Capstone Project - III

Supervised ML- Classification

Bank Marketing Effectiveness Prediction

BY

Yashwant B. Raul

(yashwantraul24@gmail.com)

Mayur S. Marathe

(marathemayu1990@gmail.com

Sanket Gawali

(sanketgawali23@gmail.com)

Content

- Project Overview
- Feature Summary
- Exploratory Data Analysis
  - **Analysys of Categorical Variable**
  - **Analysis Of Continuous Variable**
  - **Analysis Of Dependent Variable**
  - **Correlation Analysis**
- Model Building
- Conclusion

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 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).

Feature Summary

- ✔ **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','unknow')
- ✔ 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')

Feature Summary

- ✔ **Related with the last contact of the current campaign:**
- ✔ 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')

Analysis Of Categorical Variable

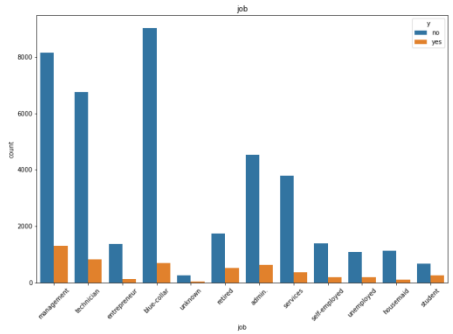
Categorical Variable = Job, Marital, Education, Default, Housing, Loan, Contact, Month, Poutcome, Y

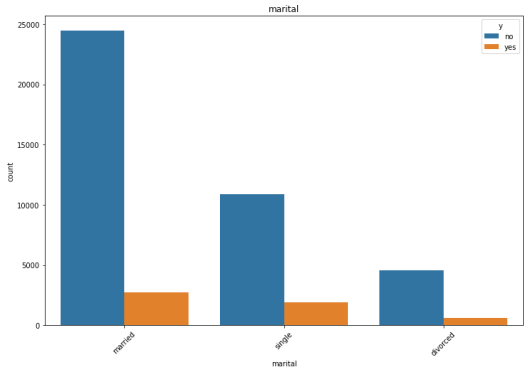
Analysis Of Categorical Variable

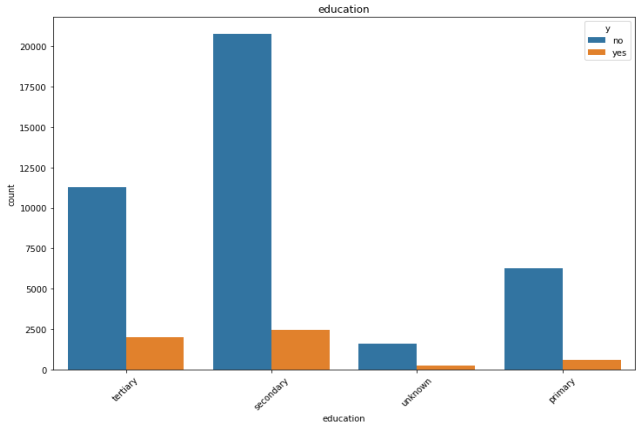
- **Management and blue-collar** has the highest distribution and **Housemaid and students** has the lowest.
- We have maximum data of **married people**.
- **Secondary** and **tertiary** education background clients are highest in the dataset.
- Client who has no credit in default is maximum and who has is very low near to 1 %.
- Client with **No personal loan** are more in dataset.
- Previous outcome is unknown in maximum cases might be the because there was no proper reason given for the same.
- We have maximum data available for the month of **May, june, july** and **august** and very less in Dec.
- **Y** which is our **target variable** : we can see there are more no results than yes.
- As it is **classification problem** and we have class imbalance which is the problem we have to solve this class imbalance before training model.

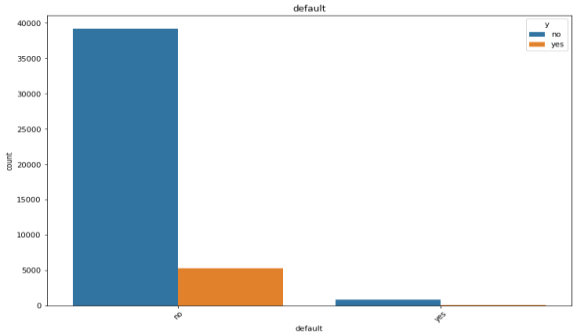
Analysis Of Categorical Variable

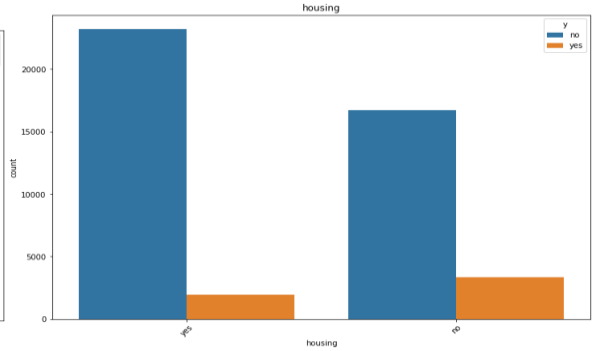
Relation Between Categorical Variable And Target Variable

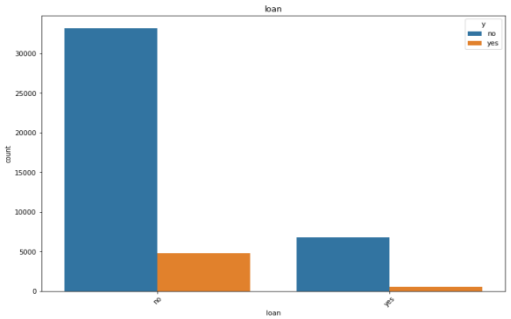


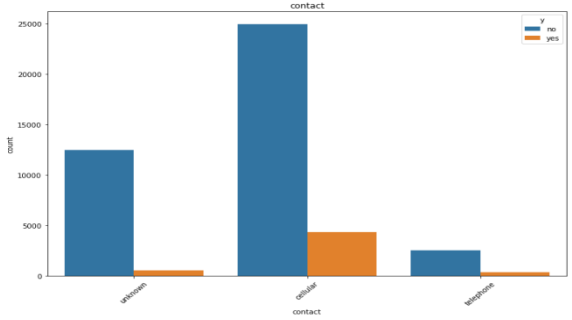


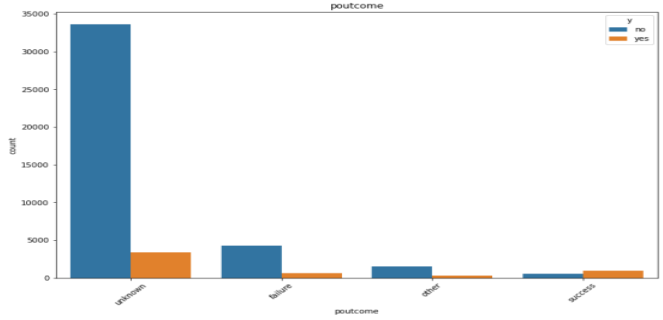






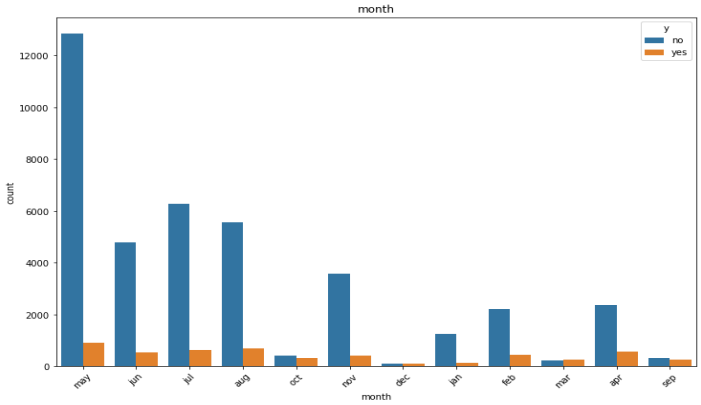


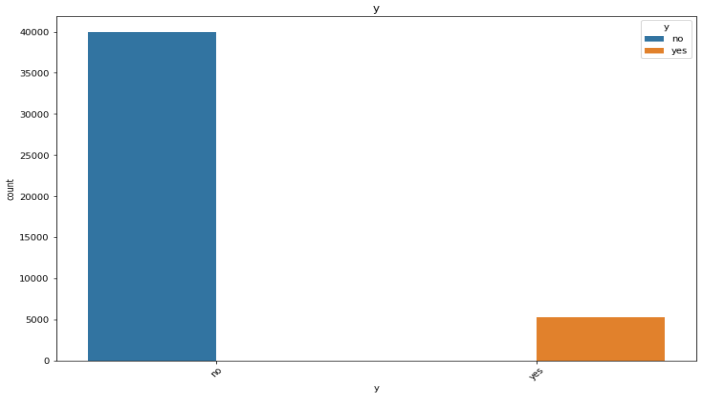




Analysis Of Categorical Variable

Relation Between Categorical Variable And Target Variable





Clients who has **housing loan** seems to be not interested much on deposit.

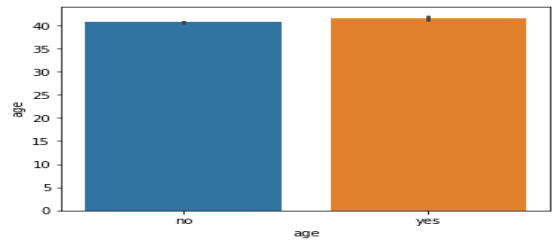
If pre campaign **outcome(i.e poutcome)**=success then, there is high chance of client to show interest on **term deposit**.

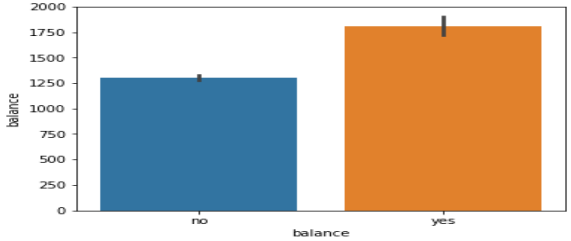
Married and **Single** have more interest in deposit.

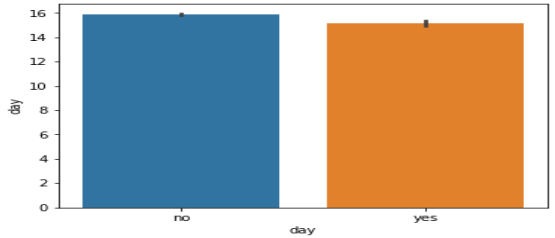
Cellular communication is more effective in comparison to other communication types.

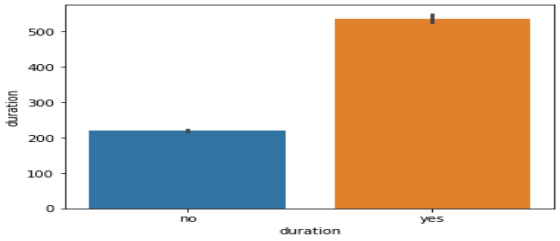
Analysis Of Continuous Variable

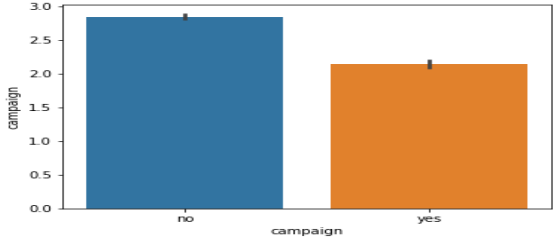
Continuous Variable = Age, Balance, Day, Duration, Campaign, Pdays, Previous

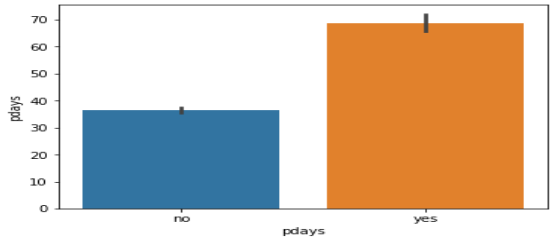


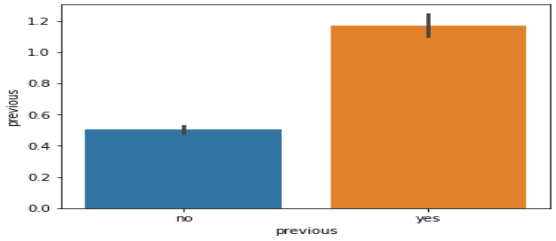












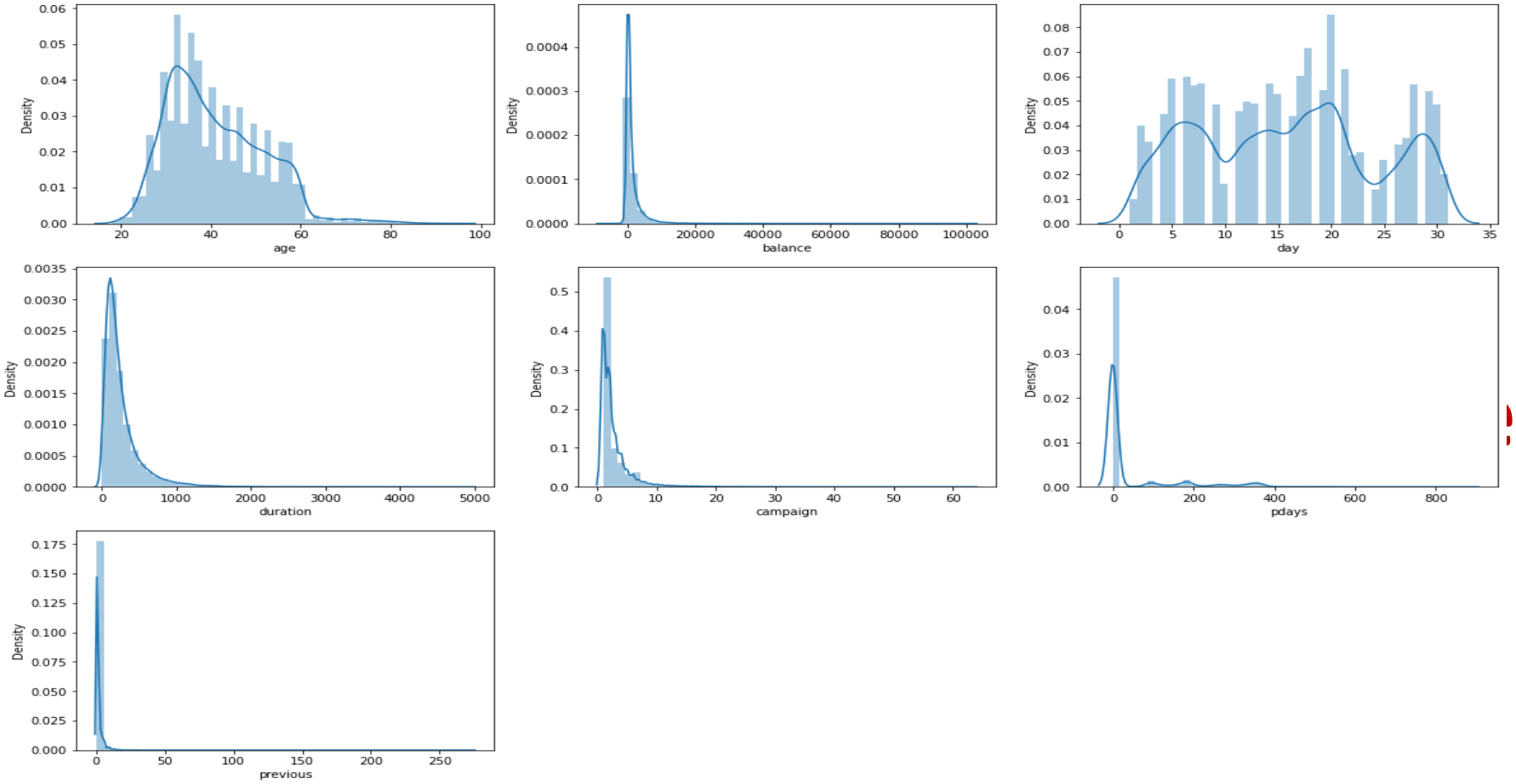
Client shows interest on **response**(term deposit) who had discussion for longer duration

Calls with large **duration** has more tendency for conversion

People were mostly **contacted** once but also some people also connect more number of time.

Analysis Of Continuous Variable

Distribution Of Data

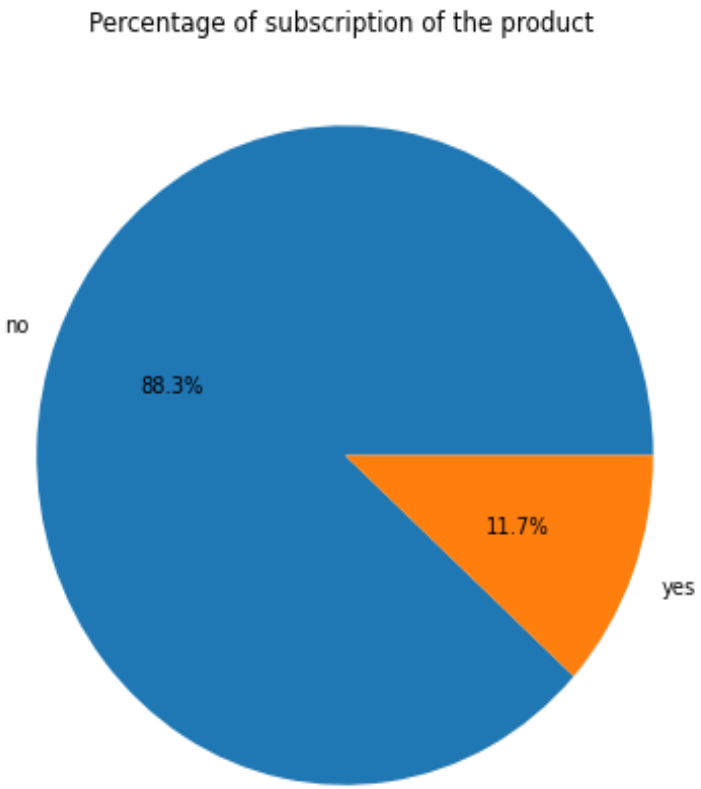


It seems **age** and **days** are Normally distributed.

Balance, Duration, Campaign, pdays and **previous** are skewed towards right and seems to have some outliers.

Analysis Of Dependent Variable: Y

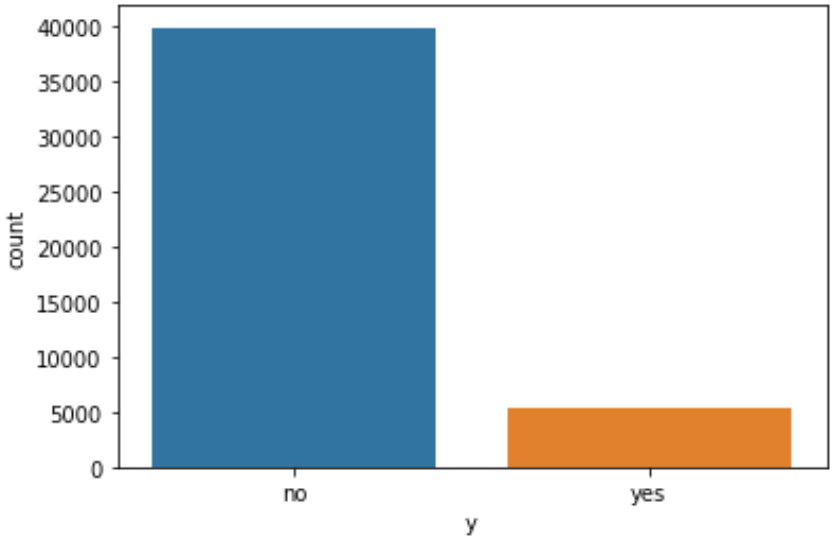
Available distribution of data



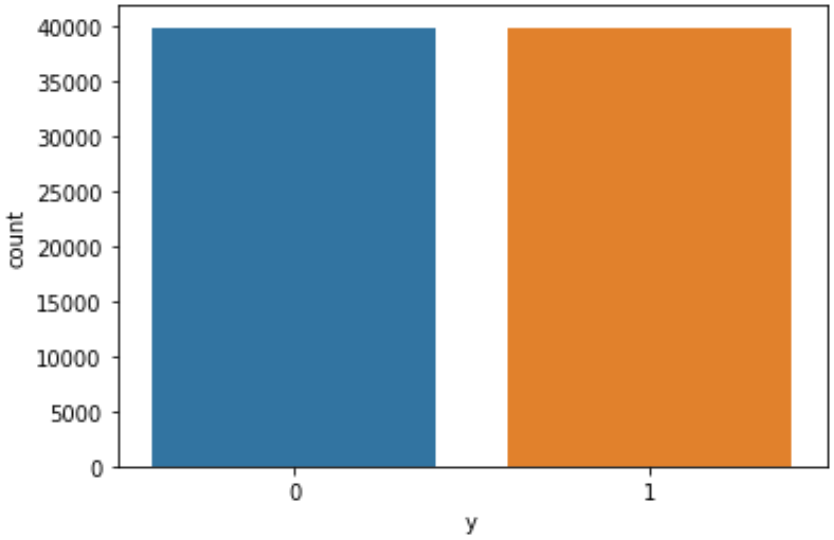
The rate of percentage of subscription of product is shown in fig.
88.3 % people not subscribed the product where 11.7% people subscribed the product
This is our target variable as we seen that here is **class imbalance problem**

Analysis Of Dependent Variable: Y

Imbalance Data

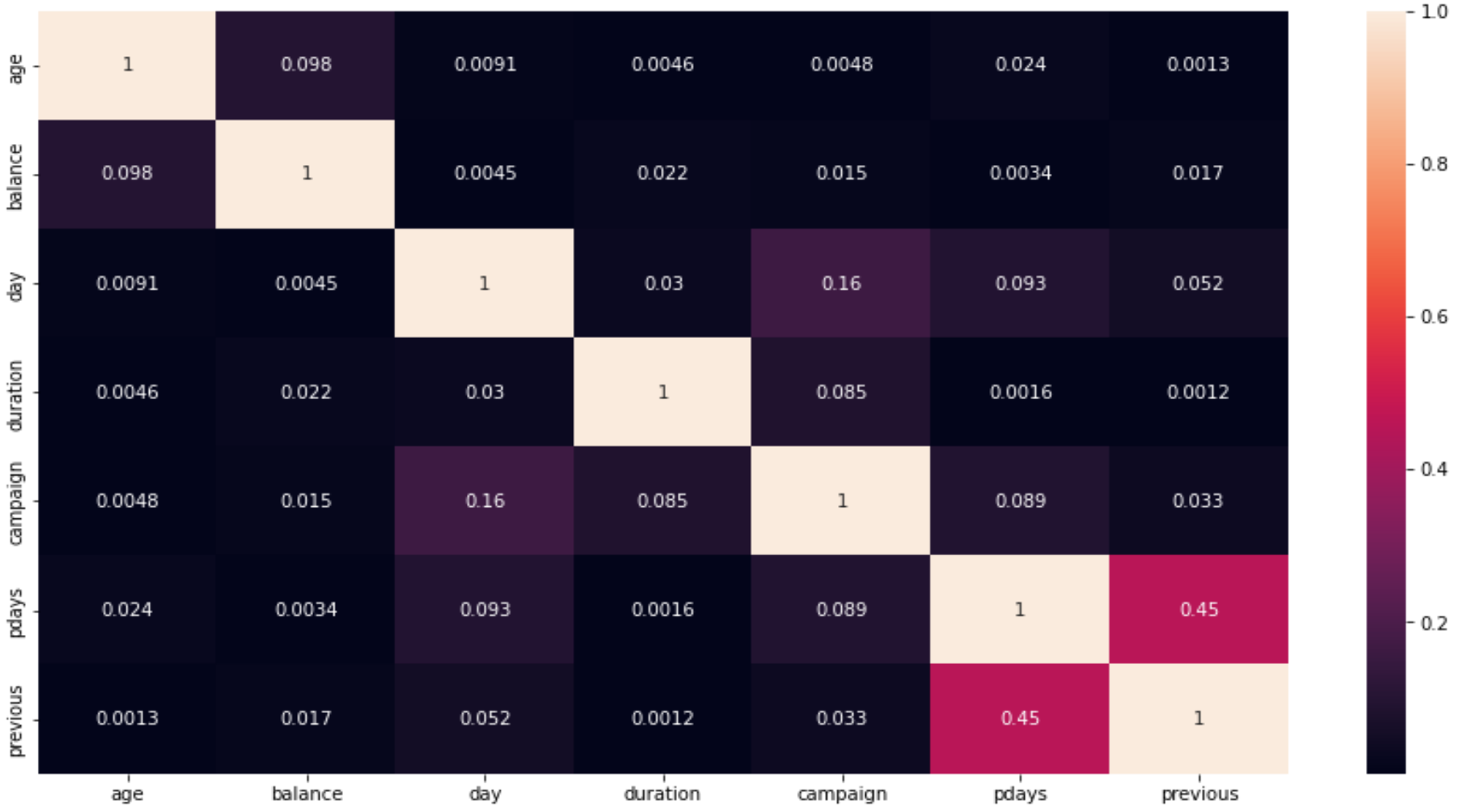


After Balancing



As it is **classification problem** and we have class imbalance which is the problem so we solve this class imbalance before training the model.

Correlation Analysis



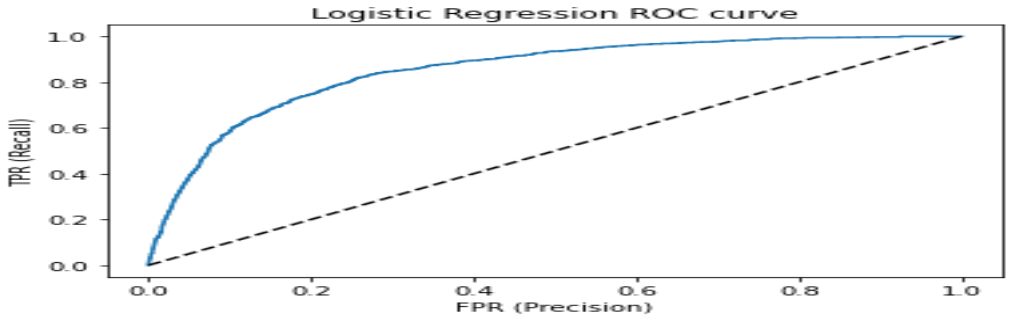
There is not any variable highly correlated.

Model Building

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier
- Gradient Boosting Classifier
- K Neighbours Classifier
- XGBoost
- Naïve Bayes Classifier

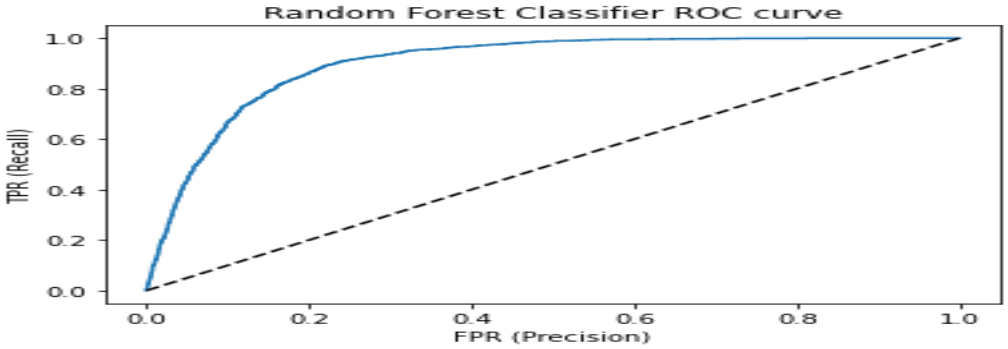
LOGISTIC REGRESSION

Training accuracy Score : 0.9294
Testing accuracy Score : 0.8790



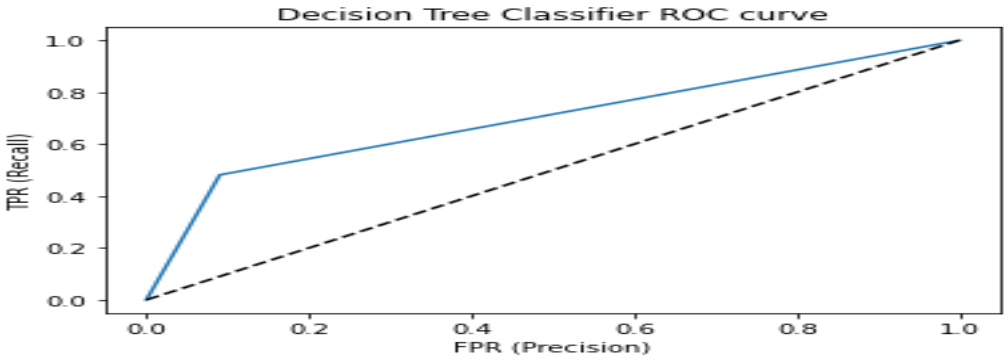
RANDOM FOREST CLASSIFIER

Training accuracy Score : 1.0
Testing accuracy Score : 0.89628



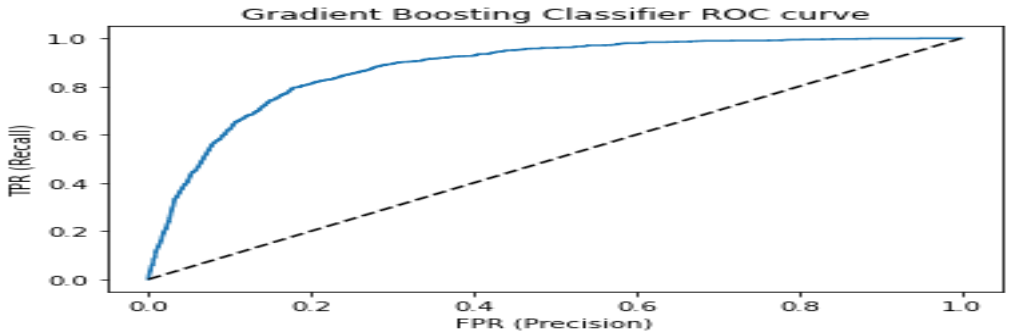
DECISION TREE CLASSIFIER

Training accuracy Score : 1.0
Testing accuracy Score : 0.8592



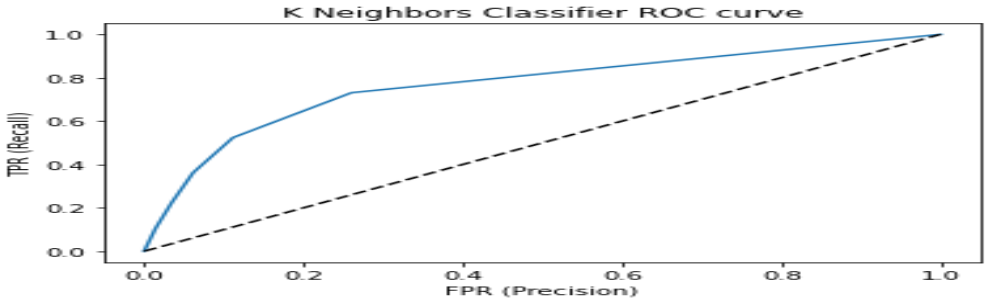
GRADIENT BOOSTING CLASSIFIER

Training accuracy Score : 0.9319
Testing accuracy Score : 0.8816



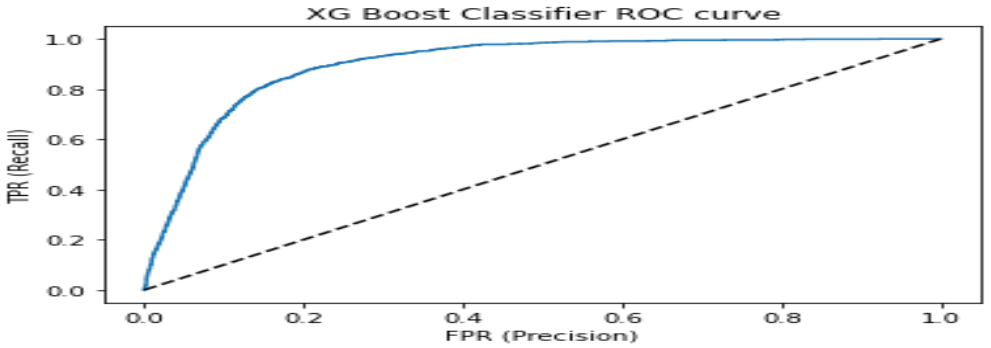
K NEIGHBOUR CLASSIFIER

Training accuracy Score : 0.9409
Testing accuracy Score : 0.8707



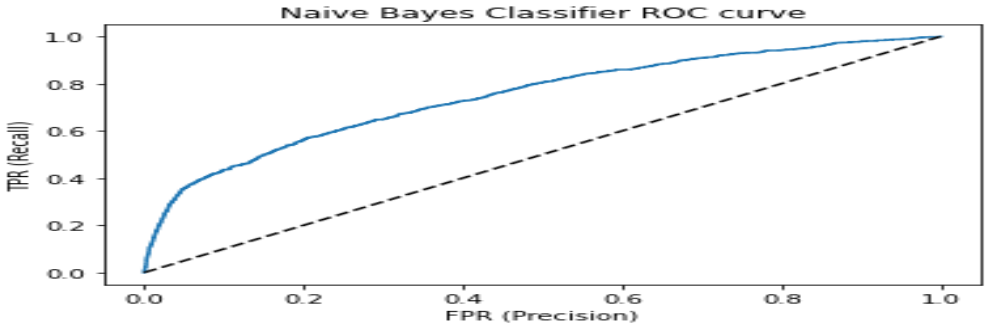
XGBOOST

Training accuracy Score : 0.9324
Testing accuracy Score : 0.8824

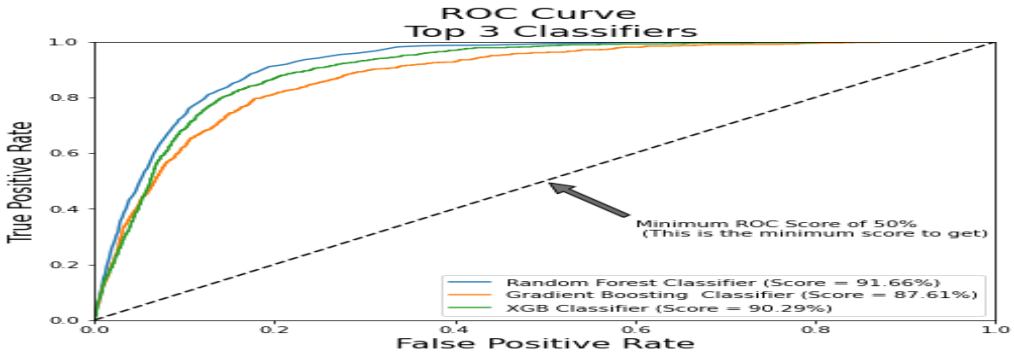


NAIVE BAYES ALGORITHM

Training accuracy Score : 0.8421
Testing accuracy Score : 0.8047



COMPARISON IN RANDOM FOREST CLASSIFIER, GRADIENT BOOSTING CLASSIFIER AND XGB CLASSIFIER



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