**Abstract :**

This document presents the development of a machine learning (ML) model for predicting the effectiveness of bank marketing campaigns. The model was developed using a dataset from a Portuguese banking institution that was collected from direct marketing campaigns.

The dataset contains a variety of features, including customer demographic and behavioural attributes. The ML model was trained using the different algorithms and evaluated using the accuracy, F1-score, precision, recall, and roc\_auc\_score metrics.

**Table of contents :**

* Abstract
* **Table of contents**
* **List of terms**
* **List of terms**
* **Acknowledgments**
* Problem Statement

#### **Introduction**

#### **Background**

#### **Materials and apparatus**

#### **Procedure**

#### Conclusion

**List of terms :**

Feature engineering :- the pre-processing step of machine learning, which extracts features from raw data.

Univariate analysis :- Univariate analysis explores each variable in a data set, separately.

*Bivariate analysis* :- The bivariate analysis is will measure the correlations between the two variables.

*Multivariate analysis* :- is a Statistical procedure for analysis of data involving more than two types of measurement or observation.

**Acknowledgments :**

Thank you AImabetter for your help and support in completing my ML project. Your guidance and advice was invaluable and I am grateful for all the help you provided.

**Problem Statement :**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

### ***Attribute Information*** :

* age: age of client
* job : type of job
* marital : marital status
* education: qualification of client
* default: has credit in default? housing: has housing loan? (categorical: 'no','yes','unknown')
* loan: has personal loan?
* contact: contact communication type
* month: last contact month of year
* day: last contact day of the month
* duration: last contact duration, in seconds
* campaign: number of contacts performed during this campaign and for this client
* pdays: number of days that passed by after the client was last contacted from a previous campaign
* previous: number of contacts performed before this campaign and for this client (numeric)
* poutcome: outcome of the previous marketing campaign
* y (target Variable) - has the client subscribed a term deposit? (binary: 'yes','no')

#### ****Introduction :****

This machine-learning model predicts whether clients will sign up for term deposits or not. Exploratory data analysis was computed, and some useful insights were discovered using various graphs. Different algorithms are implemented to get the optimal model for accurate predictions.

**Background :**

The idea behind this model is to assist the bank in setting targets for which clients will subscribe to term deposits to save both time and money.

**Materials and apparatus:**

*Materials:*

1. Computer: A desktop or laptop computer with an operating system capable of running machine learning algorithms is essential for any machine learning project.
2. Data: Datasets are an essential component of any machine learning project. The type of data required for a specific project will depend on the type of machine learning task it is being used for.
3. Software: Machine learning algorithms need to be implemented in a programming language. Popular choices are Python used.
4. Libraries: Libraries are collections of pre-written functions that can be used in the implementation of machine learning algorithms. Popular libraries include scikit-learn used.

*Apparatus:*

1. GPU: Graphics processing units (GPUs) are specialized hardware used to speed up the training process of machine learning algorithms. GPUs can significantly reduce the time required to train a machine-learning model.

**Procedure :**

**[1] Import, Loading and** Inspection of Data:

After importing the dataset, we look at its columns and shape. The info() method is used to verify variables and associated datatypes for null values. Using the describe() function, we can determine the fundamental characteristics of each *variable, such as the mean, median, count, and so on.*

We can better comprehend the meaning of the variable thanks to the supplied variable description. This helped us understand datasets.

[2] Handling duplicated values :

Fortunately, there aren't any duplicate values in the dataset, but if there are, you can get rid of them with the drop\_duplicated() method.

[3] Handling null values :

There are unknown tagged values in some features that are considered null. Those null values are replaced with modes and features with more than 50% null values, which are useless and are removed from the dataset.

[4] Handling outliers :

Boxplot and distplot are used to detect outliers.The interquartile range approach is used to eliminate outliers from data*.*

[5] Feature engineering and data wrangling:

Using feature engineering, we generate new variables from the original one.

[6] Exploratory data analysis :

We use a count plot, bar plot, line plot, heatmap, box plot, and distribution plot for exploratory data analysis.

Univariate, bivariate, and multivariate analyses were performed using statistical techniques, and the results revealed insightful information.

[7] Feature Encoding :

Label encoding and one-hot encoding techniques are used to convert a categorical variable to an integer that a machine learning model can understand. Label encoding is used for variables having very few categories, and one hot encoding is used for variables having many categories.

#### **[8] Data Scaling :**

#### Data scaling is carried out to make it simple for a model to understand and learn the problem. The original distribution's shape is maintained by MinMaxScaler. The data's original information is not significantly altered. Thus we used MinMaxScaler to scale the dataset.

#### [9] Fitting different ML models :

#### On datasets, algorithms such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machine, XGBoost, K Nearest Neighbor, Naive Bayes, Support Vector Machine, and Artificial Neural Network are used to build the model.

#### The several classification models were trained using the Sklearn library, and predictions were made using the test dataset. Model fitting is carried out using GridSearchCV to improve the model's accuracy and predictability.

[10] Evaluation of Model :

Metrics including accuracy, precision, recall, F1-score, confusion matrix, and roc\_auc\_score were used to further assess the model. Model XGBoost tops all classification evaluation metrics among all different implemented models.

#### **[11] Feature Importance :**

#### The presence or absence of a housing loan has a significant impact on the model output used to predict whether a client will subscribe to a term deposit or not.

#### A higher feature importance score for features like housing, month\_jun, and month\_jan, indicates that those specific features have a greater influence on the model output used to predict whether or not a client will subscribe to a term deposit.

[12] Model Explainability :

In order to give additional insight into how an ML model arrived at its final output, SHAP was used to explain the model.

It describes which variables affect output favorably or unfavorably, which variables are crucial for generating output, and which variables are useless for the model.

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.

**Conclusion :**

We tested numerous machine-learning models and assessed them using different classification evaluation metrics. The XGBoost classification model comes the closest to having the highest scores on all classification evaluation metrics.

The XGBoost classification model is the perfect and well-trained model for predicting whether clients will subscribe to term deposits or not because of the model's high accuracy, precision, recall, and roc\_auc\_score.