# Capstone Project Submission

| |
|---|
| **Team Member's Name and Email :** |
| Name : Ajit Sharad Mane<br><br>Email : ajitmane36@gmail.com<br><br>Contribution – Individual |
| **GitHub Repo link :** |
| Github Link :- https://github.com/ajitmane36/Bank-Marketing-Effectiveness-Prediction-ML-Classification.git |
| **Summary :** |

# Bank Marketing Effectiveness Prediction

This project focuses on utilizing machine learning techniques to predict the effectiveness of bank marketing campaigns. The data used in the project is provided by a Portuguese banking institution and includes input variables such as age, job, marital status, education, and balance etc.

The goal of this project is to develop a classification model that can accurately predict the effectiveness of bank marketing campaigns. Through the use of machine learning algorithms and techniques, the model will be able to classify a client's response to a campaign as either positive or negative. This project will provide insight into how different input variables can affect the effectiveness of bank marketing campaigns and help banks better target their customers.

The data set contained details about bank marketing campaigns. Descriptive statistics were computed for each variable as part of the analysis, and visualizations were made to investigate the relationships between the various variables. We created a number of graphs, such as a distplot, count plot, bar plot, pair plot, heatmap, and boxplot, to gain insight from the dataset.

There are 45211 observations in this dataset and 17 columns with the following names: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, and y (target variable). There are no duplicate values in this dataset. The 10 categorical variables in this dataset are: employment, marital, education, default, housing, loan, contact, month, poutcome, and y. This dataset contains seven numerical variables: age, balance, day, duration, campaign, pdays, and prior. For the variables job, education, contact, and poutcome, the number of unknown tagged values are 288; 1857; 13020; and 36959, respectively. Unknown tagged values can be treated as null since they are not defined and can be taken out of features by treatment.

Replaced null values with their respective modes for features like contact, education, and job. Moreover, features with more than 50% null values were eliminated because they were useless and negatively impacted model performance. Outliers are treated using the interquartile range for the variables age, balance, duration, campaign, p-days, and previous.

After doing univariate, bivariate, and multivariate analyses, we discovered insights that are as follows

- The average client is between the ages of 25 and 60, but the majority of bank term deposits are made by clients between the ages of 30 and 36.
- Most clients with blue-collar jobs do not subscribe to bank term deposits (20.52%), but most clients with managerial jobs do (2.88%).
- Most of the clients are married. Clients who are married are the most likely to subscribe to term deposits, and they are also the least likely to subscribe to term deposits.
- Most of the clients are married. Clients who are married are the most likely to subscribe to term deposits, and divorced clients are less likely to subscribe to term deposits.

- Clients who are more educated than the primary are more likely to sign up for a term deposit.
- Most of the clients who subscribed to term deposits have no credit in default.
- The majority of clients who have signed up for a term deposit do not have any housing loan.
- If a client has a housing loan, there is a 51% chance that they will not subscribe to a term deposit.
- Clients are more likely to subscribe to the term deposit if they do not have any personal loans.
- If the client has a personal loan, there is a greater chance that they will not subscribe to a term deposit.
- The clients who were contacted with celluler are mostly subscribed to term deposits.
- Less than one percent of total clients contacted per day subscribe to term deposits.
- In May, June, July, August, and April, more than 1 percentage of clients subscribed to the term deposit, but other than this month, less than 1 percentage of clients subscribed to the term deposit.
- In June, July, August, and April, more than 1 percentage of clients subscribed to the term deposit, but other than this month, less than 1 percentage of clients subscribed to the term deposit. May's subscriber rate is more than double that of the other months of the year, a difference of more than 2 percentage.
- No one has signed up for term deposit if they have received more than three phone calls. Less than three times contacted clients who signed up for term deposits.
- Only 11.7% of total clients sign up for term deposits, which means that there is an 88.3% chance that clients will not subscribe to term deposits**.**
- Most clients who have management-related jobs and a tertiary degree have subscribed to the term deposits.
- Customers with a secondary education are the second most likely to subscribe to term deposits.
- Clients are more likely to subscribe to term deposits if they spend more time on the phone.
- Average of 400 seconds required to convey clients' intent to subscribe and make a term deposit
- A customer is more likely to sign up for a term deposit if he is entirely debt-free.
- Customers are less likely to choose a term deposit if they already have both types of loans.

Label encoding used with just a few categories for the categorical variables marital, education, default, housing, loan, contact, and y. For the categorical variables job and month, which contain many categories, one hot encoding utilized. Detect a class imbalance beacause discovered that the number of clients who subscribed to term deposits is 11.7% lower than the number of clients who did not (88.3%). Class imbalance was handled successfully using the Synthetic Minority Oversampling Technique (SMOTE). MinMaxScaler used to scale the dataset to the same length.

Models trained on the dataset include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machine, XGBoost, K Nearest Neighbor, Naive Bayes, Support Vector Machine, and Artificial Neaural Networks. The performance of each algorithm improved after being trained with cross validation after being fitted twice using train test split and cross

validation.

The XGBoost classification model trained using cross-validation is the ideal model and well-trained for predicting whether the client will subscribe to a term deposit or not due to its high accuracy (0.93), precision (0.93), recall (0.93), F1 score (0.93), and rou auc score (0.93), which is close to 1.

For features like housing, month_jun, and month_jan, a higher feature importance score means that those variables have a greater impact on the model's prediction of whether or not a client will sign up for a term deposit. We explained the model using the SHAP technique. We can draw the conclusion that lower values of the majority of the input features have a positive impact on the model's prediction, whereas higher values of the majority of the input features have a negative impact.

We faced a challenge when building models: that model may require extensive tuning and testing to ensure the accuracy and reliability of the predictions.

: