

# **Capstone Project**

## **Bank Marketing Effectiveness Prediction**

**Done by**  
**Amit Kundu**  
**&**  
**Sananda Biswas Chatterjee**

# CONTENTS

- Introduction
- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Model Implementation
- Evaluation Metrics
- Conclusion

# INTRODUCTION

A bank's corporate objectives, as well as the external environment, are taken into consideration while planning, organising, and providing the services that consumers required.

Depositing money into a financial institutions account is a part of the term deposit. Term deposits have a variety of minimum deposit requirements and frequently have short terms, ranging from a month to a few years.

Investors who purchase term deposits should be aware that they can only withdraw their funds after the term has passed. In rare circumstances, the account holder may permit to an early cancellation or withdrawal if the investor provides several days' notice. Additionally, early termination will impose a price.

# 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  $y$ ).

# DATA SUMMARY

**The Dataset contains 17 Features with 45211 observation.**

## **Categorical Features**

- Marital - (Married , Single , Divorced)
- Job - (Management,BlueCollar,retired etc.)
- Contact - (Telephone,Cellular,Unknown)
- Education - (Primary,Secondary,Tertiary)
- Month - (Jan,Feb,Mar,Apr,May etc.)
- Poutcome - (Success,Failure,Other,Unknown)
- Housing - (Yes/No)
- Loan - (Yes/No)
- Default - (Yes/No)

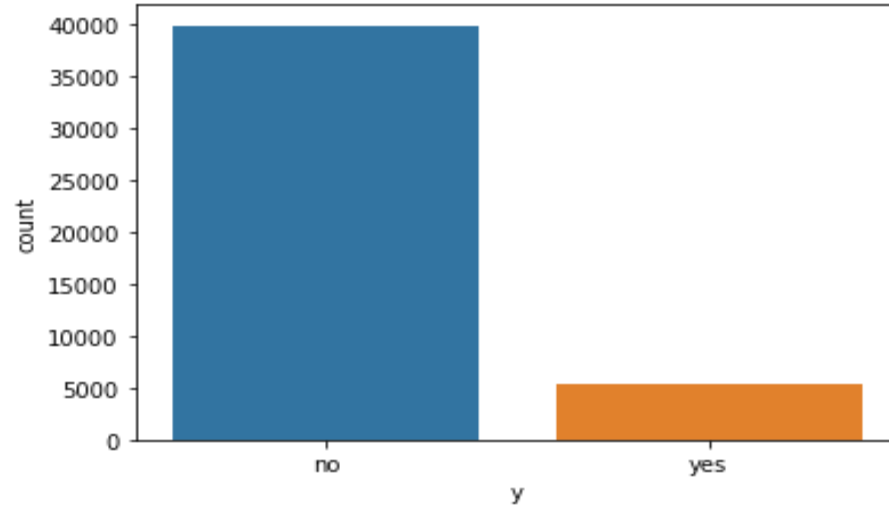
## **Numerical Features**

- Age
- Balance
- Day
- Duration
- Campaign
- Pdays
- Previous

## **Desired target**

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

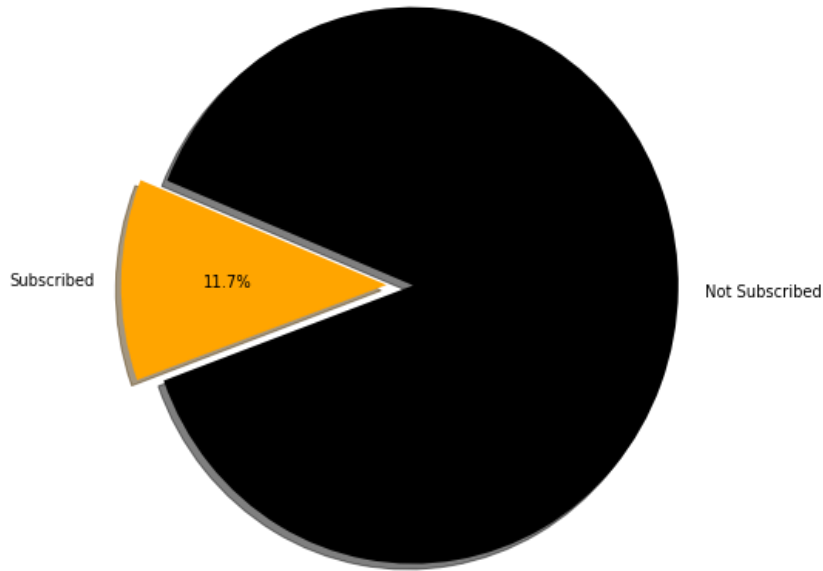
## Exploratory Data Analysis (Target)



- As We Can See that our data is highly imbalanced, because majority of the data points belong to 'no' class.

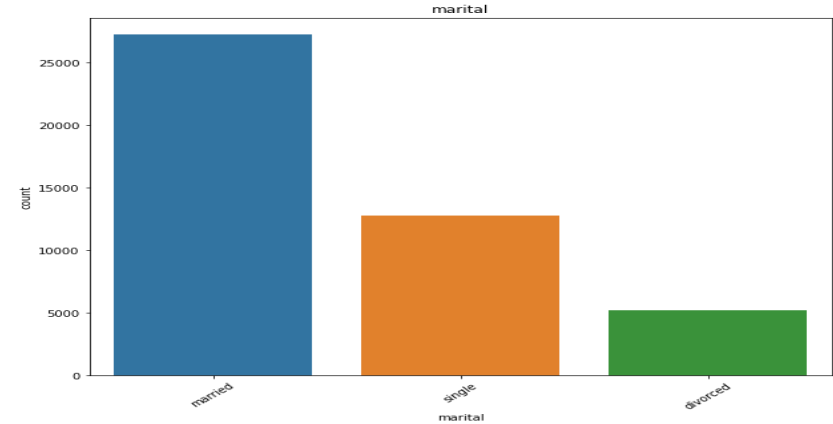
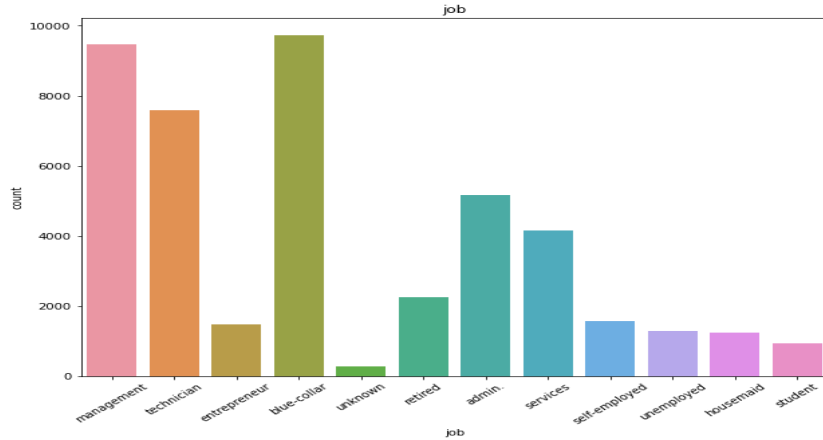
## EDA(Continued...)

Proportion of Subscribed & Not Subscribed term Deposit



- This information shows that 88% of customers did not subscribed term deposit.
- So, we can say that the percentage of people subscribing to the term deposit is quite low.

# EDA Of Categorical Variables

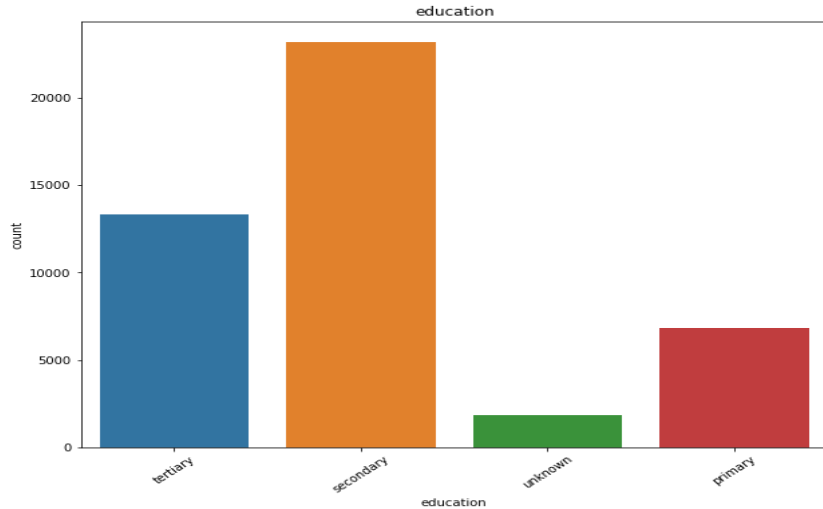


- There are 12 different job profiles in this plot. The top 3 job profiles, contributing for 60% of the total records, are “management,” “blue-collar,” and “technician.”
- Management-related occupations have subscribed more for the deposits.

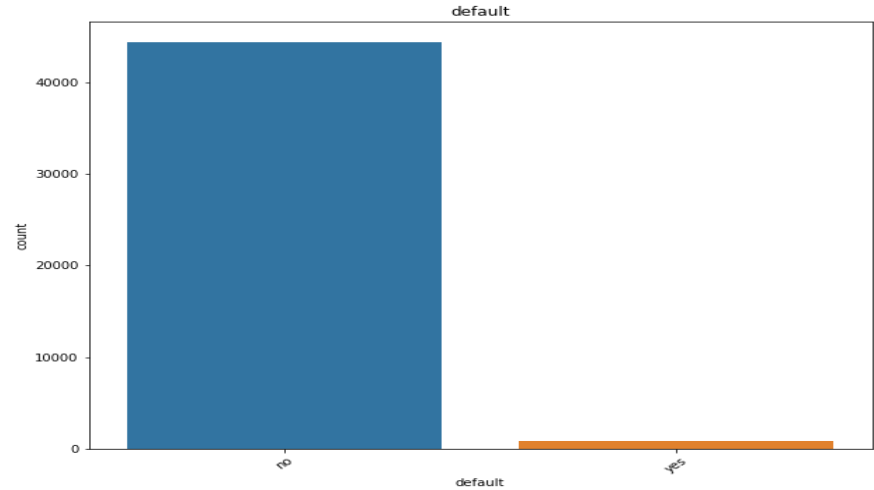
- In this chart married clients has high number of records compared to others.
- Compared to others marital status, married people have subscribed more for deposits.



## EDA(Continued)

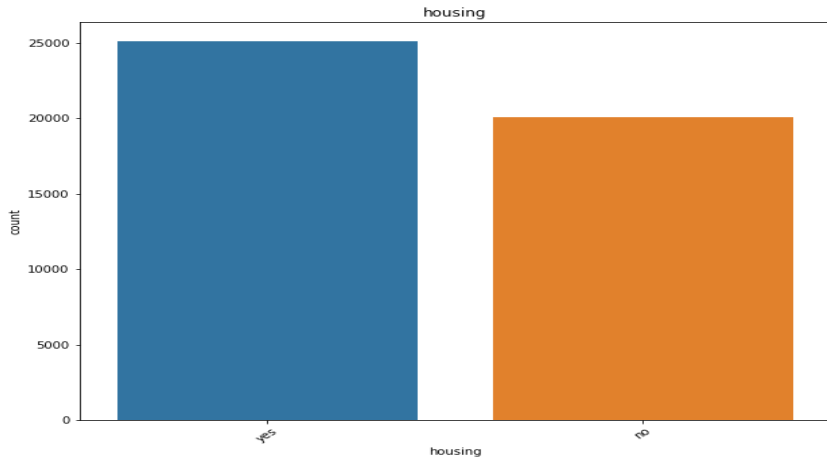


- We can see that clients with secondary education backgrounds are more in numbers as well as people with secondary education qualification has subscribed most for deposit.

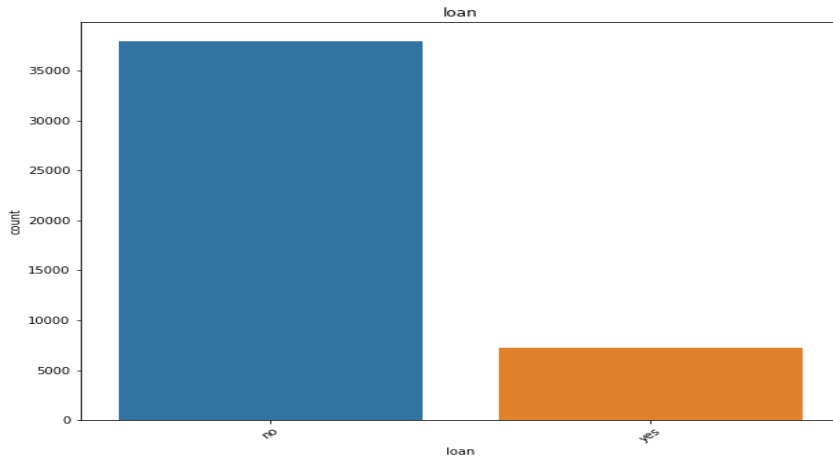


- Only a small percentage of defaulting clients are contacted because the majority of those with default status of "no" haven't subscribed for bank deposits.

## EDA(Continued)

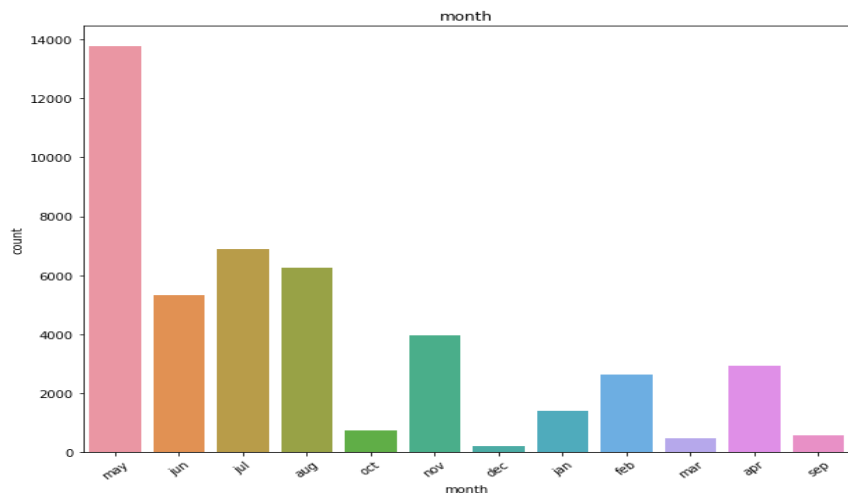


- Client who has housing loan seems to be not interested much on deposit.
- The majority of the clients has a housing loan.

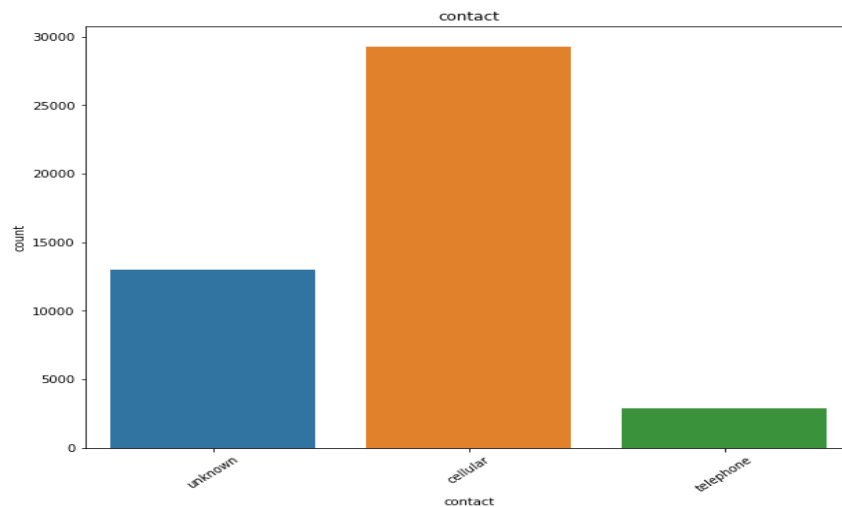


- People who do not have personal loans are more likely to choose term deposits.

# EDA(Continued)

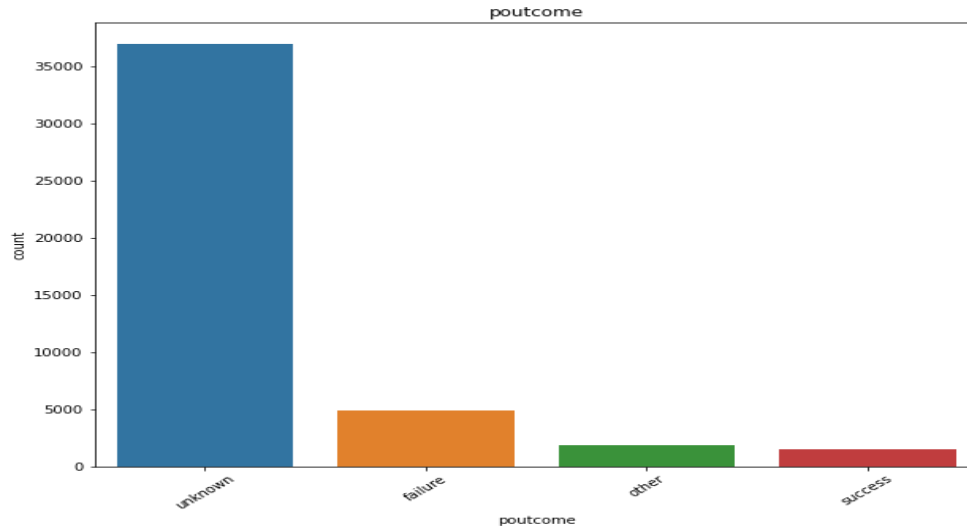


- The month of "May" showing the highest amount of marketing activity.



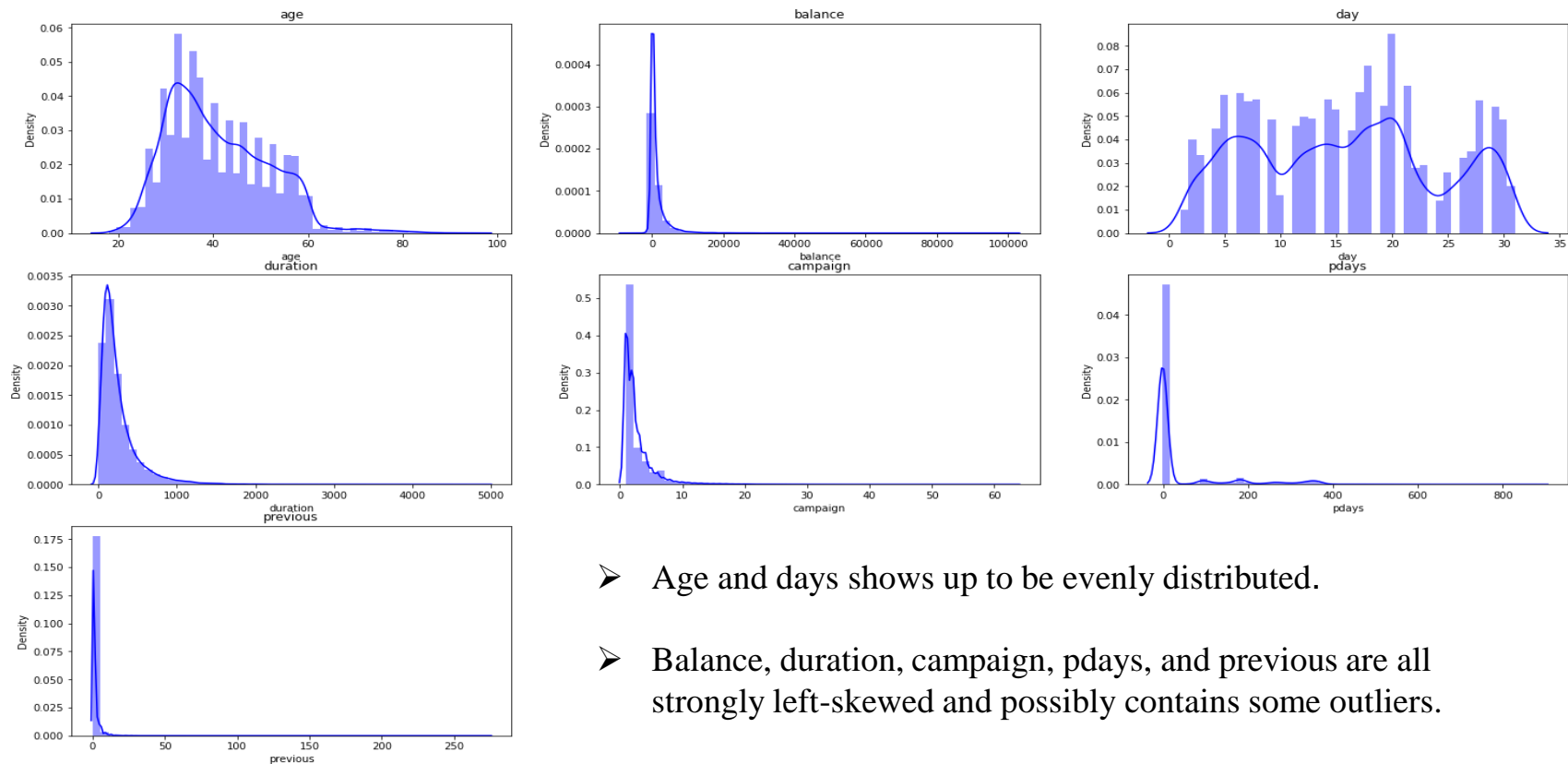
- Most people are connect through the cellular than telephone.
- People who were contacted by the bank on their mobile devices subscribed to the deposits more.

## EDA(Continued)



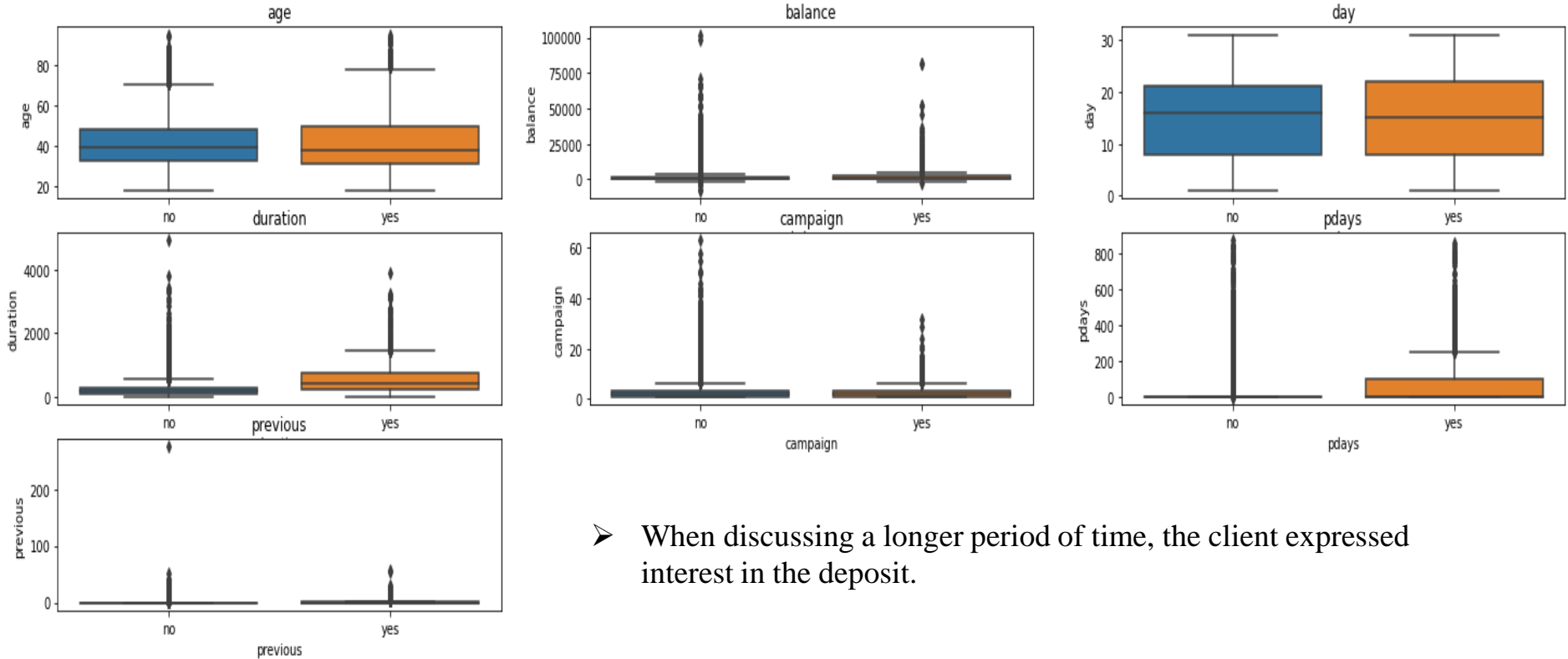
- Majority of the poutcome of the previous campaign is Non-Existent.
- People whose previous poutcome is non-existent have actually subscribed more than any other group.

# Distribution of Numerical Features

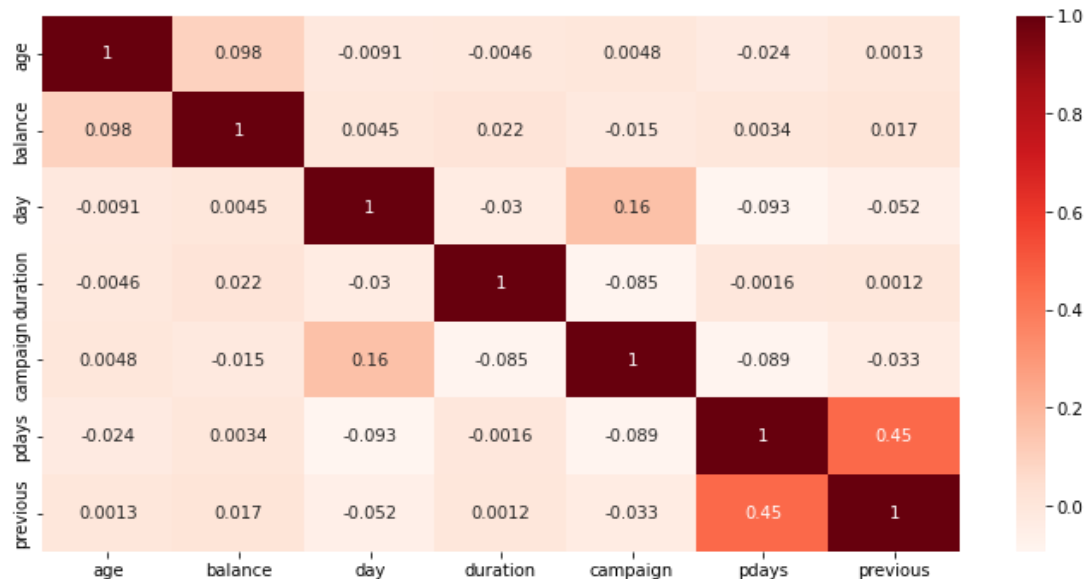


- Age and days shows up to be evenly distributed.
- Balance, duration, campaign, pdays, and previous are all strongly left-skewed and possibly contains some outliers.

# Relation between numerical Features and Labels



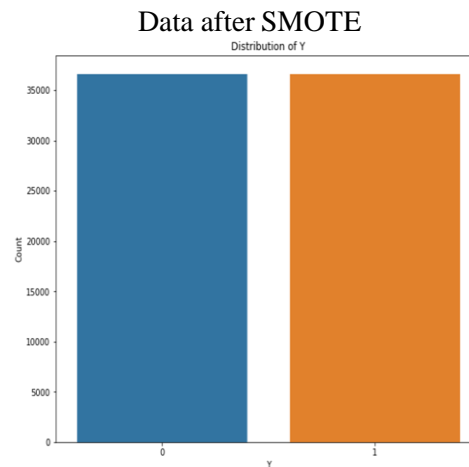
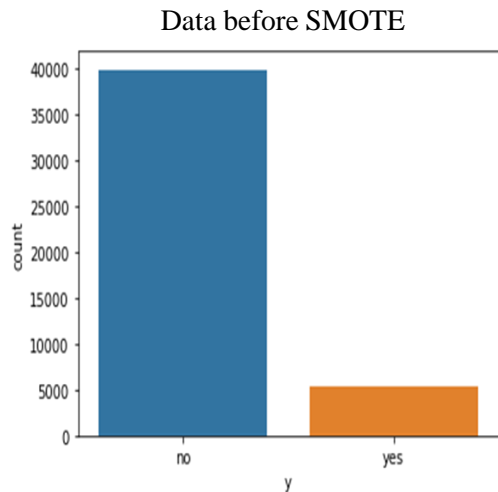
# Correlation between numerical features



- pdays and previous have the highest correlation with each other (0.45).
- It seems that there is no feature that strongly correlates with other features.

# SMOTE

- SMOTE stands for Synthetic Minority Over-sampling Technique.
- It is an over-sampling technique in which new synthetic observations are created using the existing samples of the minority class.
- 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.





# **MODEL BUILDING**

**We are going to deploy the model of:**

**1) Logistic Regression**

**2) Random Forest Classifier**

**3) Decision Tree Classifier**

**4) K-Nearest Neighbors**

**5) XGBoost Classifier**

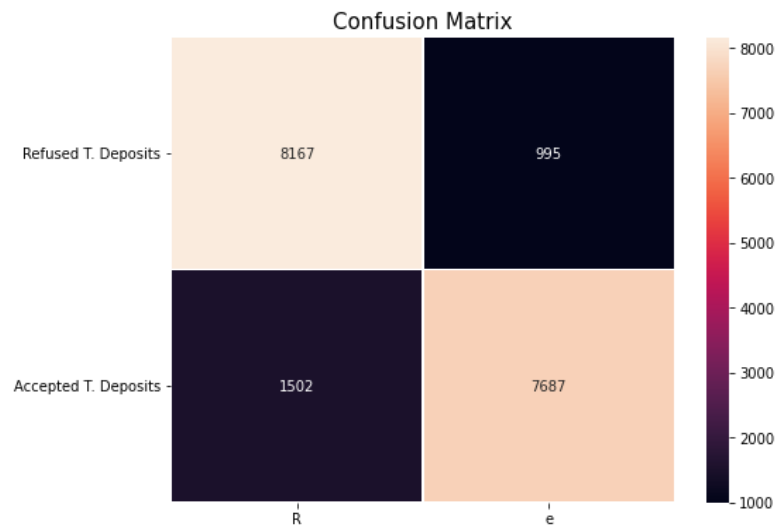
# Logistic Regression

Training accuracy	0.8675
Testing accuracy	0.8639

## Classification report

	precision	recall	F1-score	support
0	0.88	0.84	0.87	9669
1	0.84	0.89	0.86	8682
Accuracy			0.86	18351
Macro avg	0.86	0.87	0.86	18351
weighted avg	0.87	0.86	0.86	18351

## Confusion matrix



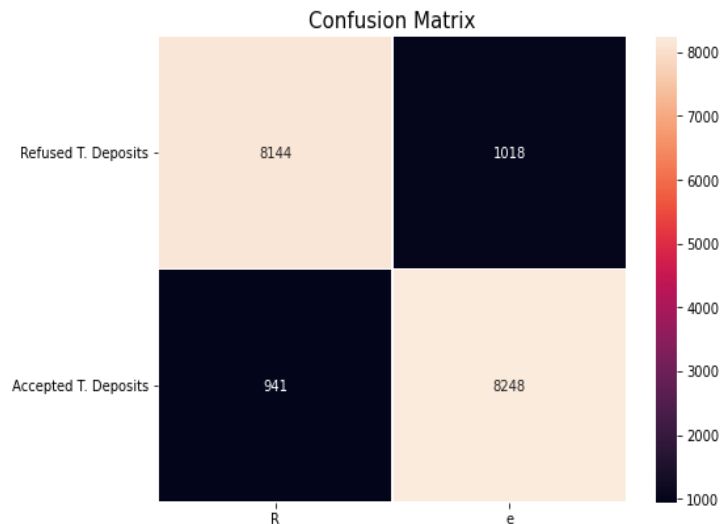
# Random Forest Classifier

Training accuracy	0.9916
Testing accuracy	0.8932

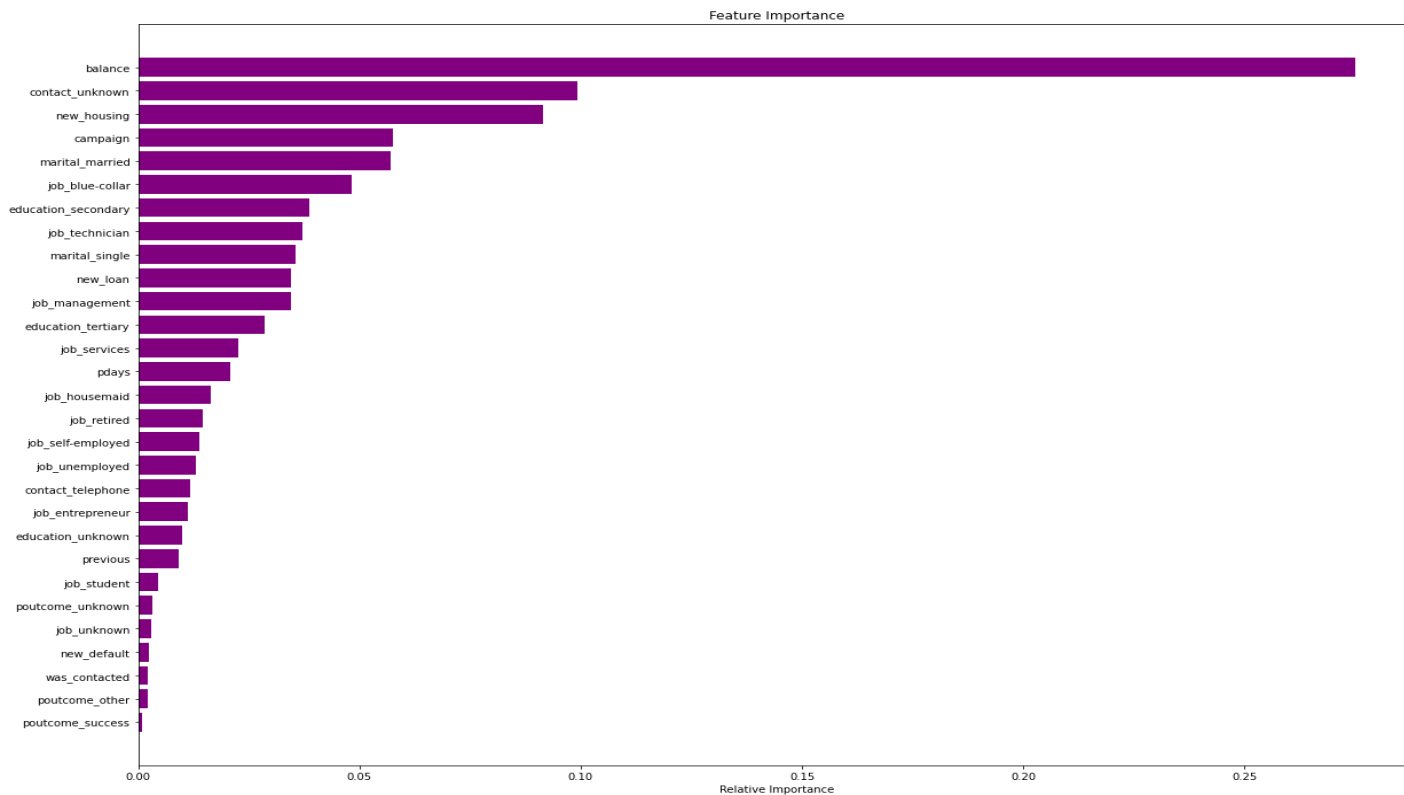
## Classification report

	precision	recall	F1-score	support
0	0.89	0.90	0.89	9085
1	0.90	0.89	0.89	9266
Accuracy			0.89	18351
Macro avg	0.89	0.89	0.89	18351
weighted avg	0.89	0.89	0.89	18351

## Confusion matrix



# Random Forest Feature Importance



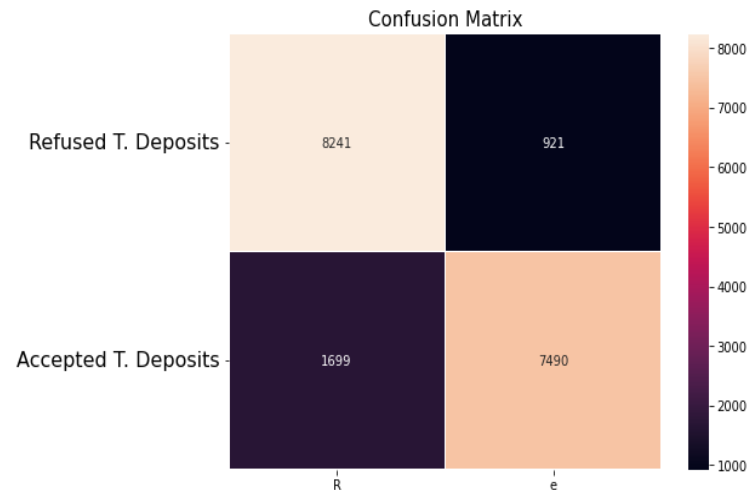
# Hyperparameter Tuning of Random Forest

Training accuracy	0.8596
Testing accuracy	0.8572

Classification report

	precision	recall	F1-score	Support
0	0.90	0.83	0.86	9940
1	0.82	0.89	0.85	8411
Accuracy			0.86	18351
Macro avg	0.86	0.86	0.86	18351
weighted avg	0.86	0.86	0.86	18351

Confusion matrix



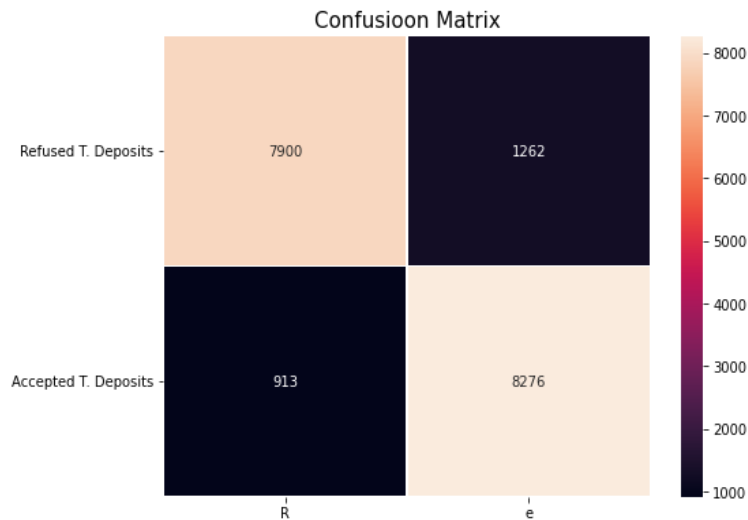
# Decision Tree Classifier

Training accuracy	0.9917
Testing accuracy	0.8814

## Classification report

	precision	recall	F1-score	Support
0	0.86	0.90	0.88	8813
1	0.90	0.87	0.88	9538
Accuracy			0.88	18351
Macro avg	0.88	0.88	0.88	18351
weighted avg	0.88	0.88	0.88	18351

## Confusion matrix



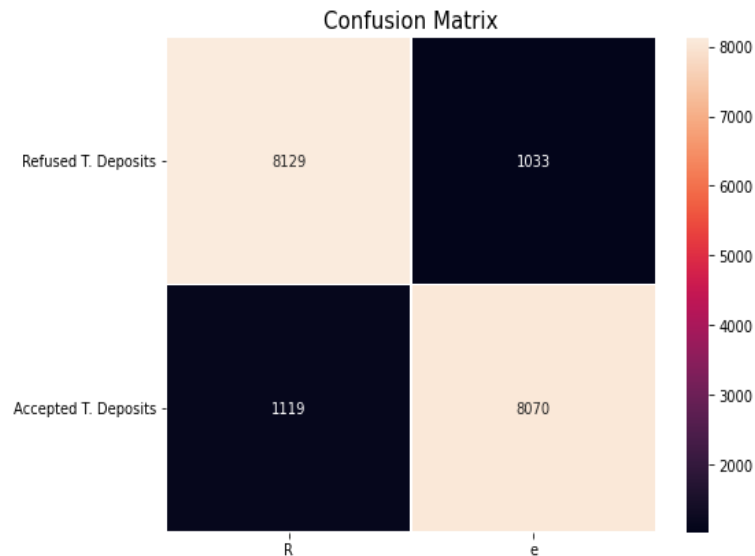
# K-Nearest Neighbors (KNN)

Training accuracy	0.9141
Testing accuracy	0.8827

## Classification report

	precision	recall	F1-score	Support
0	0.89	0.88	0.88	9248
1	0.88	0.89	0.88	9103
Accuracy			0.88	18351
Macro avg	0.88	0.88	0.88	18351
weighted avg	0.88	0.88	0.88	18351

## Confusion matrix



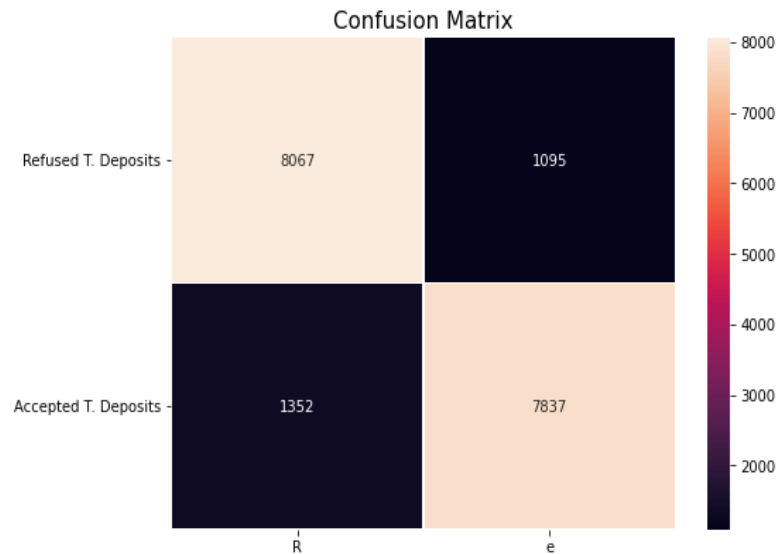
# XGBoost Classifier

Training accuracy	0.8687
Testing accuracy	0.8666

## Classification report

	precision	recall	F1-score	Support
0	0.88	0.86	0.87	9419
1	0.85	0.88	0.86	8932
Accuracy			0.87	18351
Macro avg	0.87	0.87	0.87	18351
weighted avg	0.87	0.87	0.87	18351

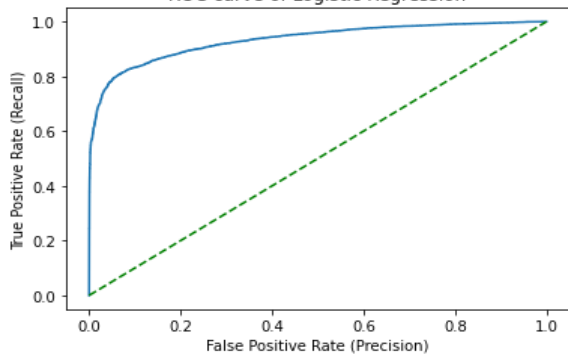
## Confusion matrix



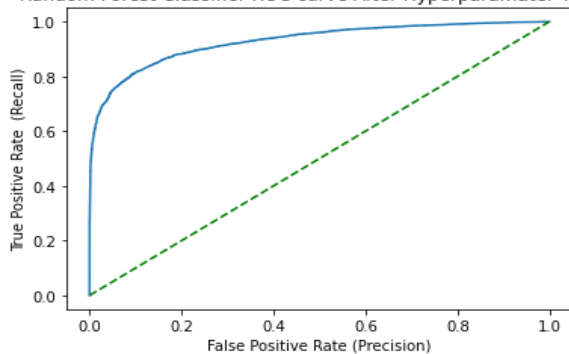


# ROC AUC Curve of different models

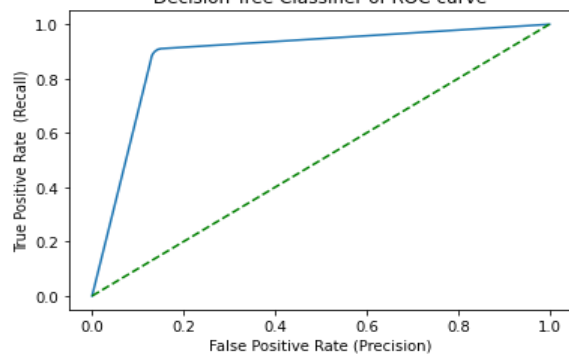
ROC curve of Logistic Regression



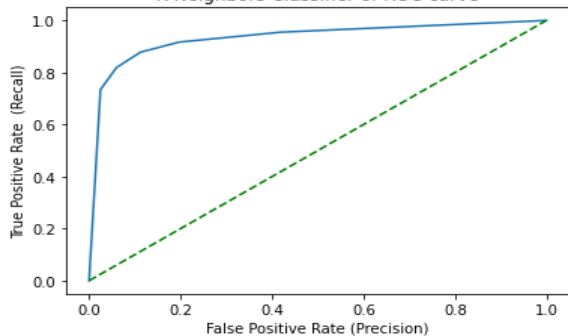
Random Forest Classifier ROC curve After Hyperparameter Tuning



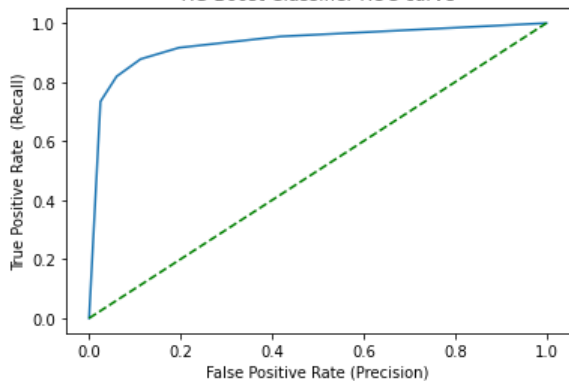
Decision Tree Classifier of ROC curve



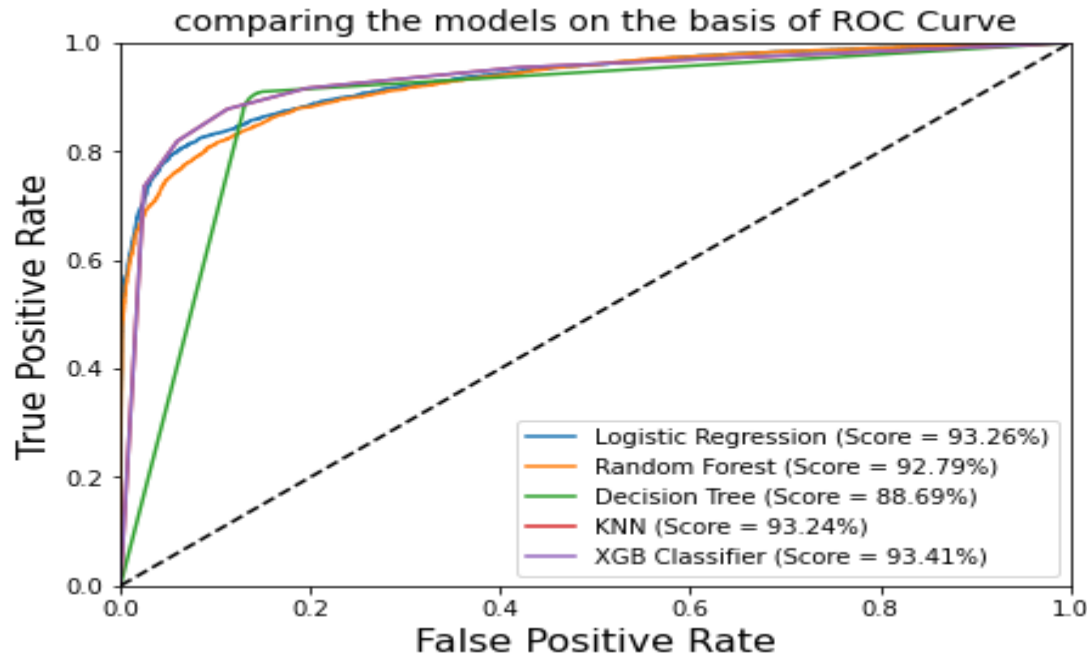
K Neighbors Classifier of ROC curve



XG Boost Classifier ROC curve



# Comparing ROC AUC curves and it's performance



# **CONCLUSION**

- **Blue-collar, management and technician showed maximum interest in subscription.**
- **Divorce people have no interest in term deposit.**
- **The majority of the customers are between the ages of 30 to 40.**
- **Most people have home loans, but only a small percentage of them choose term deposits.**
- **The outcome of the campaign is significantly influenced by the customer's account balance. We can then interact with those customers who have a balanced account balance.**
- **The model can assist us to identify customers based on whether they have made deposits or not.**
- **Instead of wasting time on the wrong customer, the model helps to target the right one.**
- **After implementation all the ML models we get maximum accuracy and ROC-AUC score in XGBoost. So we can conclude that it is the best model for us.**

**Thank you**