Capstone Project

(Supervised ML - Classification)

Bank Marketing Effective Prediction

By

Ajinkya Morade and Nitesh Gajakosh

CONTENTS

- Introduction.
- Problem Statement.
- Explanation of Features
- Exploratory Data Analysis
- Feature Engineering
- Model selection
- SMOTE
- Model Performances
- Hyperparameter Tuning of models
- ROC AUC Curve
- Comparing algorithms/Models
- Conclusions

INTRODUCTION

Bank marketing is **the design structure, layout and delivery of customer-needed services worked out by checking out the corporate objectives of the bank and environmental constraints**.
A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.
The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination or withdrawal-if they give several days notification. Also, there will be a penalty assessed for early termination.

Problem Statement

Make predictions

01

The classification Goal is to predict if the client will subscribe to a term deposit

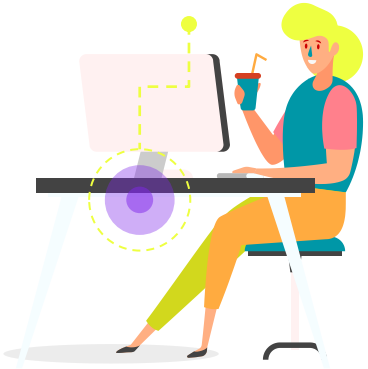
Model Development

02

Develop a **Supervised Machine Learning Model** using **Classification.**

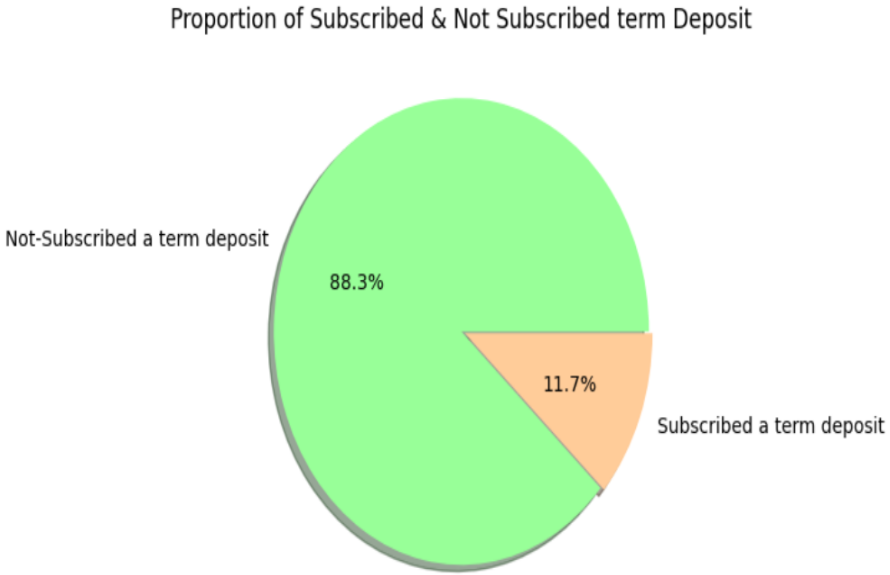
Features

01 Age
02 Job -type of a job
03 Marital -Marital Status
04 Education
05 Default- Has credit in default?
06 Housing - Has housing Loan?
07 Loan - Has a personal loan
08 Contact - Communication type
09 Month- Last contact month of the year
10 Day_of_week -Last day of week
11 Duration -Last contact duration
12 Balance



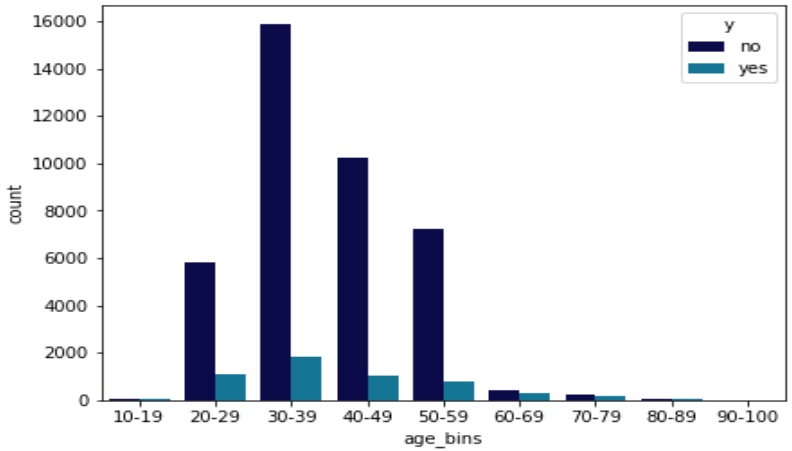
13 Campaign-No. Of contacts performed during this campaign
14 Pdays- No. of days
15 Previous-No. Of contacts performed before this campaign
16 P-outcome -Outcome of previous marketing campaign
17 Target variable(y) -Has client subscribed to term deposit

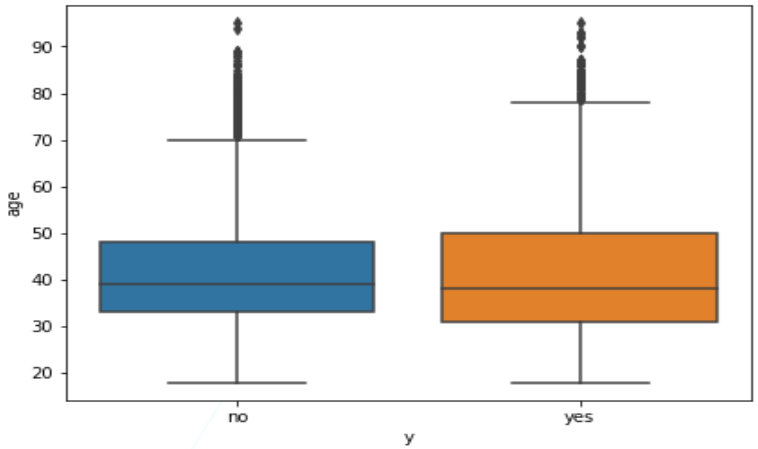
EDA



We can see from the above plot that the dataset is imbalanced, where the number of the non-subscribed class is close to 8 times the number of Subscribed class

EDA



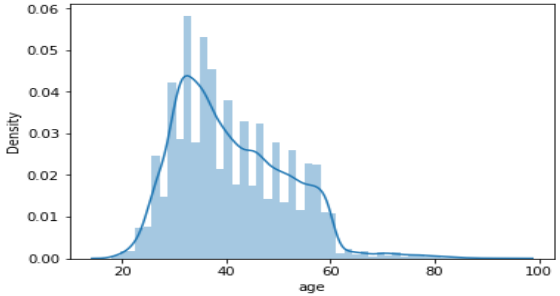


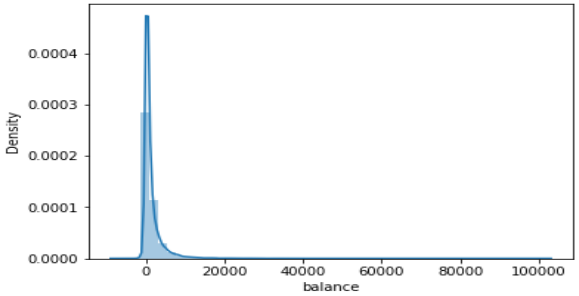
Majority of the customers are of the age group 30-39. Followed by 40-49 and 50-59

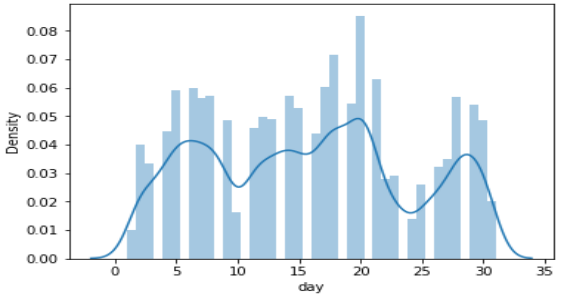
The Box Plot for the both subscribed and Not-subscribed customers looks the same
In No class,outliers are present above age 70 and For Yes class, Outliers are present above age 75

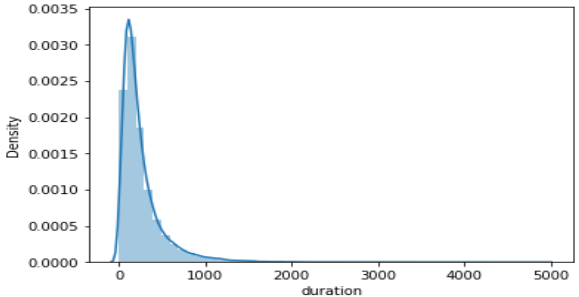
EDA Of Categorical Variables

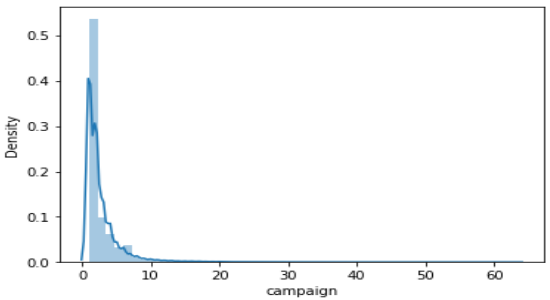
Distribution of Continuous Numerical Features

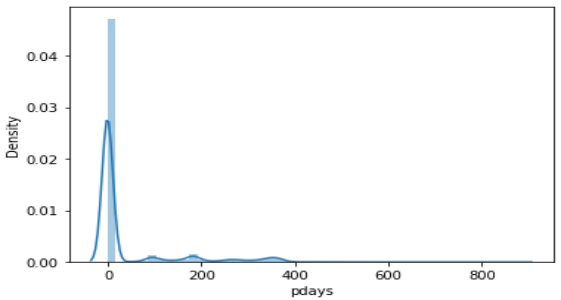


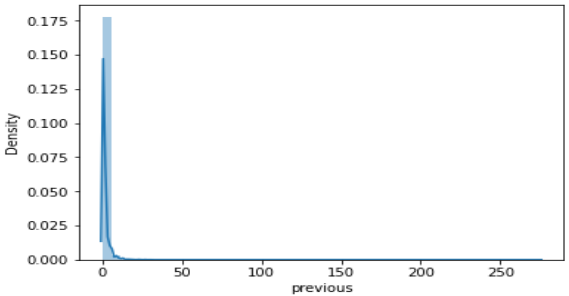






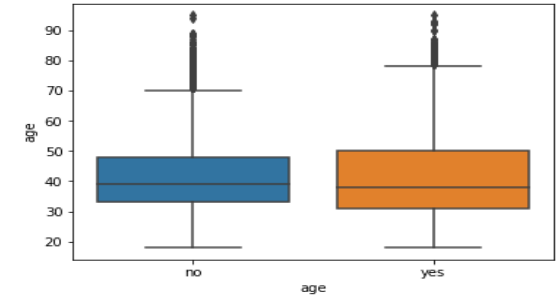


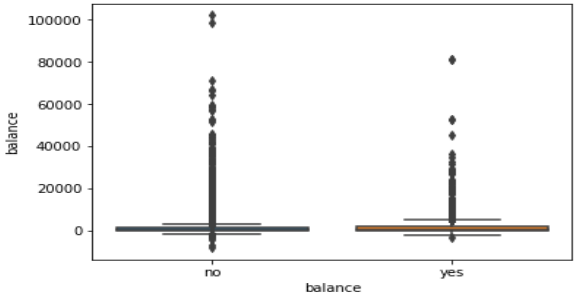


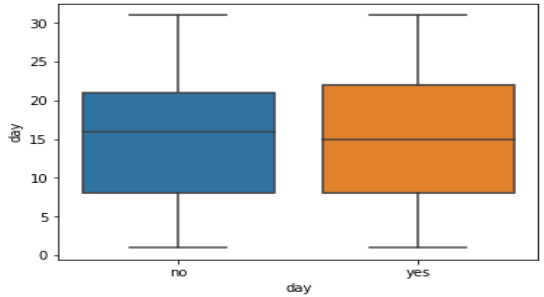


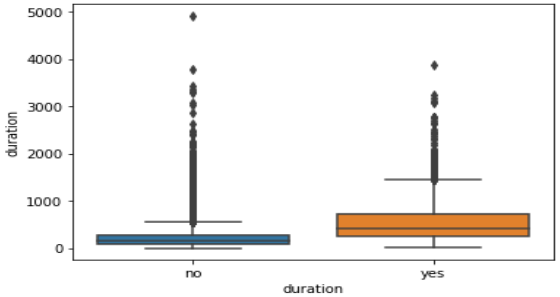
- It seems age, days distributed normally.
- Balance, duration, campaign, pdays and previous heavily skewed towards left and seems to be have some outliers.

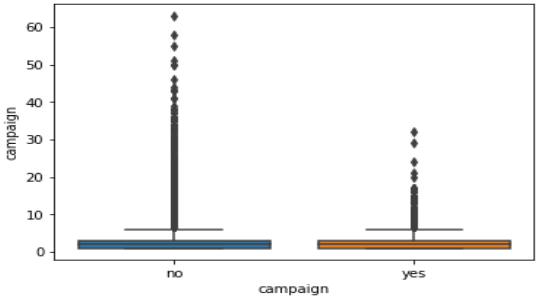
Relation between Continuous numerical Features and Labels

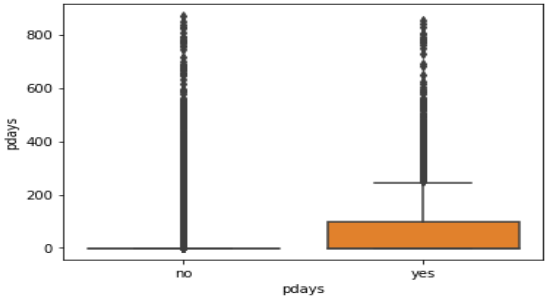


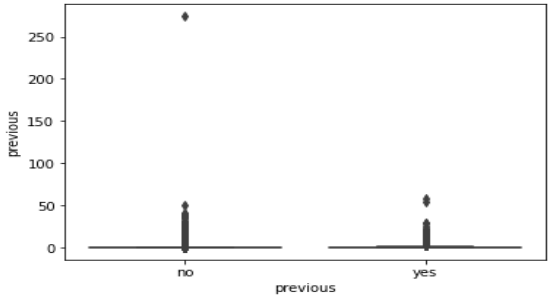






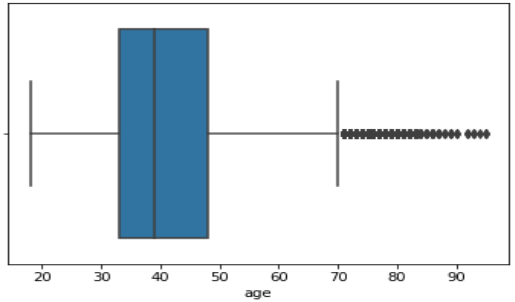


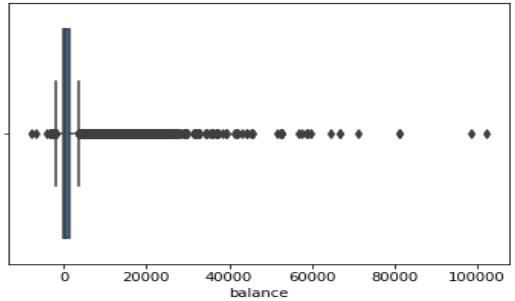


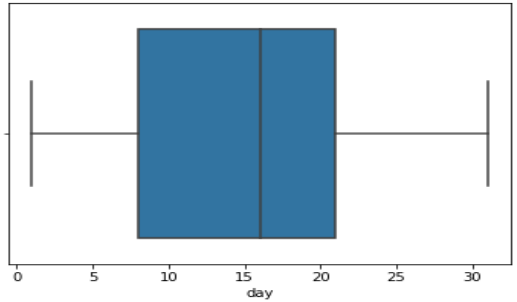


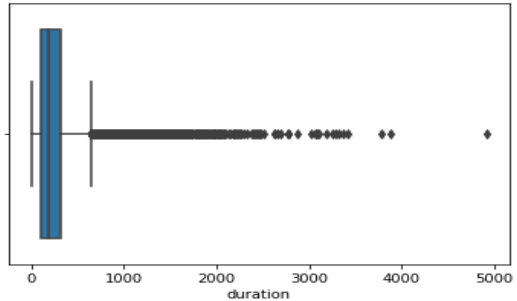
- Client shows interest on deposit who had discussion for longer duration.

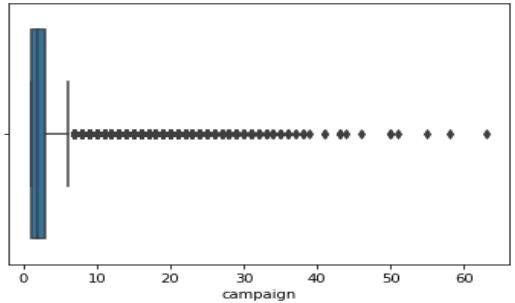
Find Outliers in numerical features

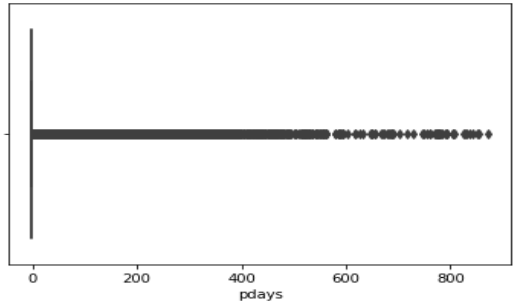


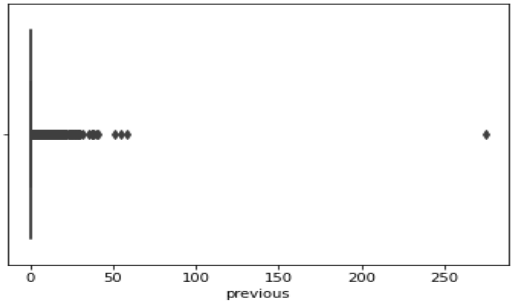






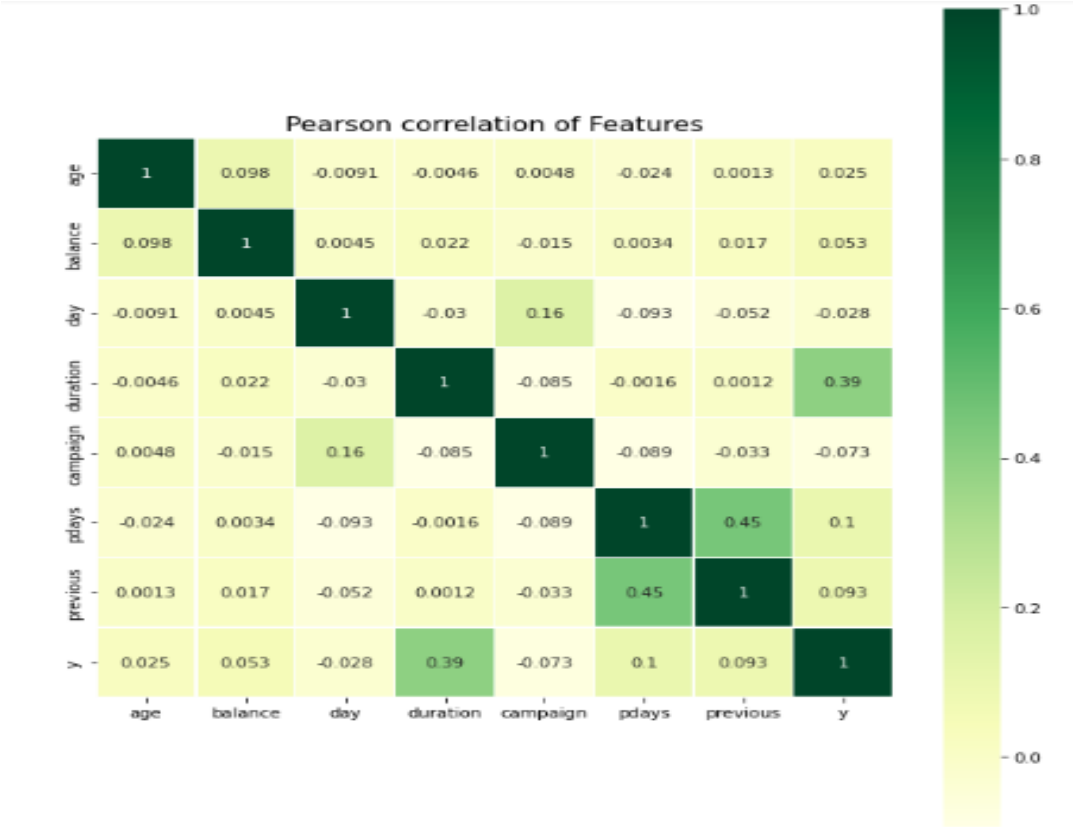






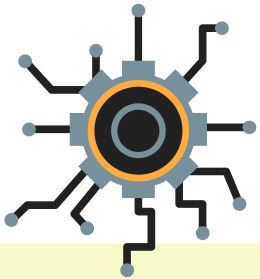
- We can conclude that age, balance, duration, campaign, pdays and previous has some outliers.

Correlation between numerical features



- pdays and previous have high correlation of 0.45 and the dependent variable has highest correlation with duration .
- It seems no feature is heavily correlated with other features.

Feature Engineering



Feature Engineering

01

Dropping columns

Default, pdays etc.

02

Label Encoding

Housing, Loan and y.

03

Getting Dummies

Job, Education, Marital Status, Contact, Month and Previous outcome

Model Selection

01

Logistic Regression

02

Decision Tree Classifier

03

Random Forest Classifier

04

Gradient Boosting Classifier

05

XGBoost Classifier

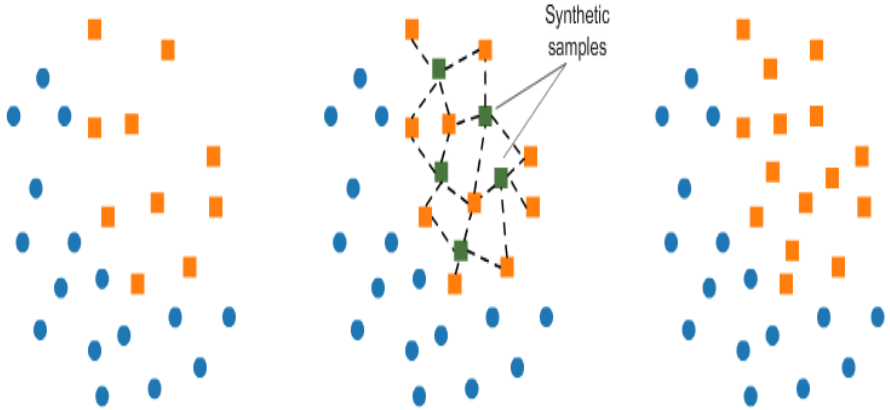
06

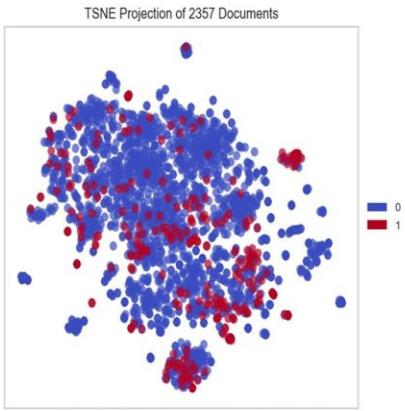
K Neighbors Classifier

07

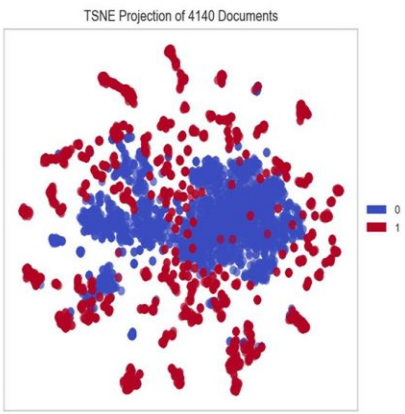
Naive Bayes Classifier

SMOTE





(a) Before SMOTE



(b) After SMOTE

SMOTE: a powerful solution for imbalanced data.
SMOTE is an algorithm that performs data augmentation by creating **synthetic data points** based on the original data points. SMOTE can be seen as an advanced version of oversampling, or as a specific algorithm for data augmentation. The advantage of SMOTE is that you are **not generating duplicates**, but rather creating synthetic data points that are **slightly different** from the original data points.

The SMOTE algorithm works as follows:

Steps Involved	
1.	You draw a random sample from minority class.
2.	For the observations in this sample, you will identify the k nearest neighbors.
3.	Take one of those neighbors and identify the vector between the current data point and the selected neighbor.
4.	You multiply the vector by a random number between 0 and 1.
5.	To obtain the synthetic data point, you add this to the current data point.

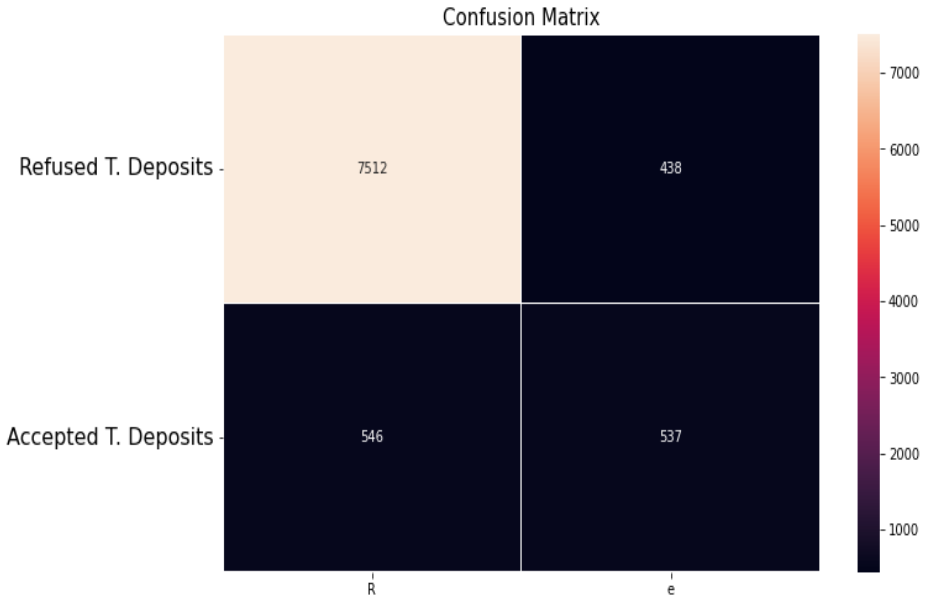
Random Forest Classifier

Classification report

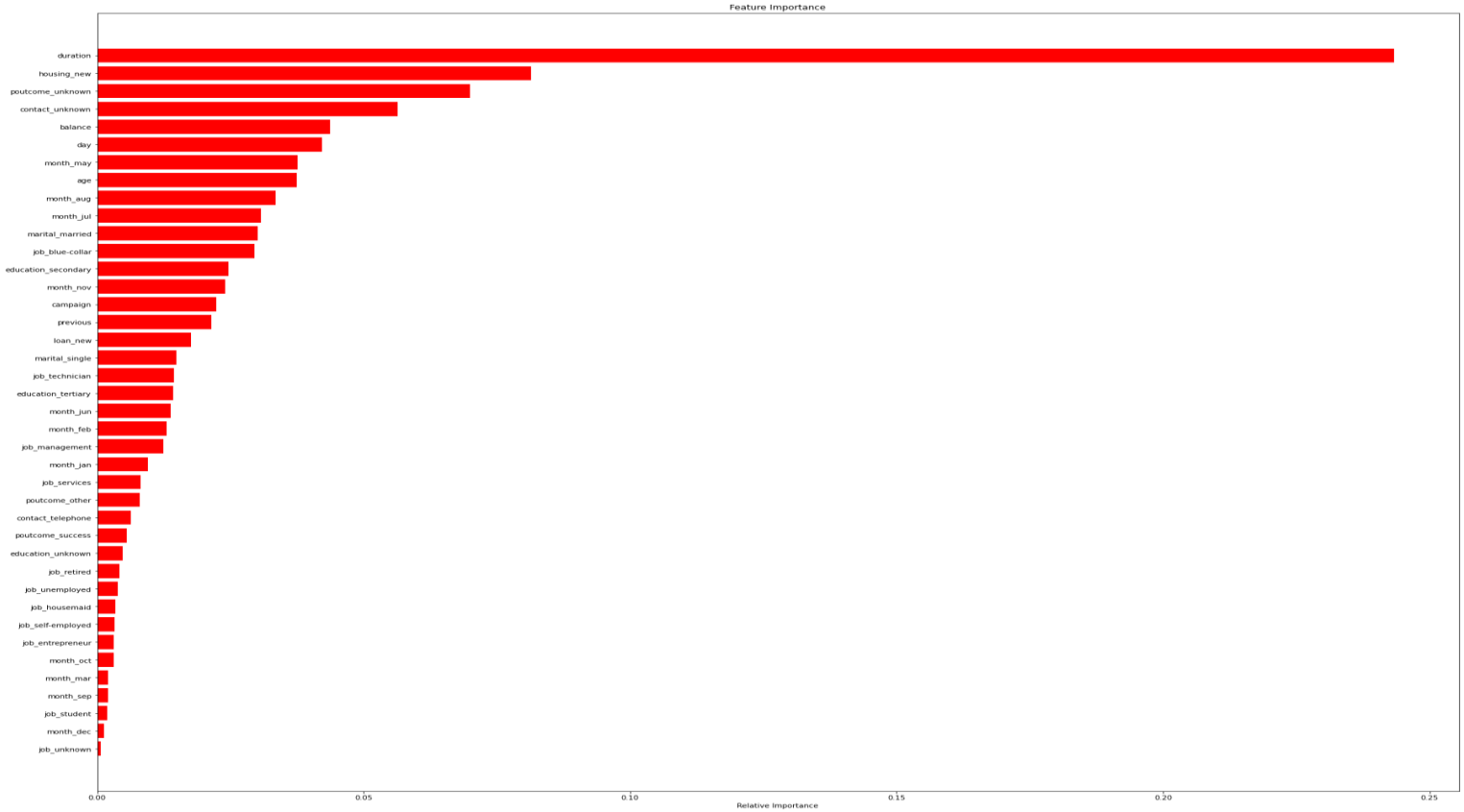
Training accuracy Score	1.0
Testing accuracy Score	0.880

	Precision	recall	F1-score	Support
0	0.94	0.93	0.94	8058
1	0.50	0.55	0.52	975
Accuracy			0.880	9033
Macro avg	0.72	0.74	0.73	9033
Weighted avg	0.90	0.89	0.89	9033

Confusion matrix



Feature Importance



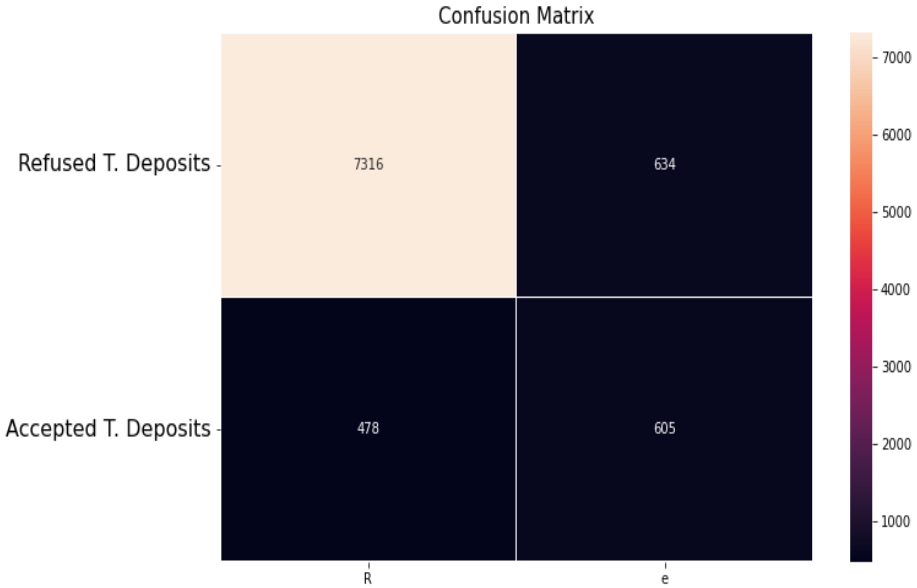
Hyperparameter Tuning

Classification report

Training accuracy Score	1.0
Testing accuracy Score	0.8910660909996679

	Precision	recall	F1-score	Support
0	0.95	0.93	0.94	8058
1	0.50	0.55	0.52	975
Accuracy			0.89	9033
Macro avg	0.72	0.74	0.73	9033
Weighted avg	0.90	0.89	0.89	9033

Confusion matrix



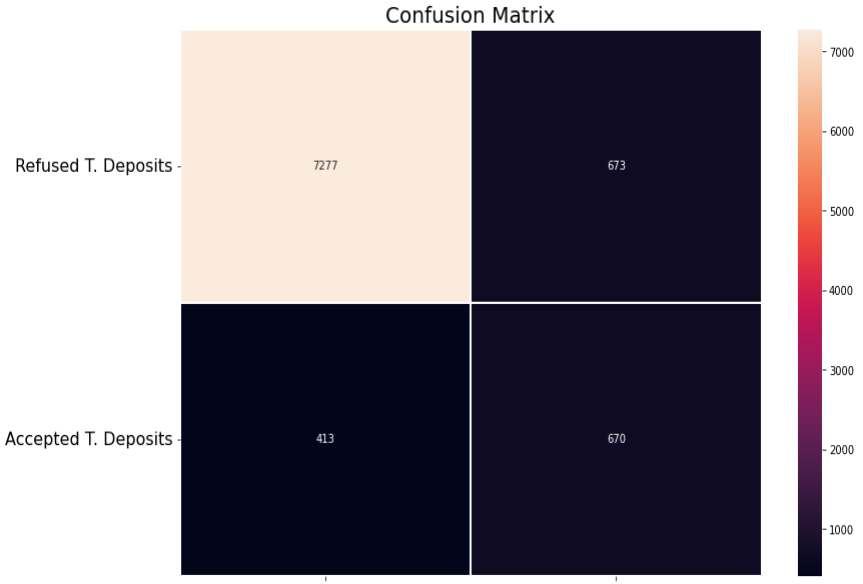
XG Boost Classifier

Classification report

Training accuracy Score	0.93
Testing accuracy Score	0.87

	Precision	recall	F1-score	Support
0	0.92	0.95	0.93	7690
1	0.62	0.50	0.55	1343
Accuracy			0.87	9033
Macro avg	0.77	0.72	0.74	9033
Weighted avg	0.87	0.88	0.87	9033

Confusion matrix



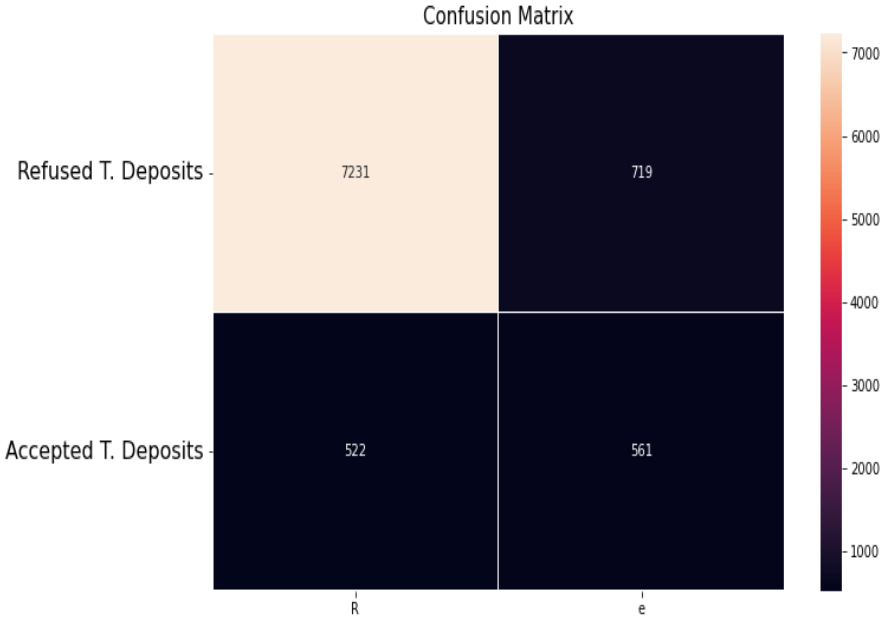
Gradient Boosting Classifier

Classification report

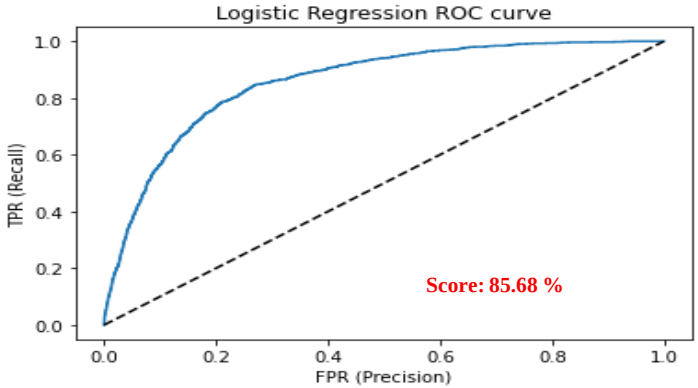
Training accuracy Score	0.93
Testing accuracy Score	0.88

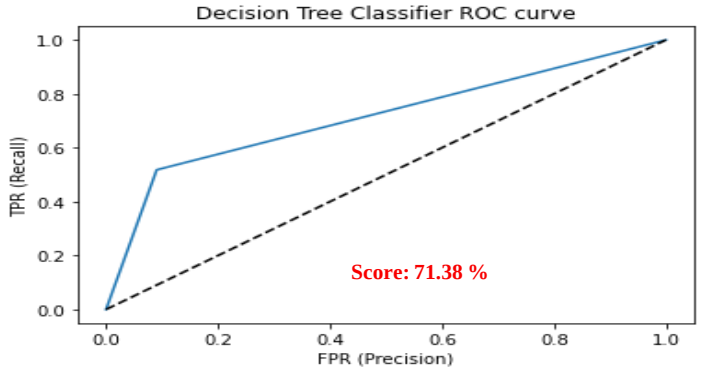
	Precision	recall	F1-score	Support
0	0.91	0.93	0.93	7753
1	0.52	0.44	0.48	1280
Accuracy			0.88	9033
Macro avg	0.71	0.69	0.70	9033
Weighted avg	0.85	0.86	0.86	9033

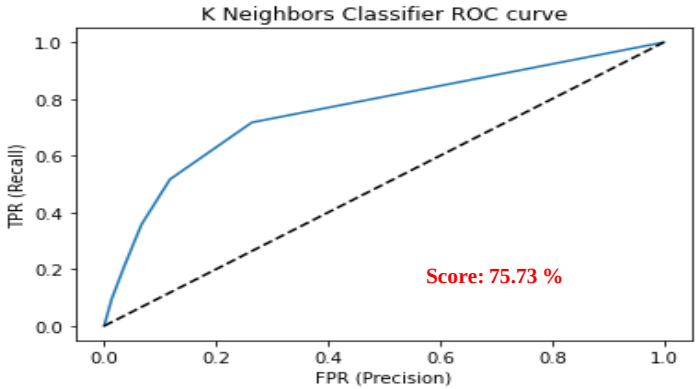
Confusion matrix

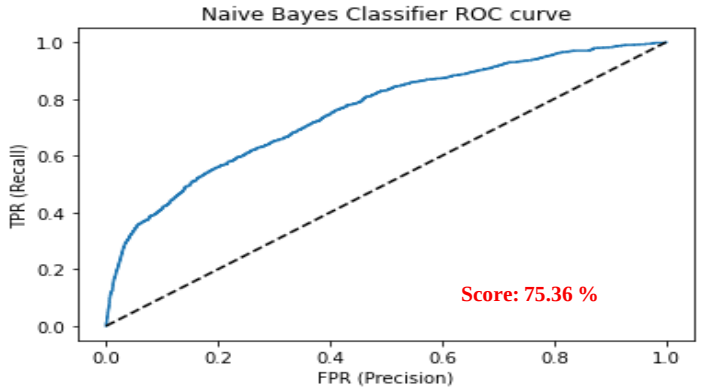


ROC AUC Curve

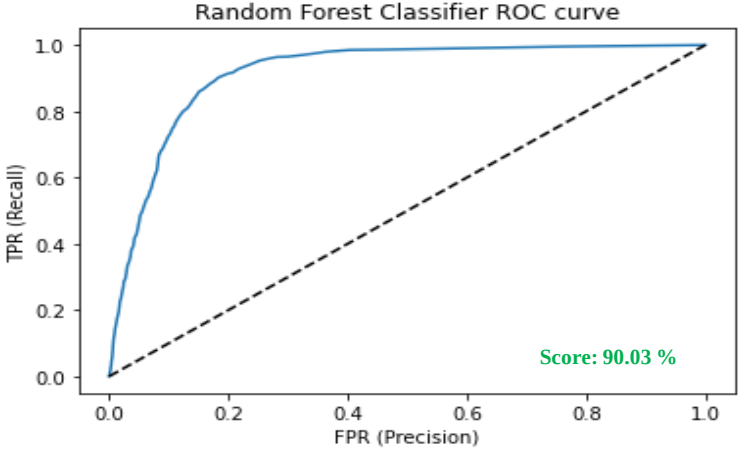


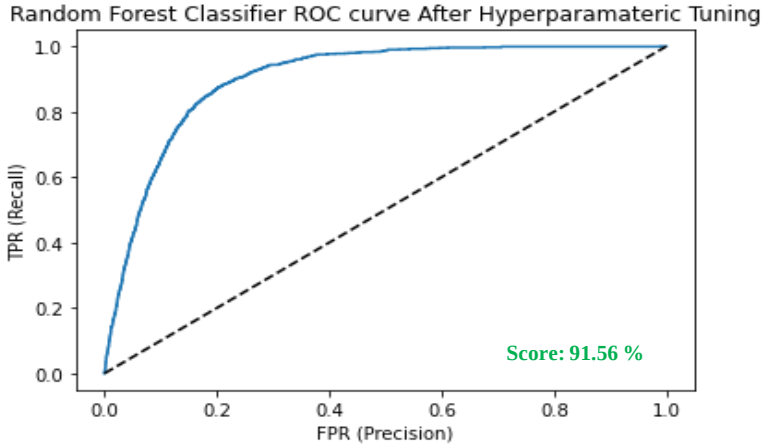


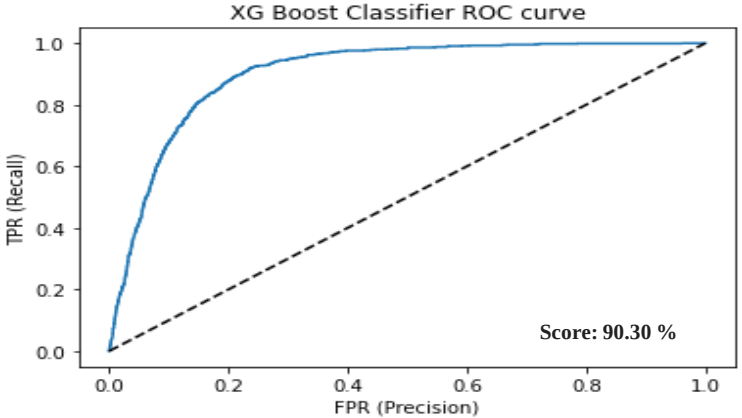


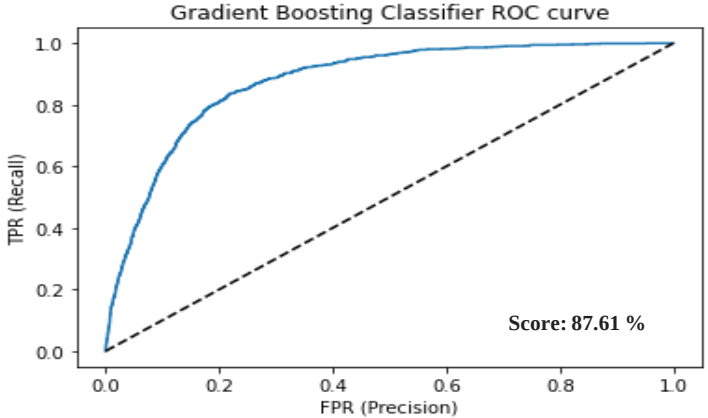


ROC AUC Curve (Top 3 accuracy models)

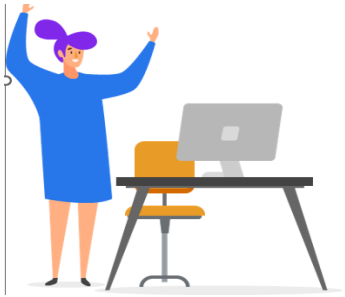


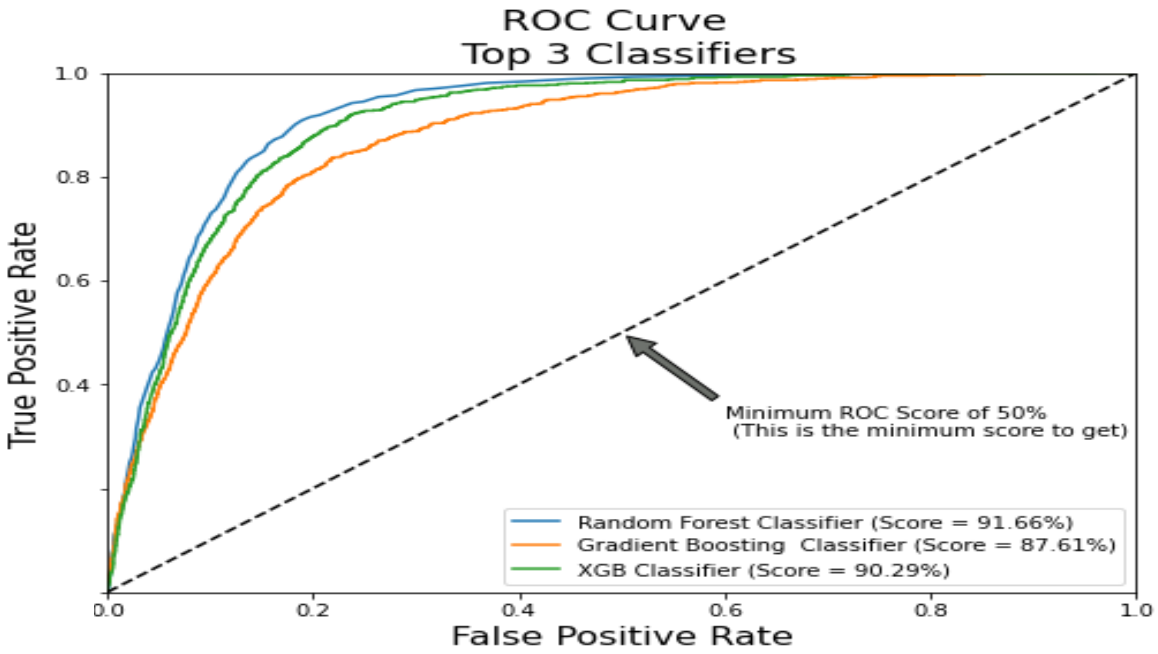






ROC AUC Curve for top 3 Models and its Performance



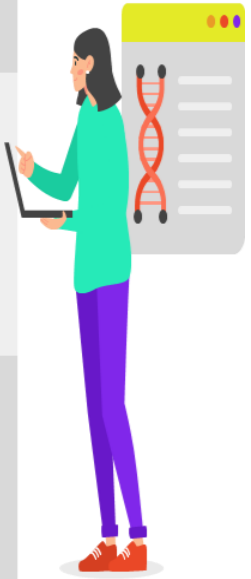


Comparison Of Models

Sr.no.	Model Name	Train Accuracy	Test Accuracy	F1 Score class 0	F1 Score class 1	ROC AUC Score
1	Logistic Regression	0.92	0.87	0.93	0.46	0.8568
2	Random Forest Classifier	1	0.89	0.94	0.52	0.9166
3	Gradient Boosting Classifer	0.93	0.88	0.93	0.48	0.8761
4	XG Boosting Classifier	0.93	0.88	0.93	0.55	0.9029
5	Decision Tree Classifier	1	0.86	0.92	0.47	0.7096
6	Naïve Byer Classifier	0.84	0.81	0.89	0.38	0.7536
7	K Neighbors Classifier	0.94	0.86	0.92	0.39	0.7573

Conclusions

Months of Marketing Activity	it will be wise for the bank to focus the marketing campaign during the months of **March, September, October and December.**
Campaign Calls	A policy should be implemented that states that no more than 3 calls should be applied to the same potential client in order to save time and effort in getting new potential clients. Remember, the more we call the same potential client, the more likely he or she will decline to open a term deposit.
Age Category	The next marketing campaign of the bank should target potential clients in their 20s or younger and 60s or older. The youngest category had a 60% chance of subscribing to a term deposit while the eldest category had a 76% chance of subscribing to a term deposit. It will be great if for the next campaign the bank addressed these two categories and therefore, increased the likelihood of more term deposits subscriptions.



Thank You