**Capstone Project**

**Supervised Ml-Classification**

**Bank Marketing Effectiveness Prediction**

## **Project Overview**

This project discusses the prediction model of Bank Marketing Effectiveness of a Portuguese Marketing institution. The marketing campaigns were based on phone calls.The classification goal is to predict if the client will subscribe to a term deposit.

First we explore the data, cleaned and preprocessed the data and then we performed the exploratory data analysis to extract information, in which we identified certain trends, relationships, correlation and found out the features that had some impact on our dependent variable and plotted the graph to visualize the impact on dependent variable. We also encoded the categorical variables.

We build the various machine learning algorithms on our split and standardized data. We tried different algorithms namely;**Logistic Regression,Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, K Neighbors Classifier, XG Boost and Naive Bayes Classifier.** We did hyperparameter tuning and evaluated the performance of the model .

We analyze the data and build the model by considering the below

**Problem Description**

### The data is related to 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 assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

## 

## **Data Description**

## **Input variables:**

### Bank Client data:

* age (numeric)
* job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
* marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
* education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
* default: has credit in default? (categorical: 'no','yes','unknown')
* housing: has a housing loan? (categorical: 'no','yes','unknown')
* loan: has a personal loan? (categorical: 'no','yes','unknown')

### Related with the last contact of the current campaign:

* contact: contact communication type (categorical: 'cellular','telephone')
* month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
* day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
* duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

### Other attributes:

* campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
* pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
* previous: number of contacts performed before this campaign and for this client (numeric)
* poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

### Output variable (desired target):

* y - has the client subscribed a term deposit? (binary: 'yes','no')

**Model Evaluation Metrics**

We Evaluated the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result. And got the results from the best suitable model for our project. Following are the evaluation metrics for our selected model:-

● Confusion matrix :- A confusion matrix is defined as the table that is often used to describe the performance of a classification model on a set of the test data for which the true values are known.

● Accuracy :- Accuracy simply measures how often the classifier correctly predicts.

● Precision :-Precision explains how many of the correctly predicted cases actually turned out to be positive

● Recall (Sensitivity) :- Recall explains how many of the actual positive cases we were able to predict correctly with our model.

● F1 Score :- It gives a combined idea about Precision and Recall metrics.

● Receiver Operator Characteristic (ROC)

● Area Under the Curve (AUC)

**Conclusion**

* Blue-collar, management and technician showed maximum interest in subscription.
* Divorce people have no interest in term deposits.
* People with secondary and tertiary education were more driven towards paying term deposits in banks.
* Generally people who don't have credit in default are interested in a deposit. Majority of the people have a home loan but only a few of them opted for a term deposit.
* Cellular communication is more effective in comparison to other communication types.
* There were maximum subscriptions in the Summer season.
* The calls with large duration have more tendency for conversion. People were mostly contacted only once.
* Majority of people were not contacted previously before this campaign and there are no significant contacts after 11 times already done.
* Success rate is high for unknown outcomes.
* We can choose our model either **Gradient Boosting Classifier, Random Forest Classifier and XG boost** to predict Effectiveness as they are showing maximum accuracy

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Logistic Regression** | **Random Forest Classifier** | **Decision Tree Classifier** | **Gradient Boosting Classifier** | **K Neighbour**  **Classifier** | **XGBoost** | **Naïve Bayes Classifier** |
| **Accuracy** | **0.87** | **0.89** | **0.85** | **0.88** | **0.87** | **0.88** | **0.80** |

**Contributor Roles**

1. **Yashwant B. Raul:- (**[**yashwantraul24@gmail.co**](mailto:yashwantraul24@gmail.co)**m)**

* Data Wrangling
* Feature Analysis
* Feature Engineering
* Model Building (Logistic Regression,Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, K Neighbors Classifier, XGBoost, Naive Bayes Classifier)

1. **Mayur S. Marathe:- (**[**marathemayu1990@gmail.com**](mailto:marathemayu1990@gmail.com)**)**

* Accessing the data
* Data visualization
* Relation of “y” with numerical variables
* Converting the categorical variables in numeric variables

1. **Sanket Gawali:-(sanketgawali23gmail.com)**

* Data visualization
* Observing and implement Categorical and Numerical data
* Converting string data to float data
* Train Test Split

**Github Link**

https://github.com/YashwantRaul/Supervised-ML-Classification

**Drive Link**

https://drive.google.com/drive/folders/1fpsMXNsPAooxneuiqP62116s0uuB7Rw9?usp=sharing