**Bank Marketing Effectiveness Prediction**

**NIKHIL MACHAVE AISHWARYA METHE ADITYA TADAS**

**Data Science Trainees**

# Abstract

Finance industry is one of the leading industries globally and have the potential to bring huge impact in the growth of nation. Thus, it is important to analyze the data or information that banking sector records about the clients. This data can be used to create connection and keep professional relationship with the customers in order to target them individually for any banking schemes. Usually, the selected customers are contacted directly through: personal contact, telephone cellular, email or any other means of contact to advertise the new services or give an offer. This kind of marketing is called direct marketingand is one of the leading marketing techniques.

Thus, in this project we trained a model that can predict that whether the client will opt for a term deposit or not using given bank-client data, data related with the last contact of the current campaign and some other useful

attributes

***Keywords: Supervised Machine Learning, Classification, Predictions, duration.***

# 1. Problem Statement

The given dataset is of a direct marketing campaign (Phone Calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (Target variable y).

We were provided with following dataset:

**Bank Client data:**

* age (numeric)
* job : type of job
* marital : marital status
* education
* default: has credit in default?
* housing: has housing loan?
* loan: has personal loan?

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

* contact: contact communication type
* month: last contact month of year
* day of week: last contact day of the week
* duration: last contact duration, in seconds (numeric).

**Other attributes:**

* campaign: number of contacts performed during this campaign
* 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
* poutcome: outcome of the previous marketing campaign

**Output variable (desired target):**

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

# 2. Introduction

Marketing is the most common method which many companies are using to sell their products, services and reach out to the potential customers to increase their sales. Telemarketing is one of them and most useful way of doing marketing for increasing business and build good relationship with customers to get business for a company. It’s also important to select and follow up with those customers who are most likely to subscribe product or service. There are many classification models, such as Logistic Regression, Decision Trees, Random Forest, KNN, ANN and Support Vector Machines (SVM) that can be used for classification prediction.

# 3. Classification Approach

After understanding the problem statement, we loaded the dataset for following operations:

* **Data Exploration**
* **Exploratory Data Analysis**
* **Feature Engineering**
* **Feature selection**
* **Balancing Target Feature**
* **Building Model**
* **Hyperparameter Tuning**

* 1. **Dataset Exploration**

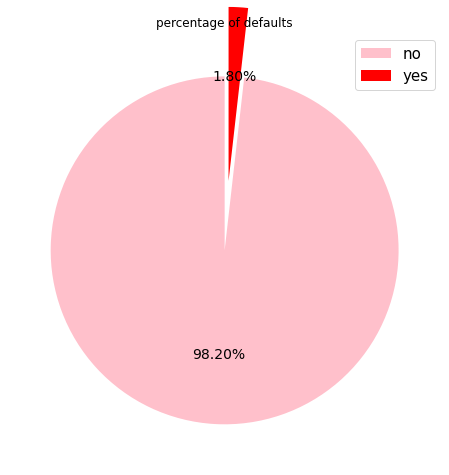
The given dataset was initially loaded for a quick overview. It was observed that our dataset contains 45211 records and 17 features. Datatypes of features was then checked and it was found that there are 7 numerical (int) and 10 Categorical (object) datatypes among which no null values and duplicated records were found in our dataset.

* 1. **Exploratory Data Analysis**

After data wrangling, we did univariate and multivariate analysis on features to understand their pattern and how they relate with target class.

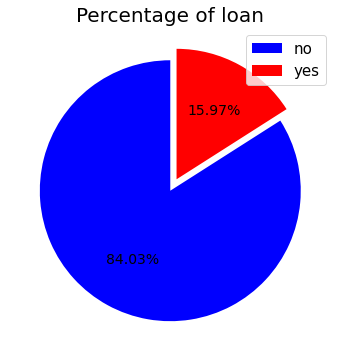
1. **Univariate Analysis**

a)



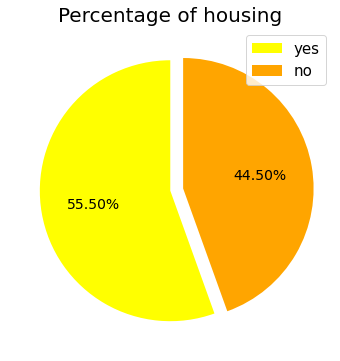
From above pie chart 98.20 % clients have not any credit default only 1.80% of clients have default in credit.

b)



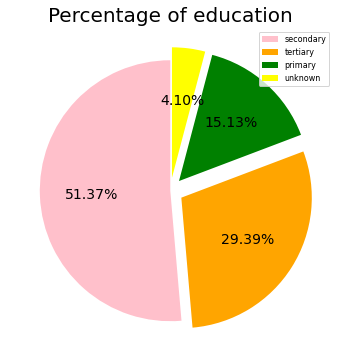
83.98 % of clients having personal load and 16.02 % clients have not any personal loan.

c)



55.58 % of clients having housing loan and 44.42 % of clients doesnot having any housing loan

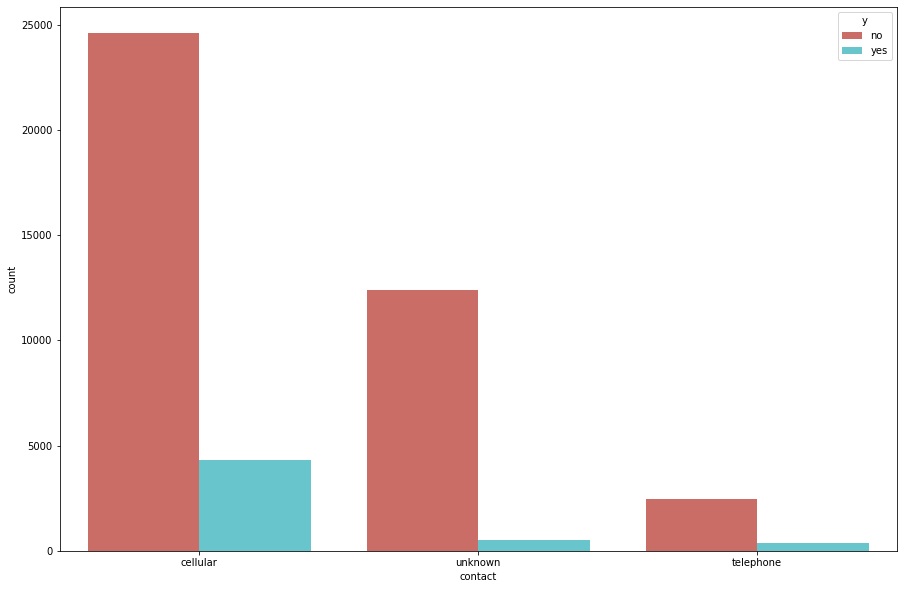
d)



51.32 % of clients are in the category of secondary education.

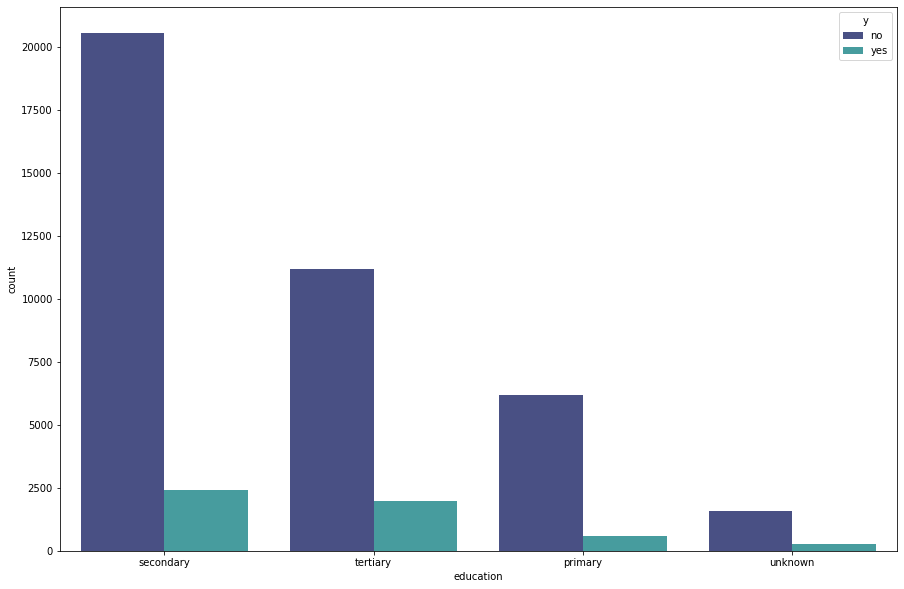
% of clients who are in the category of primary education(15.15%) is very less.

1. **Bivariate Analysis**



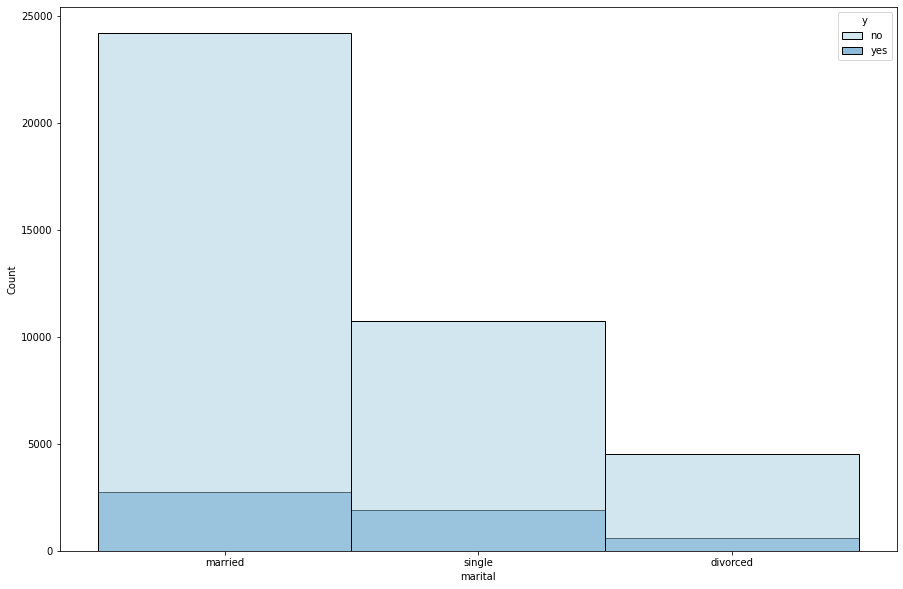
From the above count plot we can conclude that when contact communication type is cellular then there is high possibility that the client subscribe a term deposit hence the bank should contact the customer by cellular type mostly.

when the contact communication type is telephone then there was very less possibility that the client subscribe a term deposit.



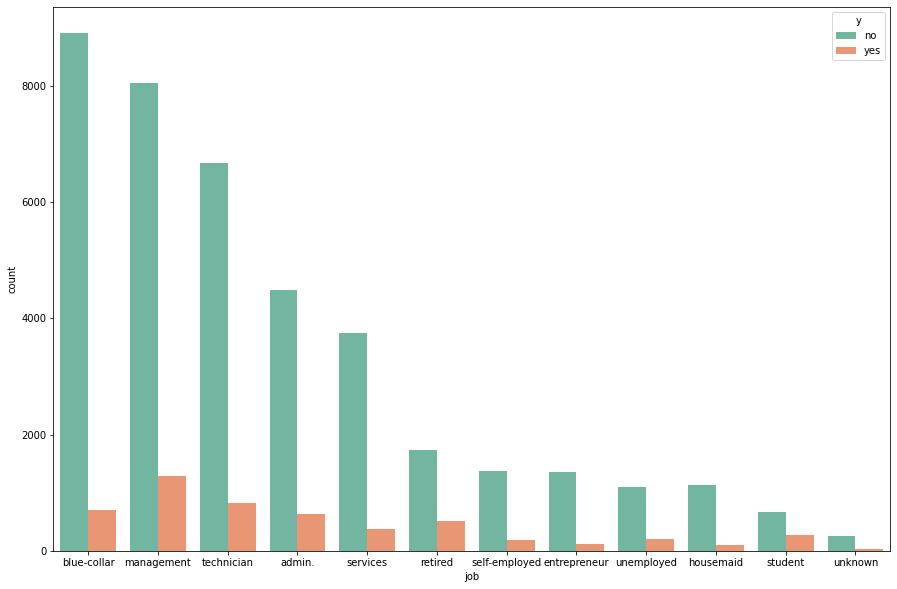
From the above bar plot we can conclude that when the customer education is tertiary and secondary then there is a high possibility that client subscribe a term deposit hence bank should approach mostly to the tertiary and secondary class education client to subscribe for term deposit.

When the education of the customer is unknown and primary those client have very low possibility to subscribe for term deposit.



Most of the clients who are married and single had subscribed for the term deposit therefore , When marital status of client is 'Single' and 'married' then there are high possibility that those clients subscribe a term deposit .Bank should target 'single' and 'married' client both to subscribe for term deposit.

when clients marital status was devorced those clients did not subscribe for the term deposit much that’s why , When the client marital status is divorced then there is very less chance that these clients agrees to subscribe for term deposit.



Most of clients are from the job called as 'blue collar, management, technician and admin, when the client jobs are Management , technician, blue\_ collar, admin services the there is high chance that those customers subscribe for term deposit so that bank should prefer salaried persons most to approach for term deposit.

when the client is retired person we can see high probability to subscribe term deposit hence retired client has high possibility that they subscribe for term deposit bank should communicate mostly to retired person to subscribe for term deposit.

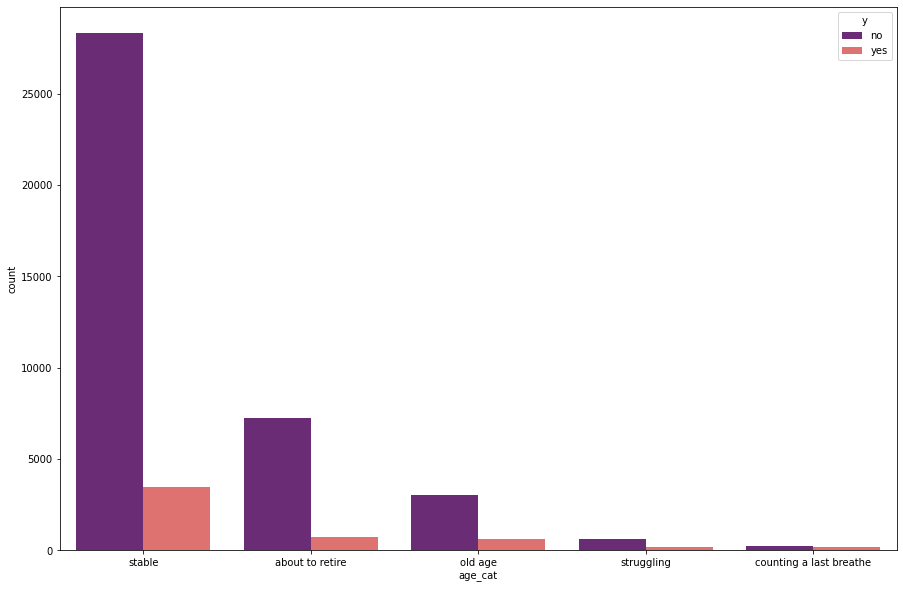
when a clients are self employed and entrepreneur we can see less probability for subscribe to term deposit as well as when a clients have a category house maid , unemployed and student and unknows there are least possibility that those customers agree to subscribe for term deposit.

**3.3 Feature Engineering** Feature engineering is one of the important steps in model building and thus we focused more into it. We performed the following in feature engineering

**a) Dealing with outliers**

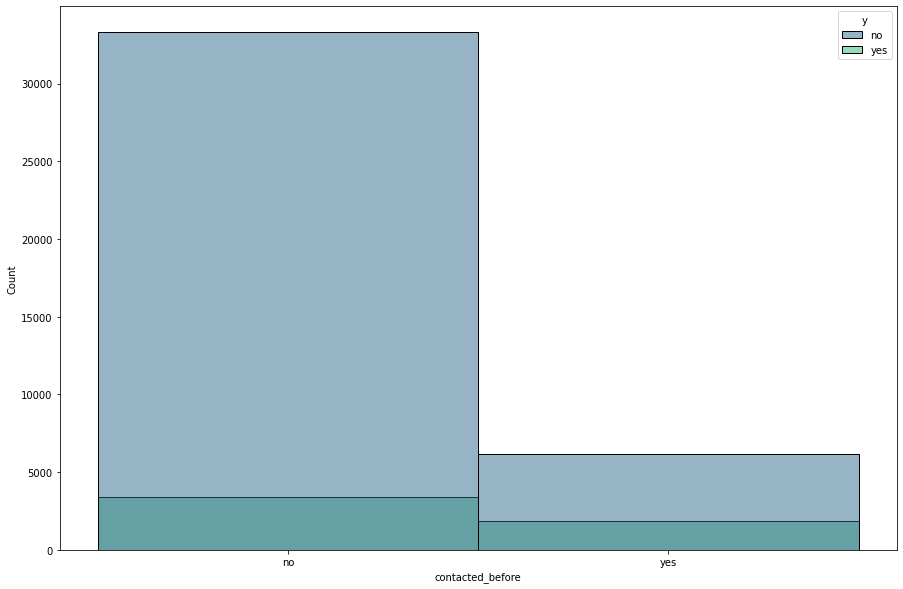
After looking at the plots above we removed the outliers

* In **duration** we removed those observation with no output and duration> 2000s
* In **campaign** we removed campaigns> 20
* In **previous** we removed observations for previous contacts> 11

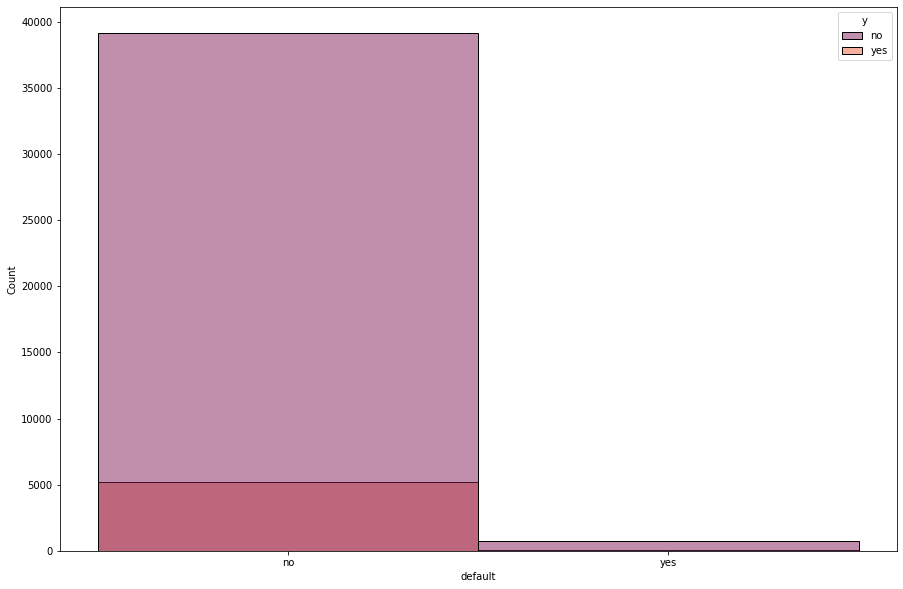


from the above plot we can conclude that when the client age categories are 'stable' ,'old age' and 'about to retire' then their is very high possibilty that those category person subscribe for a term deposit.

when clients category is struggling and counting last breathe then there is very less possibility that a customer subscribe for term deposit

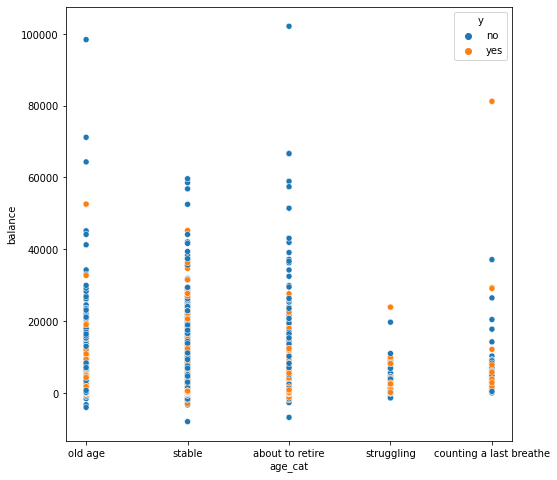


from the following plot bank has not contacted most of the clients ,the clients which bank not contacted before have high posibility that they suscribe for term deposite than a client which bank contacted before.

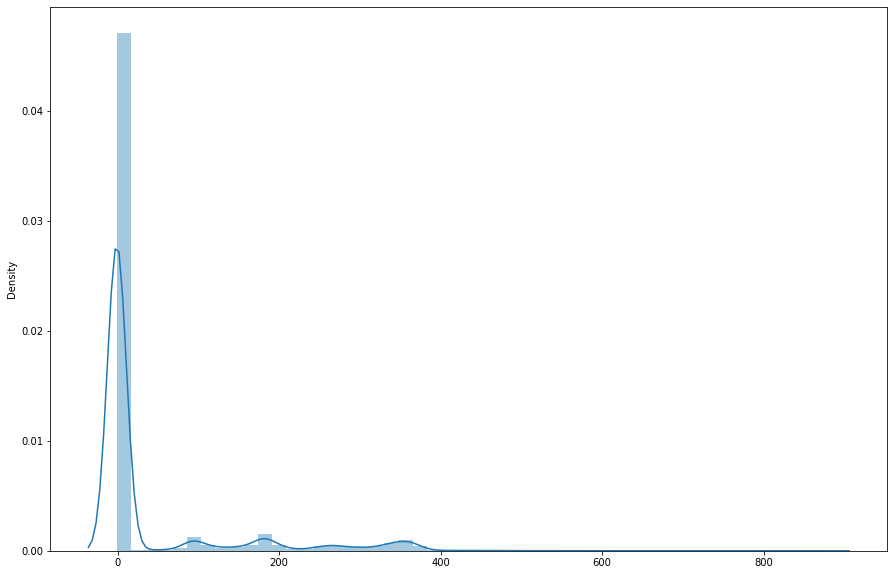


Most of the clients in our dataset was not credit defaulter so that when the client has credit is not in default then there is high possibility that customer suscribe for term deposite.

when the client is credit default there is very less possibility that a customer suscribe for term deposite.

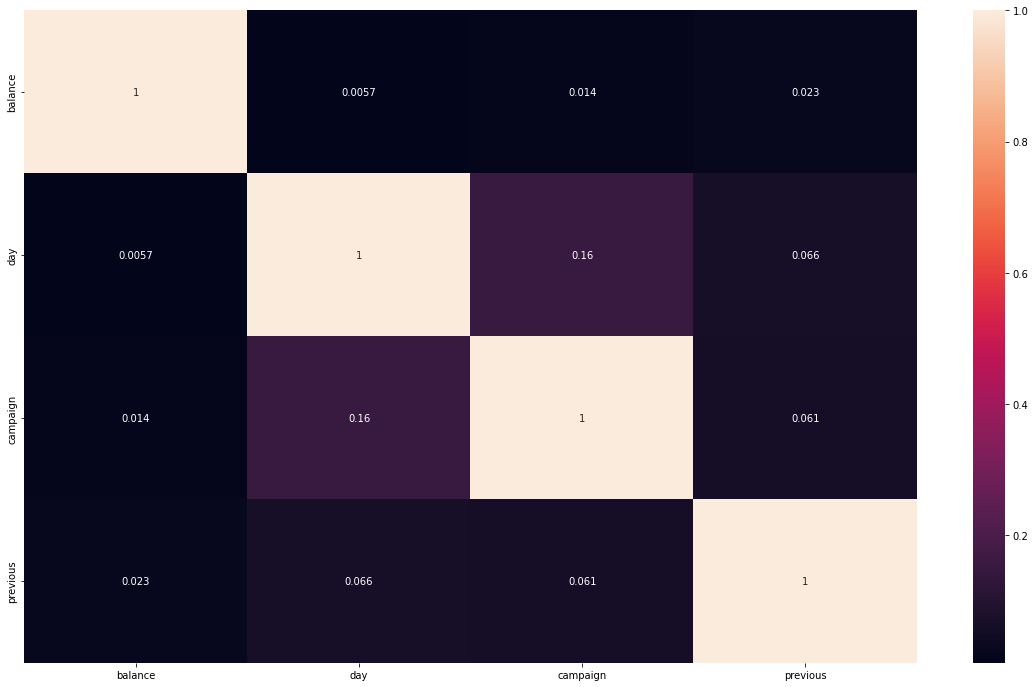


From the above plot we can roughly conclude that when that balance was from 500-35000 (in the middle range) then those customer subscribed for the term deposit so we can say that high balance or low balance will not be predict that client will subscribed for term deposit or not



From the above plot we can see that the pdays have most of the values are 0 and less than 0 so we have to drop that column for better prediction of our model

3.4) Correlation Map



* 1. **Model Building**

To start with building first we dealt with highly imbalanced data using SMOTE and then feature standardization.

The target variable contains highly imbalanced labeled data in the 88:12 ratio. Using SMOTE which is basically used to create synthetic class samples of minority class to balance the distribution of target variable. The target variable balanced for modeling.

Before Over Sampling, counts: Features (35767, 43) and Label (35767,)

0 31570

1 4197

Name: y, dtype: int64

0 7893

1 1049

Name: y, dtype: int64

After Over Sampling, counts: Features (63140, 43) and Label (63140,)

0 31570

1 31570

Name: y, dtype: int64

**Feature Standardization**

Standardization typically means rescales data to have mean of 0 and standard deviations of 1. To bring all values from independent variables in same scale. Using standard scalar, the independent variables transformed.

There are several classification models available for prediction/classification. In this project we used following models for classification Algorithm’s

1. **KNN**
2. **Random Forest**
3. **XGBOOST**
4. XGBOOST with Hyperparameter tunning

**3.6.1 K-Nearest neighbors (KNN)**

K-Nearest Neighbor is a non-parametric supervised learning algorithm both for classification and regression. The principle is to find the predefined number of training samples closest to the new point and predict the correct label from these training sample. It’s a simple and robust algorithm and effective in large training datasets.

Following are steps involved in KNN.

1. Select the K value.
2. Calculate the Euclidean distance between new point and training point.
3. According to the similarity in training data points, distance and K value the new data point gets assigned to the majority class.

Cross validation score [0.76655256 ,0.7773232 , 0.77723971, 0.78239813, 0.7752171 ]

KNN Test accuracy Score 0.7889885276288763

Precision Recall f1-score support

0 0.84 0.71 0.77 9981

1 0.75 0.87 0.80 9980

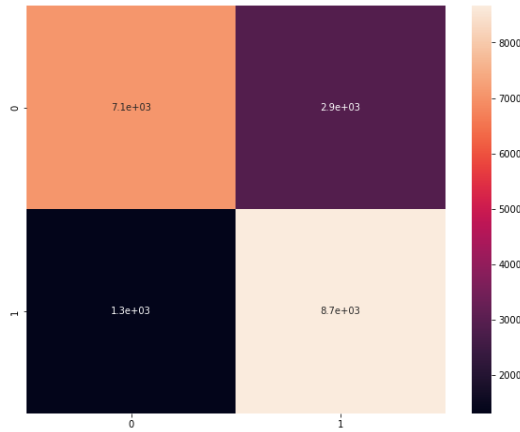
Accuracy 0.79 19961

Macro avg 0.80 0.79 0.79 19961

Weighted vag 0.80 0.79 0.79 19961

array([[7080, 2901],

[1311, 8669]])



**3.6.2 Random Forest**

Random forest is a Decision Tree based algorithm. It’s a supervised learning algorithm. This algorithm can solve both type of problems i.e. classification and regression. Decision Trees are flexible and it often gets overfitted. To overcome this challenge Random Forest helps to make classifications more efficiently. It creates a number of decision trees from a randomly selected subset of the training set and averages the-final outcome. Its accuracy is generally high. Random forest has ability to handle large number of input variables.

Cross Validation score [0.89997495 0.90030893 0.89922351 0.8995491 0.90405812]

Random Forest Test accuracy Score 0.9070186864385552

Precision recall f1-score support

0 0.87 0.96 0.91 9981

1 0.95 0.86 0.90 9980

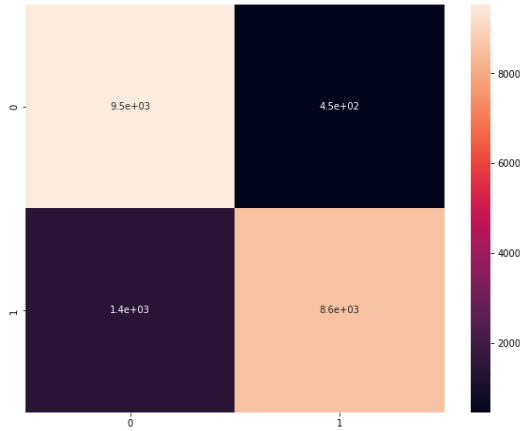
accuracy 0.91 19961

macro avg 0.91 0.91 0.91 19961

weighted avg 0.91 0.91 0.91 19961

array([[9533, 448],

[1408, 8572]])



1. XGBOOST

XGboost is the most widely used algorithm in machine learning, whether the problem is a classification or a regression problem. It is known for its good performance as compared to all other machine learning algorithms.

Cross validation score [0.93186942 0.93270435 0.93170243, 0.9261022 , 0.93470274]

Test accuracy Score 0.9356244677120384

precision recall f1-score support

0 0.90 0.98 0.94 9981

1 0.97 0.90 0.93 9980

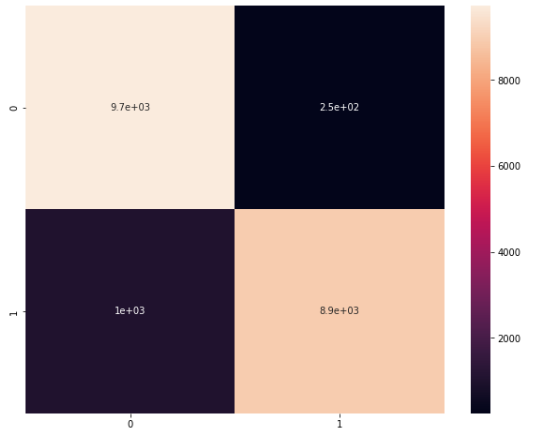
accuracy 0.94 19961

macro avg 0.94 0.94 0.94 19961

weighted avg 0.94 0.94 0.94 19961

array([[9734, 247],

[1038, 8942]])



**Hyperparameter tuning of XG Boost Classifier**

Cross validation score [0.93137419 0.93454667 0.93270997 0.93737475 0.93269873 0.93102872

0.92334669 0.93036072 0.93670675 0.93219773]

xgb\_hypertuned Test accuracy Score 0.9358749561645208

precision recall f1-score support

0 0.91 0.97 0.94 9981

1 0.97 0.90 0.93 9980

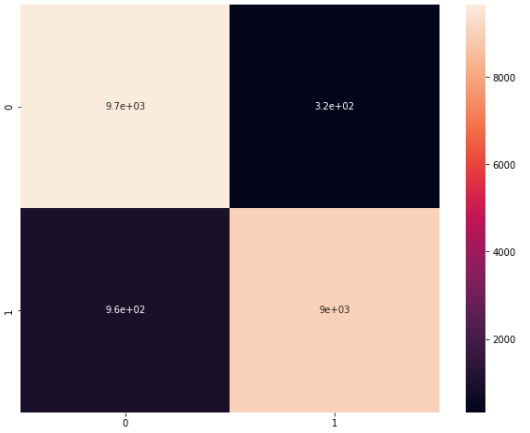
accuracy 0.94 19961

macro avg 0.94 0.94 0.94 19961

weighted avg 0.94 0.94 0.94 19961

array([[9663, 318],

[ 962, 9018]])



# 4.Model Evaluation

For classification problems we have different metrics to measure and analyze the model’s performance. In highly imbalanced target feature accuracy metrics doesn’t represents true reality of model.

**4.1 Confusion Matrix**

The confusion matrix is a tabular form metrics which tell us the truth labels classified versus to the model predicted labels. True Positive signifies the how many positive classes samples model able to predict correctly. True Negatives signifies how many negative class samples the model predicted correctly.

**4.2 Precision/Recall**

Precision is the ratio of correct positive predictions to the overall number of positive predictions: TP/TP+FP. It focus on Type 1 error.

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

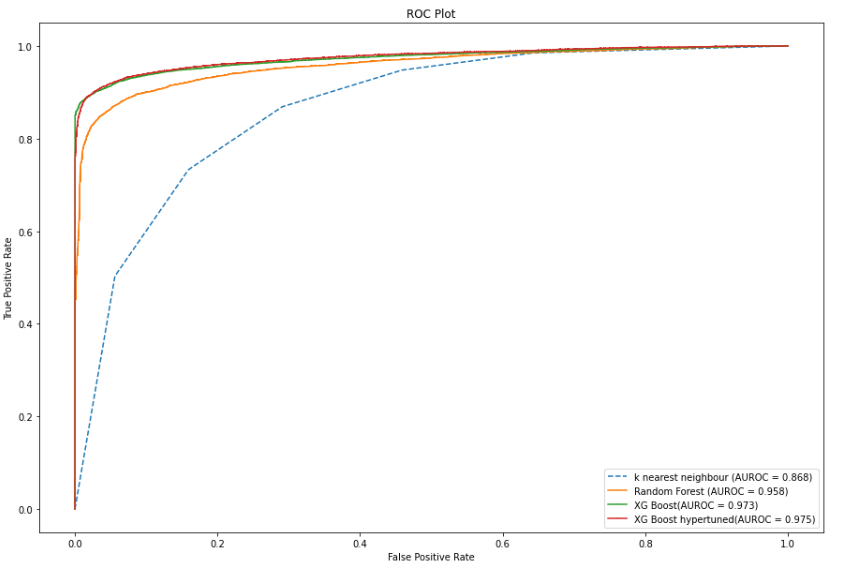
**4.3 Accuracy**

Accuracy is one of the simplest metrics to use. It’s defined as the number of correct predictions divided by the total number of predictions and multiplied by 100.

4.4 AUC ROC Score

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes

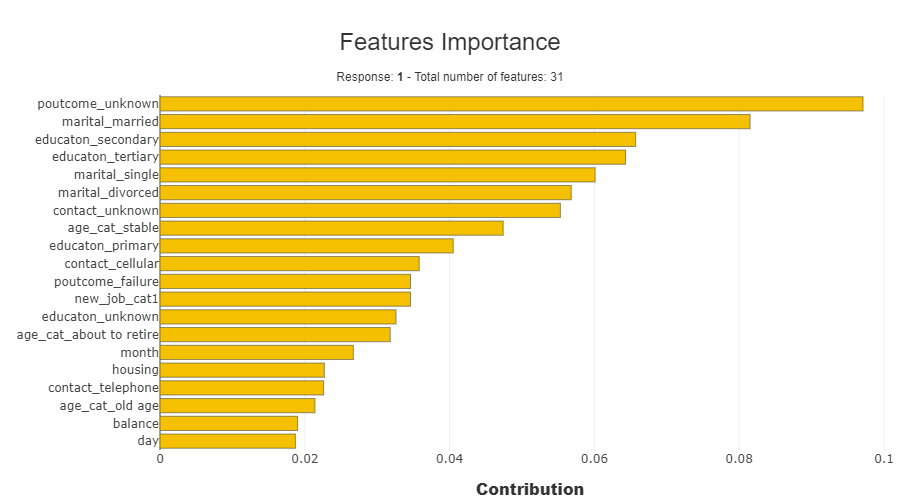
**AUC – ROC SCORE OF ALL MODELS**



**5) COMPARING MODELS**

****

6. Feature importance using shapash Explanatory



6. Conclusion and Future scope

**Conclusion-** It was a great learning experience working on a Bank dataset.

From the above model explanatory tool we have seen that poutcome

Unknown is the most important feature while predicting our target variable also from the table we can see that when the poutcome is 0 then it contribute in the negative way and increases the probability of predicting 0.

Marital married is the second most important feature for predicting target when the marital married then it will affect positively and increases the probability of predicting 1.

Also age cat stable variable affect positively on the target variable when the age of clients is stable then it will increases the probability of predicting 1 that means it higher the probability that client will subscribe for term deposit.

Also education secondary affects positively on the target variable when the client education is secondary then it increases the probability that client will agree to subscribe for term deposit.

From the above project we can conclude that XG boost classifier is the best fit classification model for predicting weather the client agree to subscribe for personal loan or not.

When we Hypertuned these XG Boost classifier the accuracy of the model increases by 1 % So it predicts 94% prediction correctly.

There are some important feature for predicting our target variable we use Shapash model explanatory to explore that features.

We visualize 20 feature which are most important while predicting target variable.

From that feature we conclude that clients age , education ,job and and marital status and outcome of previous campaign are the most important feature for predicting that weather client agree to subscribe for term deposit or not that’s why bank prefer these information to start for new campaign and to target customer.

**.** Future Scope -

Our main objective is to get good precision score for without 'duration' models and good recall score for 'duration' included model.

So, we can initially formulate the required time to converge a lead using 'duration' included models and then sort out precise leads for 'duration' excluded models using this formulated time.

Here, the idea is to find out responses for any particular record with varying assumed predefined duration range.

In this way we can help marketing team to get precise leads along with time required to converge that lead and also, those leads that have least probability to converge (if we get no positive response for any assumed duration). Thus, an effective marketing campaign can be executed with maximum leads converging to term deposit.