**Bank Marketing Effectiveness Prediction**

**Prabhat Patel, Annayan Bose**

**Data Science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Globalization and technological advancement has created an extremely competitive market. This has also affected the banks. In recent years, banking and direct database marketing have become an important strategy for understanding customer needs. The success rate of banking marketing depends on the achieved results and decisions. In order to make more accurate predictions, statistical tools and methods have been used.

This report examines how to use machine learning techniques to analyze and make predictions in banking marketing using existing dataset. The purpose of building the models is to predict whether the client will subscribe for a term deposit.

**Problem statement:**

### The data was 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 was to predict if the client would subscribe to a term deposit (variable y).

**Introduction:**

In the last few decades, machine learning (ML) has grown into one of the most significant IT and Artificial intelligence (AI) branches. This is a specific sub-group of AI based on the idea that the machine can learn by identifying patterns and make predictions in various data problems with minimum human intervention. Machine learning is a data analysis method that is widely used in various business and industrial sectors. The main reason for that because ML can build predictive models to produce better predictions and achieve the desired level of accuracy, leading to better outcomes

The aim of the project was to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing. To find how they could be used together in a process of converting raw data to effective decision making knowledge. Building the predictive models would help to predict whether the client would subscribe for a term deposit.

This report will describe the different stages of preparation and implementation of the predictive models, starting with EDA(Exploratory Data Analysis) on dataset and then implementing various machine learning techniques; in particular, logistic regression, Decision tree, Random Forest, XgBoost etc.

**Dataset and Feature:**

**Banking Dataset** -

The dataset consists of direct marketing campaigns data of a banking institution which consists of 45211 data points with 17 independent variables out of which 7 are numeric features and 10 are categorical features. The list of features available to us were given below:

## **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: 'primary', 'secondary', 'tertiary', '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 to a term deposit? (binary: 'yes', 'no')

# **Exploratory Data Analysis:**

# Jumping into the modelling part knowing nothing about the problem is not a good idea. So as a start, we did some EDA to find out more about what data we were dealing with, if there was any pattern to the data. So some of the important insights are:

* The dataset was imbalanced, where the number of negative classes is close to 8 times the number of positive classes.
* The customers who had a job of admin had the highest rate of subscribing a term deposit, but they were also the highest when it came to not subscribing. This is simply because we had more customers working as admin than any other profession.
* Majority of the customers were married. Followed by Single, divorced and unknown.
* Majority of the customers had a housing loan.
* Maximum number of people between falling between the range of 30 and 50.
* Maximum number of people have balance less than 3000 in their account very few people have balance greater than 5000.
* Maximum number of people were contacted during this campaign was less then equal to 5 times.

# **Data Pre-Processing:**

As our dataset had 10 categorical features so we were needed to encode these features into a numerical representation to apply the machine learning models.

**Label encoding**: This is one of the most popular encoding schemes to deal with the categorical features. For each unique category it provides a unique numeric value (0,1,2etc.).

## **Dealing with Missing values:**

Thankfully in our dataset there were no missing values. If there were any missing values we would have either removed the rows (if the number of missing values were less compared to the amount of data we have).

**Feature selection**: Also as it was already mentioned in a problem statement that ‘duration’ column should be dropped for realistic model prediction. So we dropped it as it would have highly affected our prediction. Performed Sample model decision tree and visualized feature importance graph, from whichever features are important we took it into consideration.

**Splitting the data:**

Now it is very important to split your dataset into train, test datasets. So we did split on 80-20 rule, that means 80% for training and 20% for testing.

## **Sampling:**

As we talked earlier also that our given dataset was highly imbalanced, and so to balance this we used the technique called sampling. Firstly, we performed over-sampling technique called **SMOTE** so that our dataset could be balanced.

**SMOTE:** Synthetic Minority Oversampling Technique is one of the most commonly used oversampling methods to solve the imbalance problem.

It aims to balance class distribution by randomly increasing minority class examples by replicating them.

SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

**Models Implementation:**

So basically we used various models to pick up the best fit for a given dataset. The models implemented are:

* **Logistic Regression**
* **K-NN**
* **Naive bayes classifier**
* **Random Forest**
* **XGBoost**

# **Performance Metric Used:**

The performance metric used for this case study is AUC ROC score also known as AUCROC(Area Under the Receiver Operating Characteristics) and Confusion matrix which consists of accuracy, precision, recall and F1 score.

**Model’s performance :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Test AUC** | **Test Accuracy** | **F1\_score** | **Precision** |
| **Logistic Regression** | **0.83** | **0.76** | **0.77** | **0.74** |
| **KNN** | **0.91** | **0.84** | **0.83** | **0.87** |
| **Naive bayes classifier** | **0.82** | **0.72** | **0.76** | **0.66** |
| **Random Forest** | **0.96** | **0.91** | **0.91** | **0.90** |
| **XGBoost** | **0.95** | **0.89** | **0.88** | **0.90** |

After implementing various models and comparing their scores as shown in the above table ,we found out that **Random Forest** and **XGBoost** were giving the best result.

**Conclusion:**

* As the given dataset was binary type classification so the normal score like accuracy ,precision etc. were not giving that good understanding and clarity so we had used **AUC** score which was able to give out the good understating and prediction.

**Feature importance:**

* 'poutcome' was the most important feature which means folks who had a good balance were more likely to take a term deposit. So they should be more focused.
* 'Default' was the least important feature. This means that bank should not focus much on the default credit client is having or not.
* Month of May had seen the highest number of clients contacted but had the least success rate. Highest success rate was observed for the end month of the financial year as well as the calendar year. So one can say that our dataset had some kind of seasonality.

**References:**

* <https://towardsdatascience.com/>
* <https://www.geeksforgeeks.org/>
* <https://stackoverflow.com/>