data to over-represent the rare case for the model.