Capstone Project

Bank Marketing Effectiveness Prediction

Done By :-

Rajni Shukla

Saquib Neyaz

Pranali Dongre

Contents

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

Problem Statement

The data is related with direct marketing campaigns 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’) subscribed 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)

Desired target

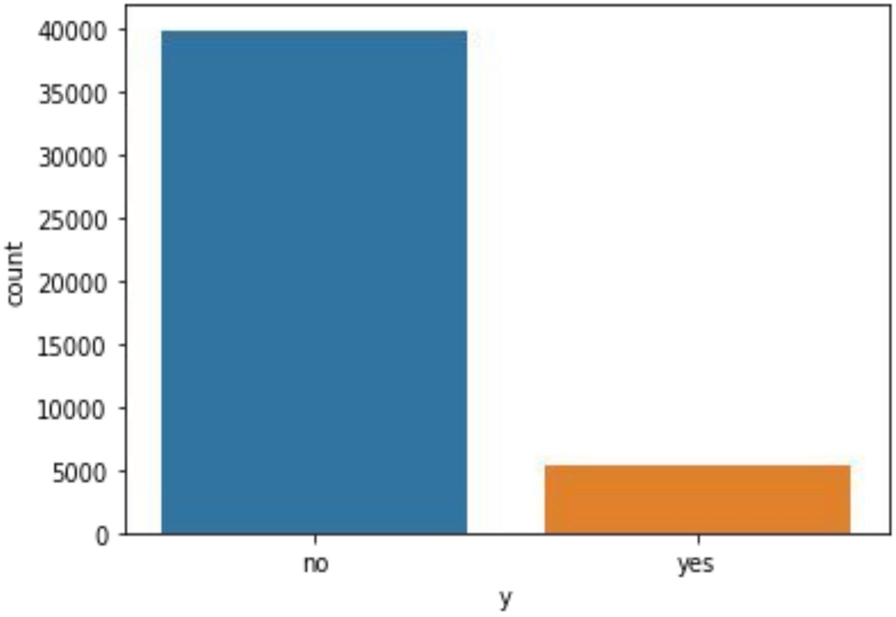
- y - has the client subscribed a term deposit? (binary: 'yes','no')

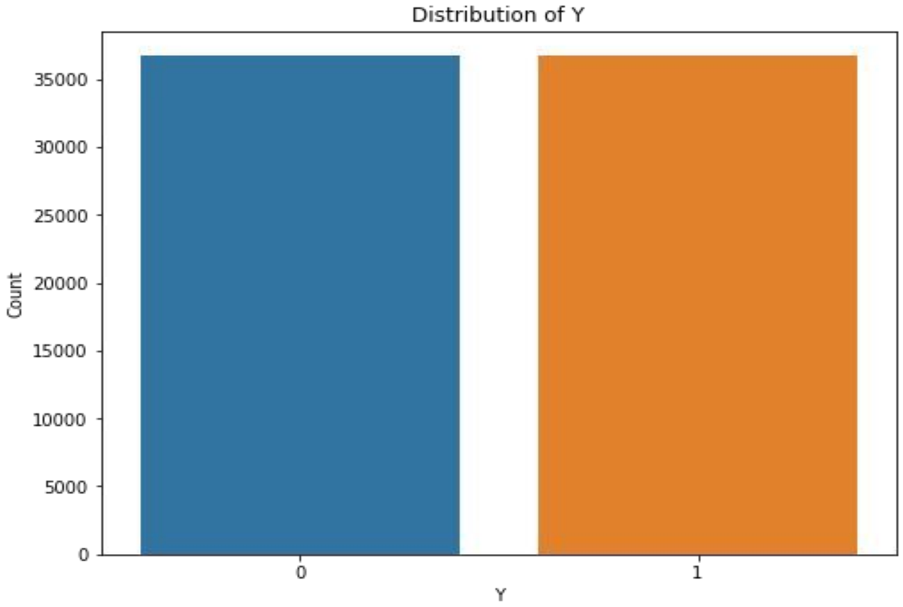
Numerical Features

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

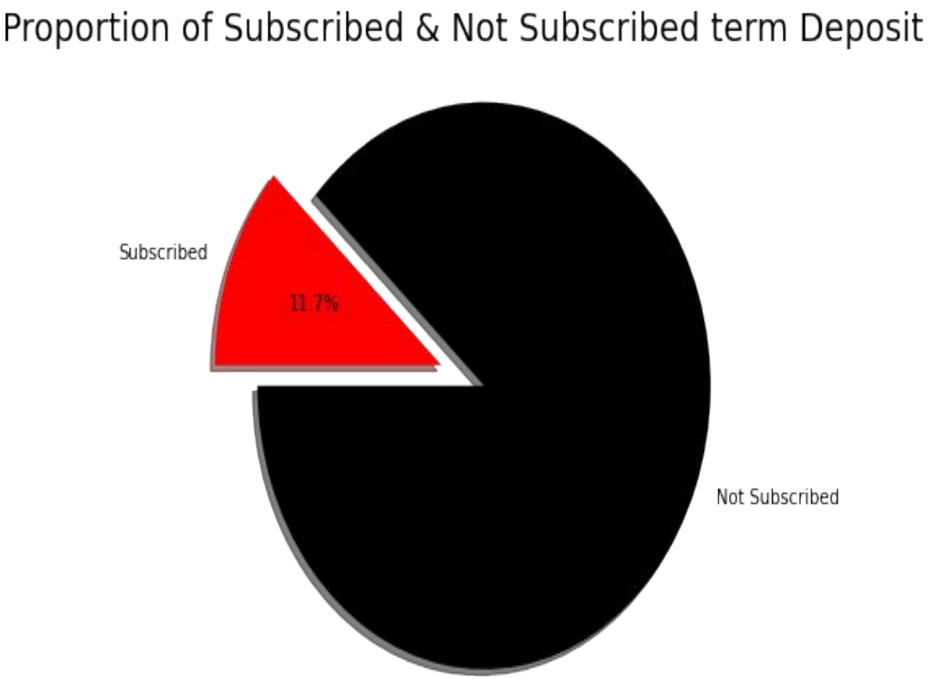
Exploratory Data Analysis (Target)

:- Before   :- After





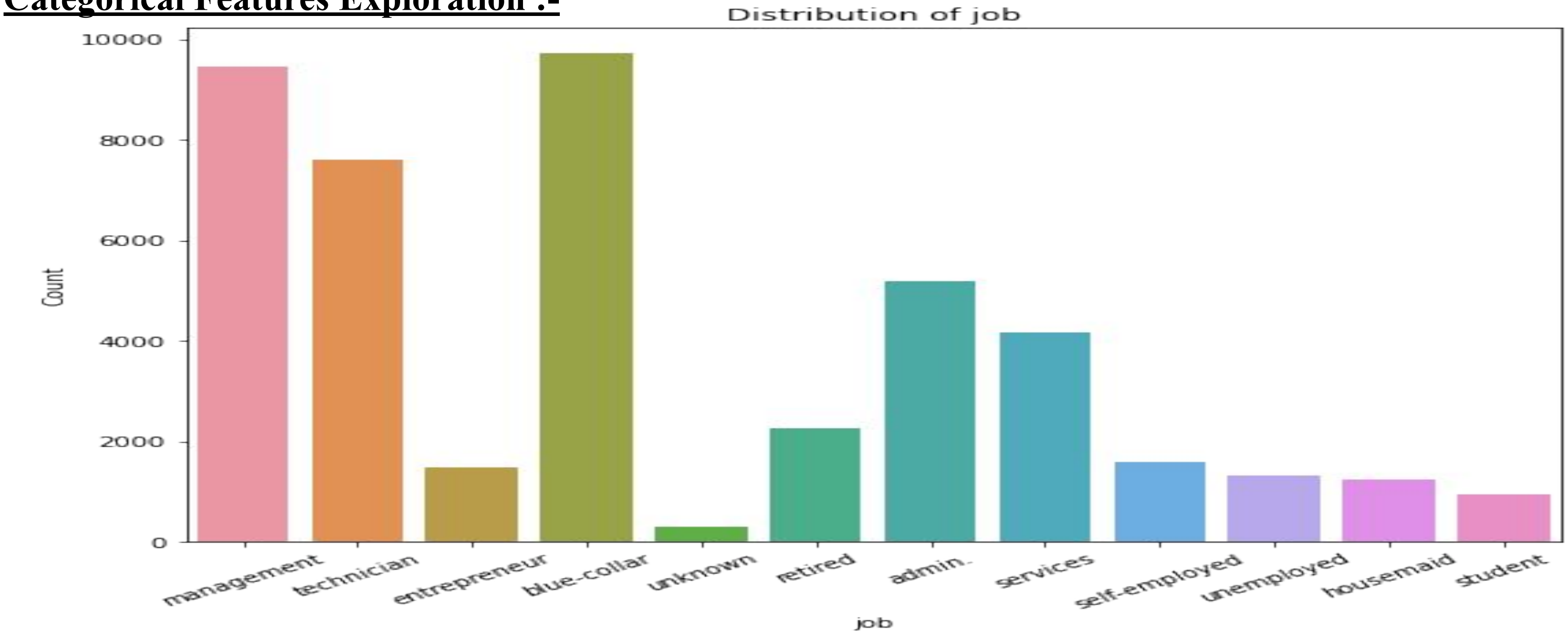
EDA(Continued…)



- From this data we can see that 88.3% customers did not subscribed for Term deposit.
- And also we can see 11.7% customer subscribed for Term deposit.
- We can say that the percentage of people subscribing to the term deposit is quite low.

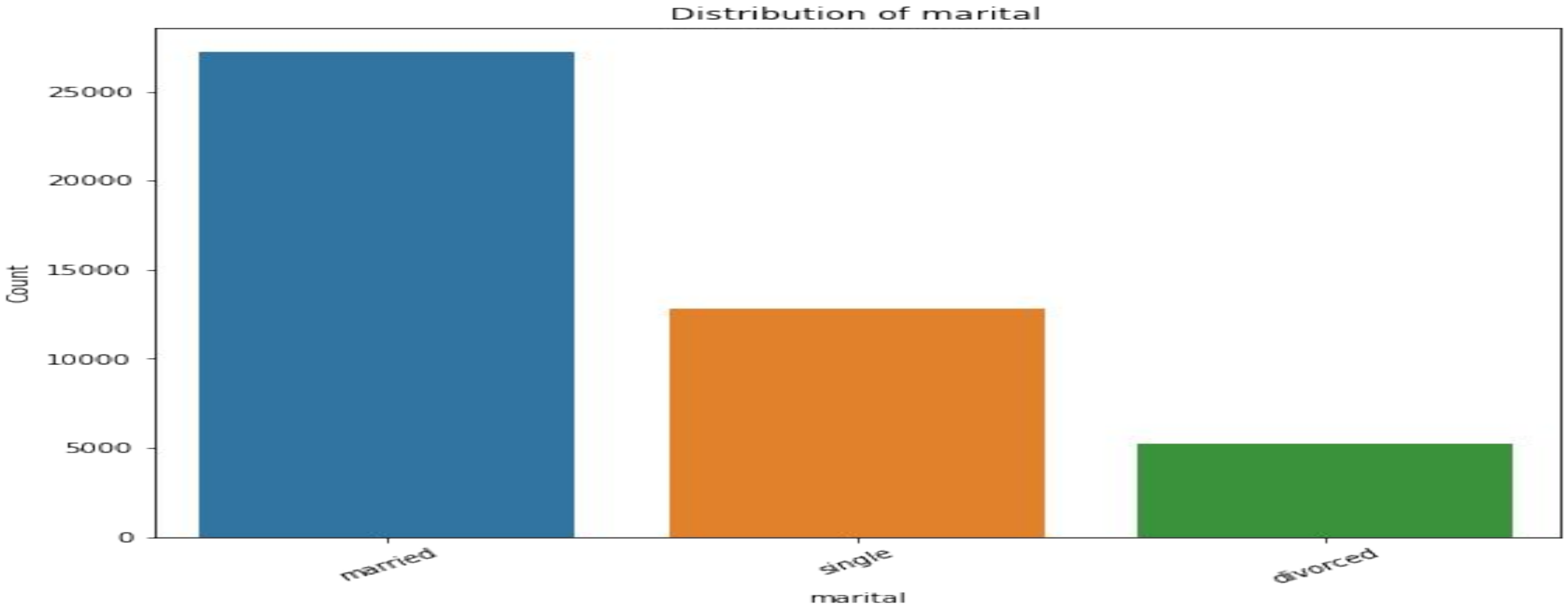
EDA(Continued…)

Categorical Features Exploration :-



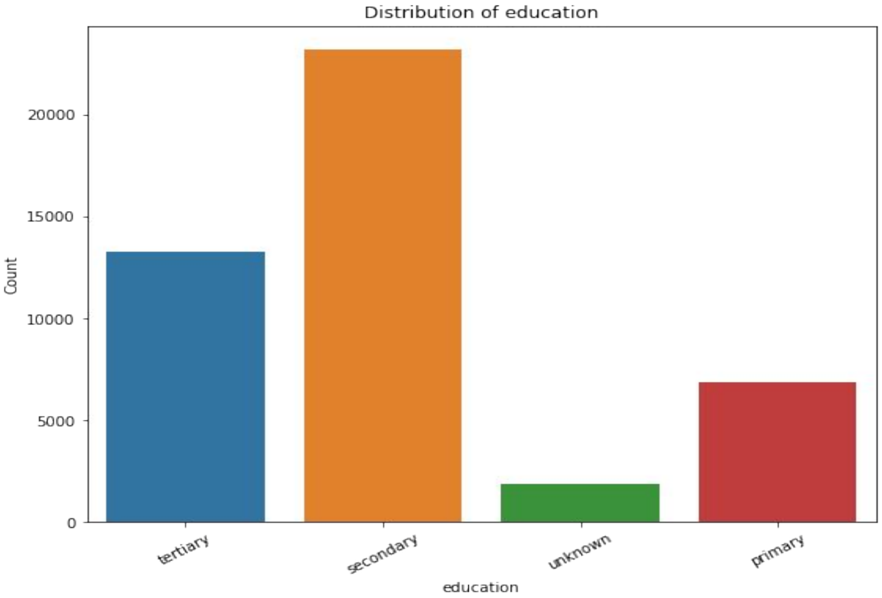
- Most of the customers have jobs as "management", "blue-collar" and "technician“.
- People with management jobs have subscribed more for the deposits.

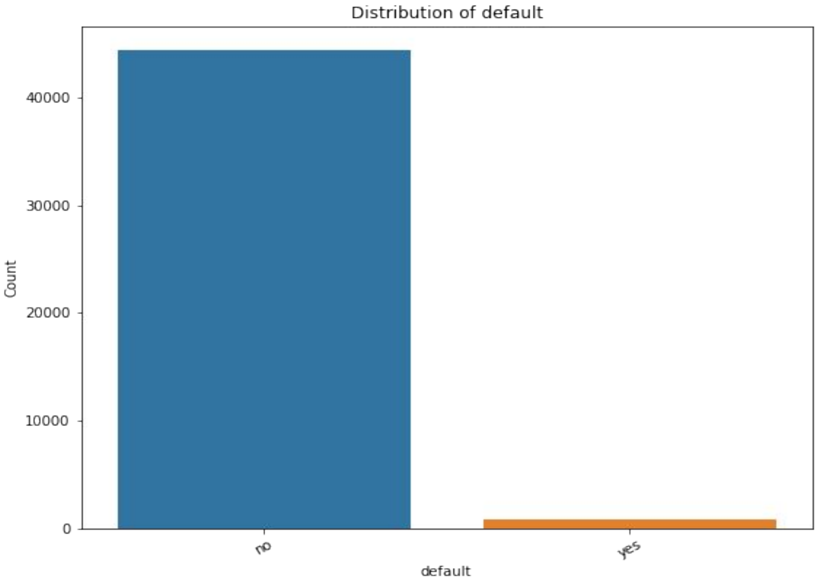
EDA(Continued…)



- Client who married are high in records.
- People who are married have subscribed for deposits more than people with any other marital status.

EDA(Continued…)

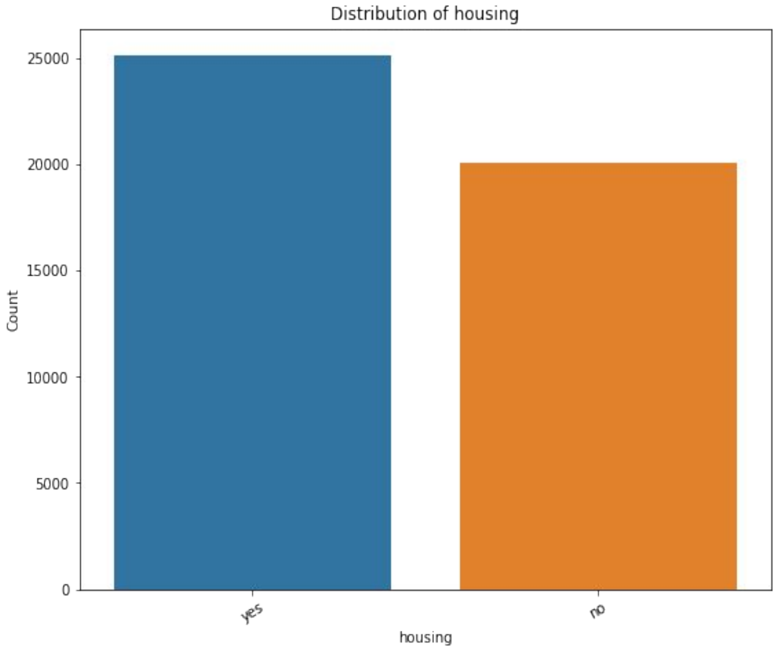




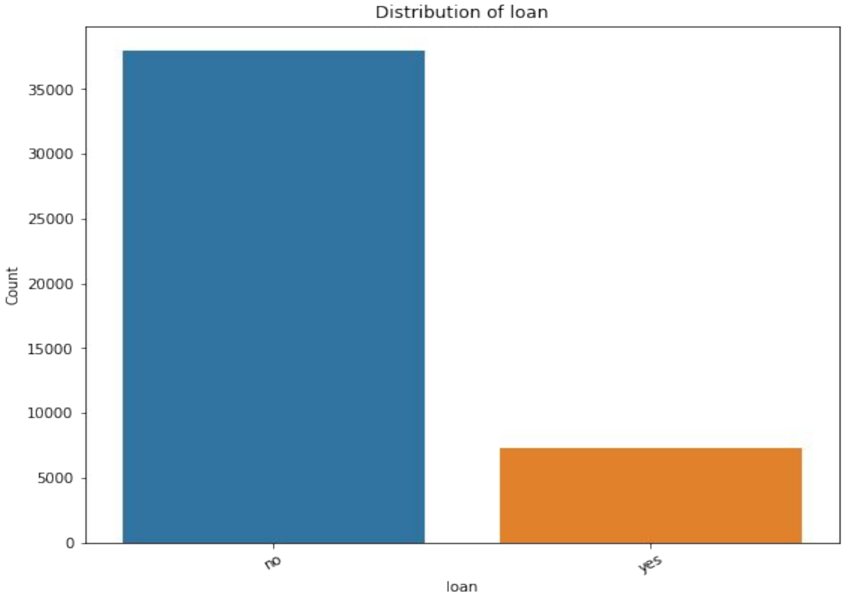
- Client whose education background is secondary are in high numbers.
- People with Secondary education qualification are the most who have subscribed for the deposits.

- Default feature seems to be does not play important role.
- People with default status as ‘no’ are the most ones who have not subscribed for bank deposits.

EDA(Continued…)

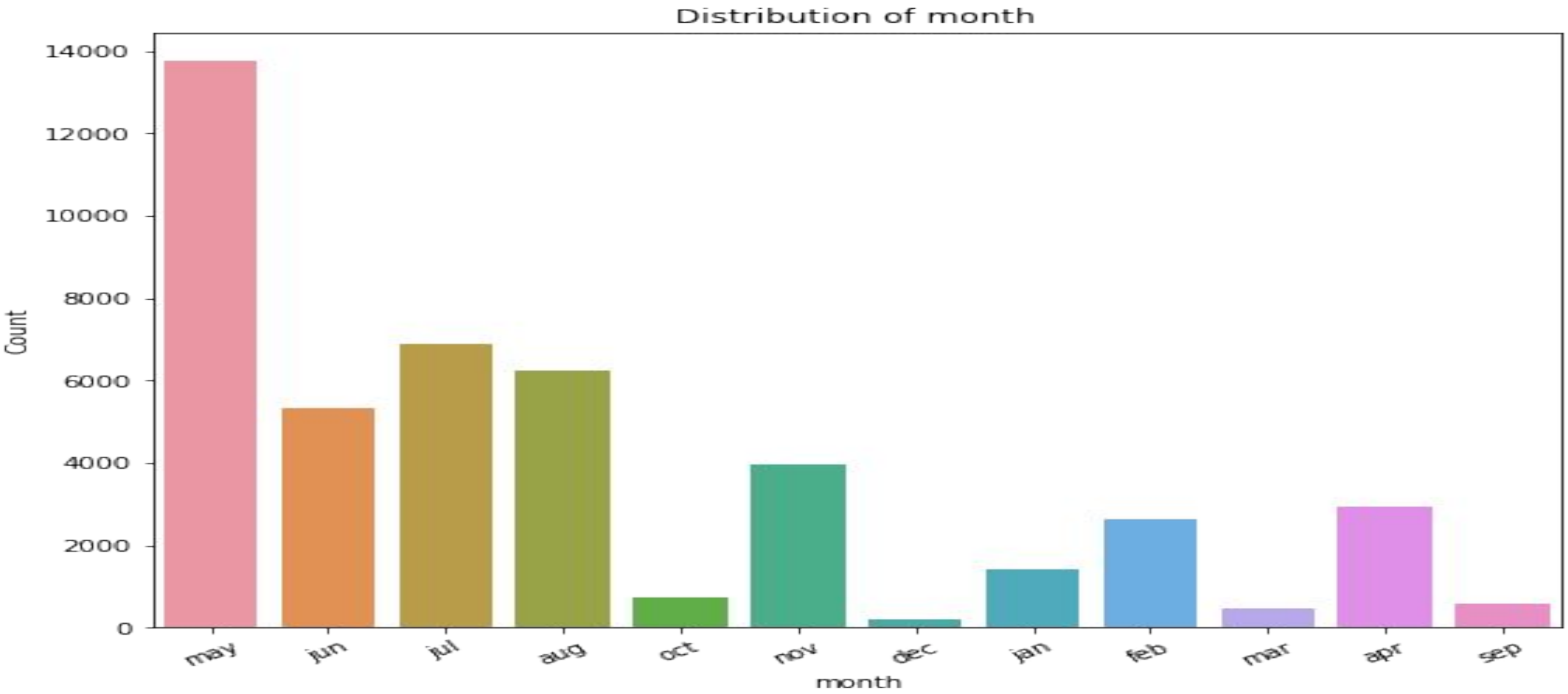


- People with housing loan are the most ones who have been contacted by the bank followed by people with no housing loan.
- Most of the client has taken the housing loan.



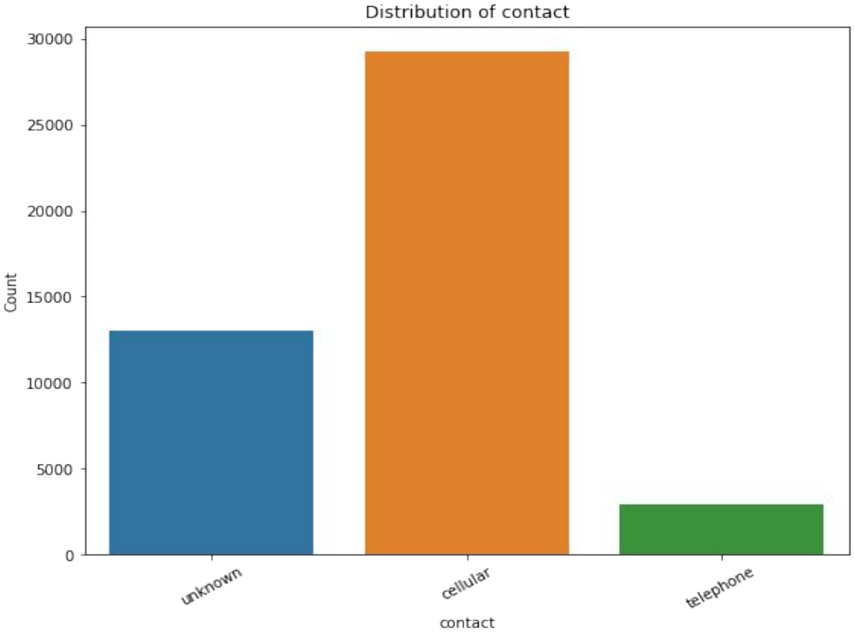
- People with no personal loan are the most ones who have been contacted by the bank for the deposits.
- People with no personal loan are the most ones who have not subscribed and are also the most ones who have subscribed for the deposits

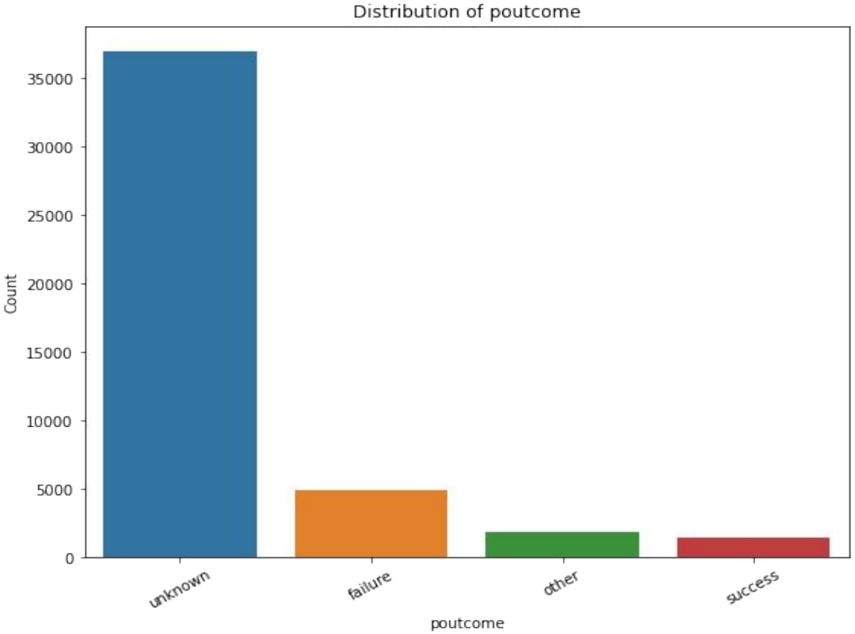
EDA(Continued…)



- Data in month of “May" is high and less in "Dec".
- The month of the highest level of marketing activity was the month of "May".

EDA(Continued…)

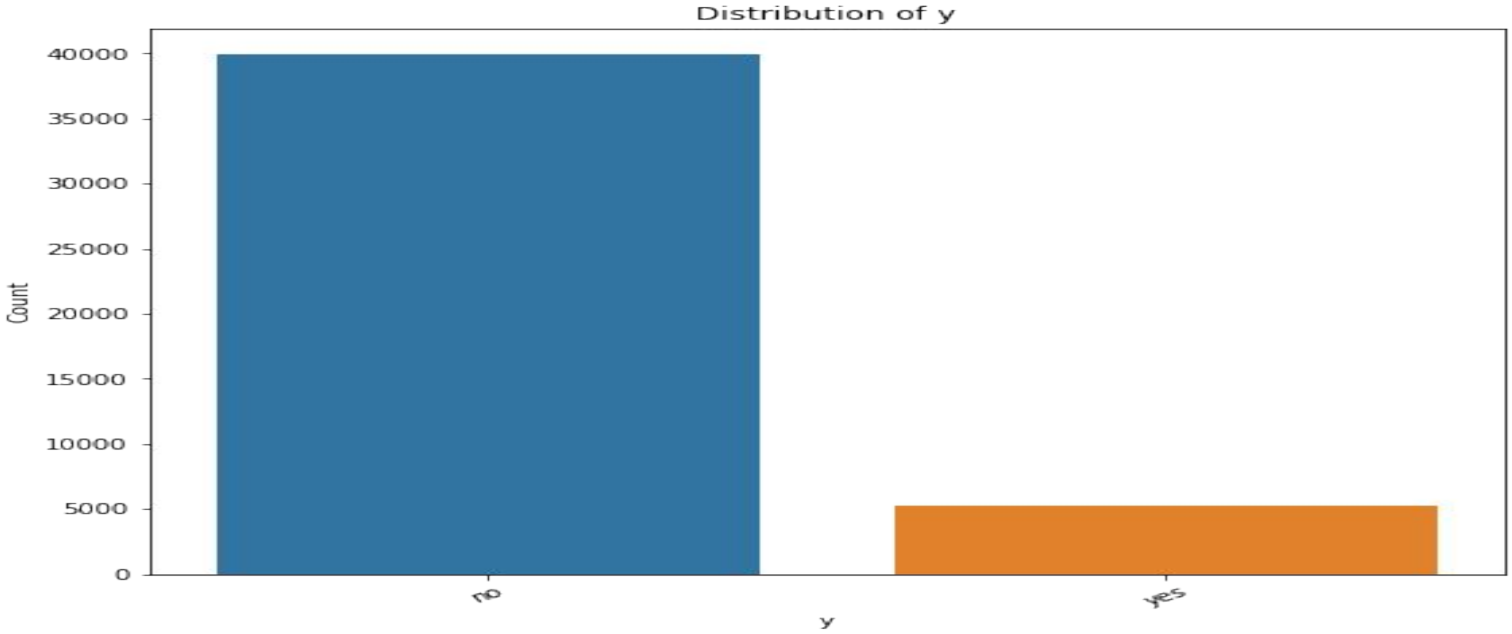




- Most people are contacted more in cellular than telephone.
- More people contacted on cellular by bank have subscribed the deposits.

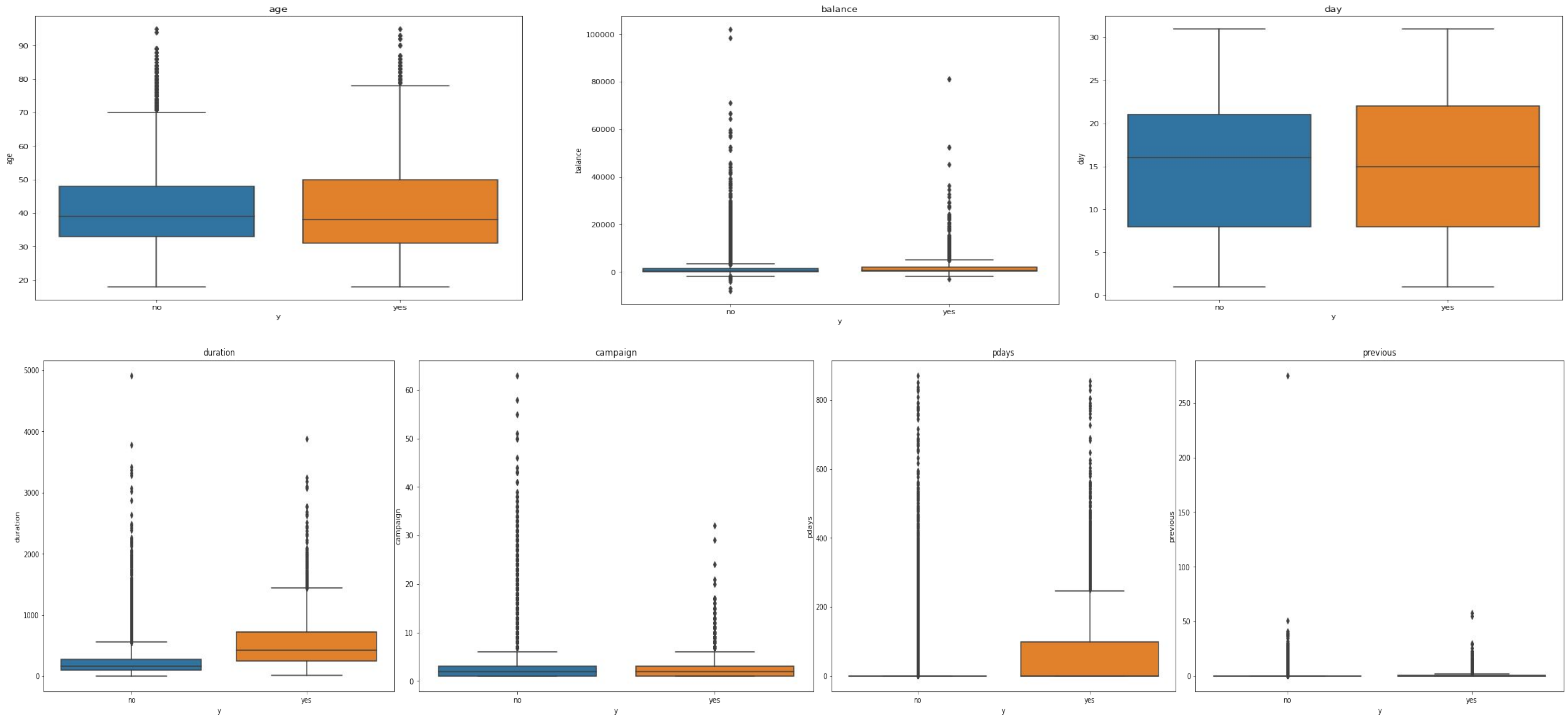
- Majority of the outcome of the previous campaign is Non-Existent.
- People whose previous outcome is non-existent have actually subscribed more.

EDA(Continued…)

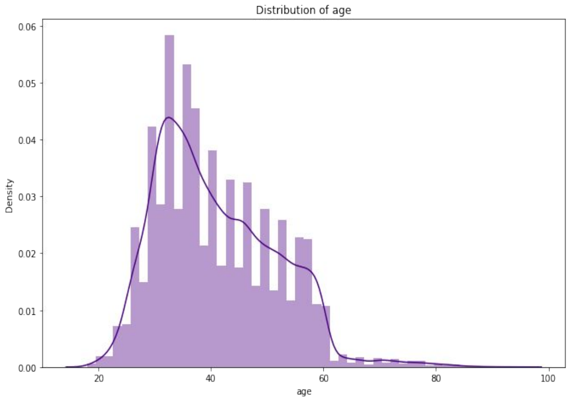


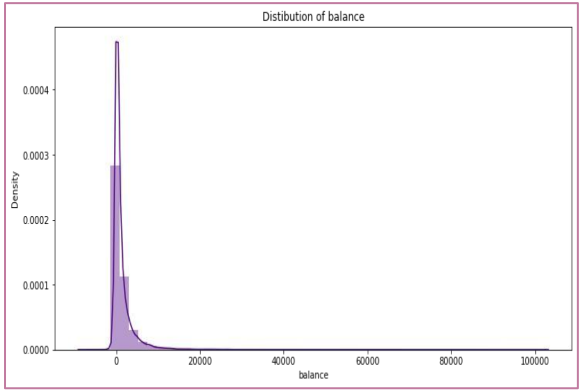
- Most people in Distribution of y count no more than yes ..

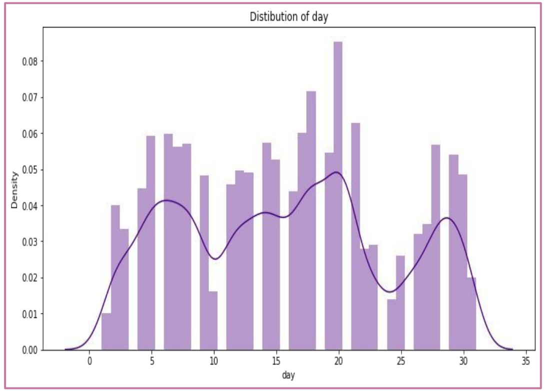
EDA(Continued…)

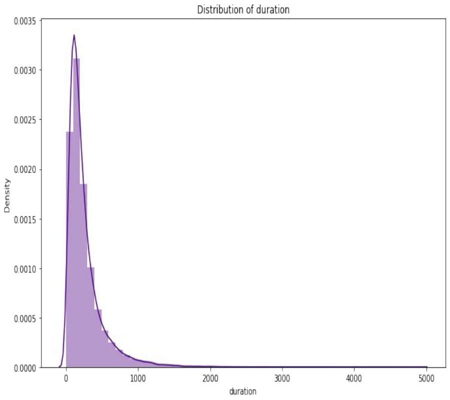


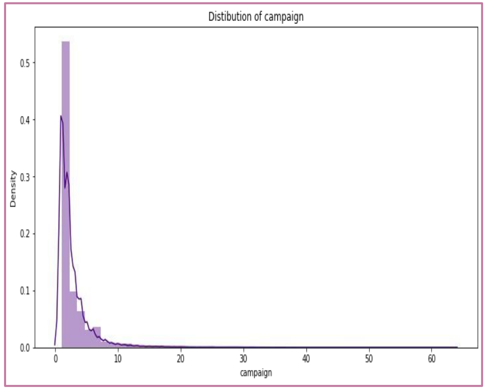
EDA(Continued..)

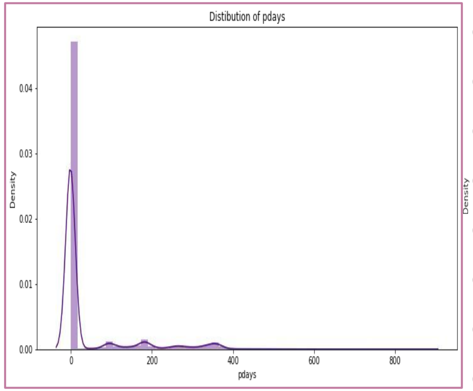

Distribution of age

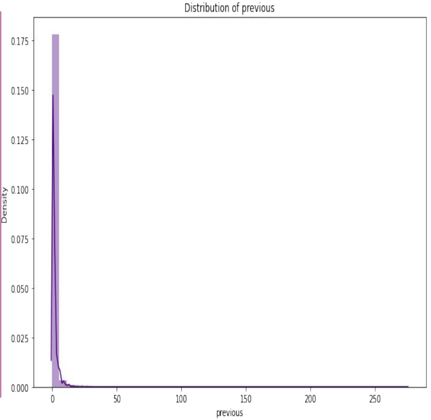

Distibution of balance


Distibution of day

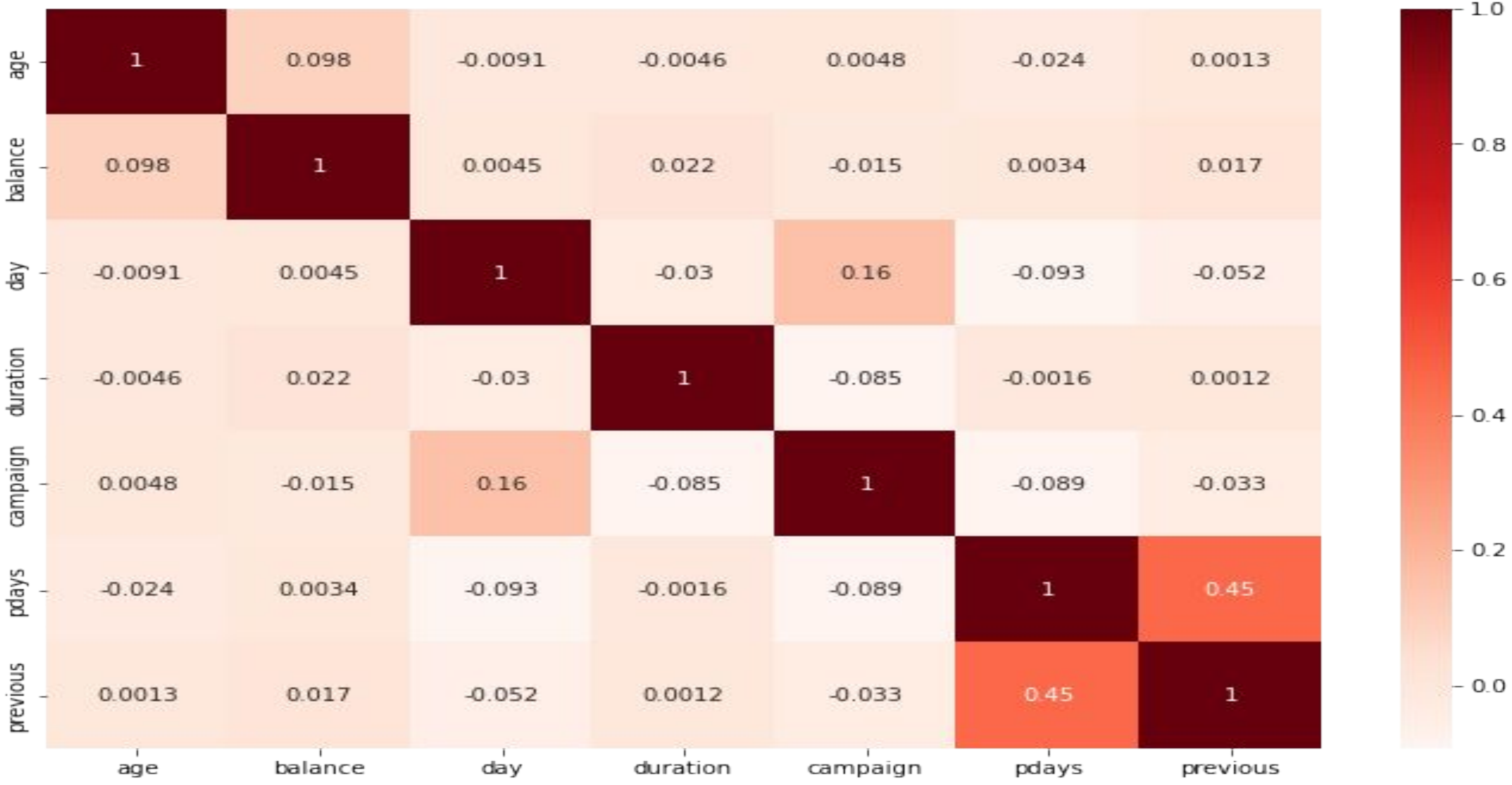

Distribution of duration


Distibution of campaign


Distibution of pdays


Distribution of previous

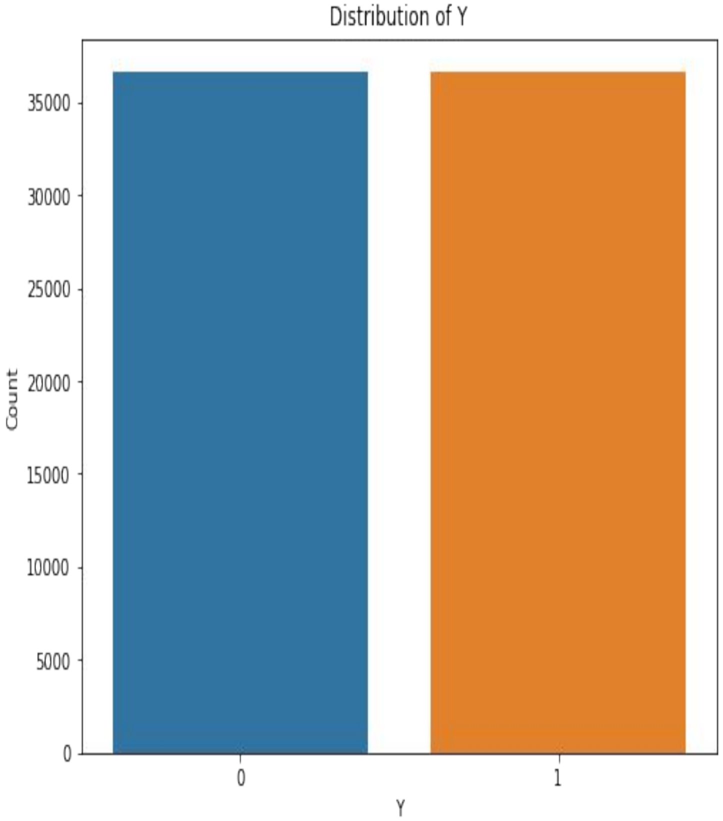
Correlation



	age	balance	day	duration	campaign	pdays	previous
age	1	0.098	-0.0091	-0.0046	0.0048	-0.024	0.0013
balance	0.098	1	0.0045	0.022	-0.015	0.0034	0.017
day	-0.0091	0.0045	1	-0.03	0.16	-0.093	-0.052
duration	-0.0046	0.022	-0.03	1	-0.085	-0.0016	0.0012
campaign	0.0048	-0.015	0.16	-0.085	1	-0.089	-0.033
pdays	-0.024	0.0034	-0.093	-0.0016	-0.089	1	0.45
previous	0.0013	0.017	-0.052	0.0012	-0.033	0.45	1

SMOTE :-

- One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short.



Model Implementation

Logistic Regression :-

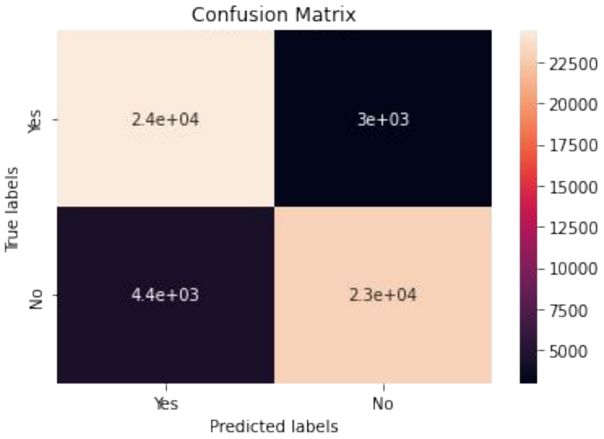
Best Parameter :-

c : 0.1

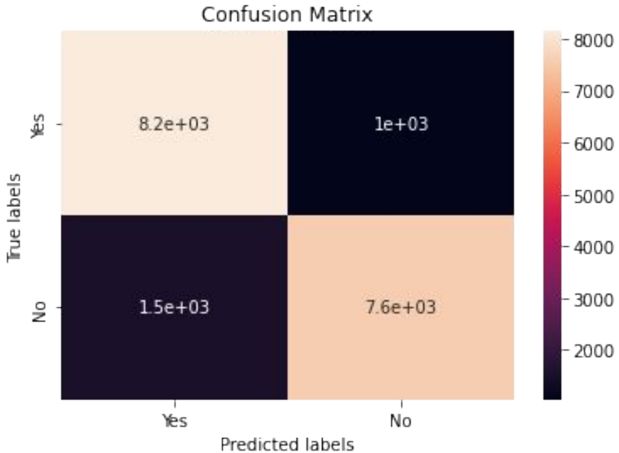
ROC-AUC Score :-

Train Data	0.94
Test Data	0.93

Confusion matrix of Train Data



Confusion matrix of Test Data



K –Nearest Neighbors :-

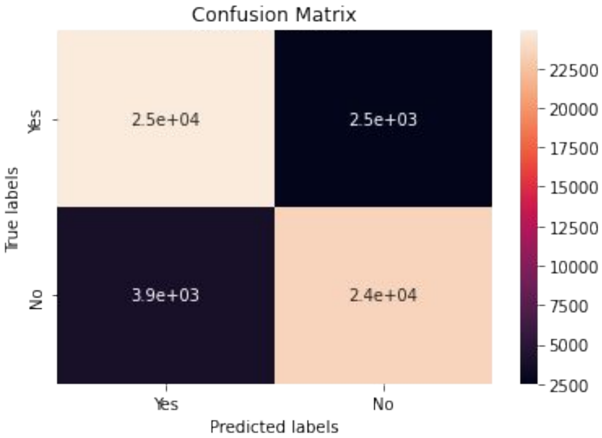
Best Parameter :-

n_neighbors : 23

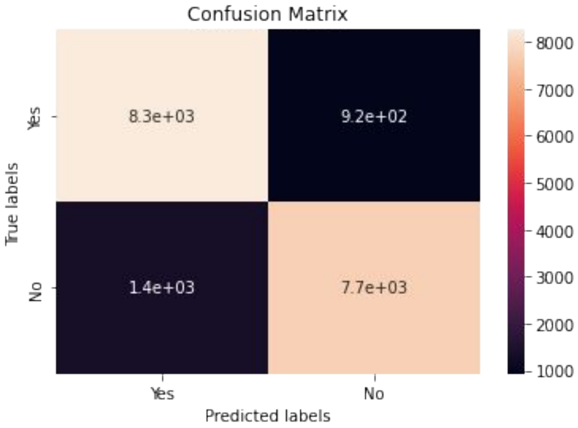
ROC-AUC Score :-

Train Data	0.95
Test Data	0.94

Confusion matrix of Train Data



Confusion matrix of Test Data



XGBoost Classifier :-

Best parameters

learning_rate : 0.5

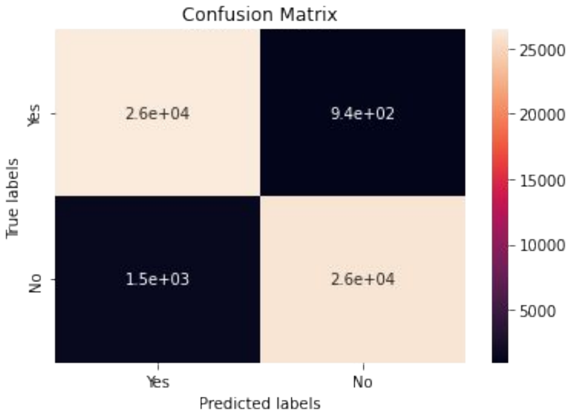
max_depth : 9

n_estimators : 125

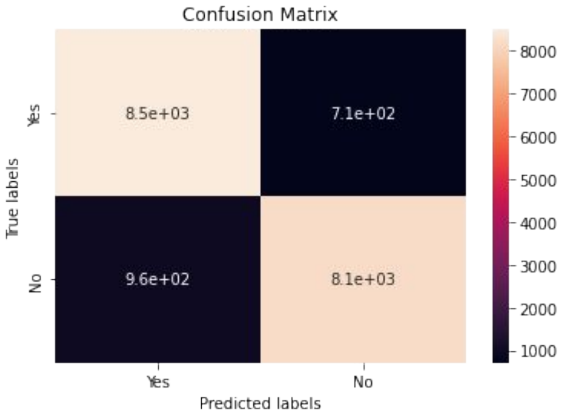
ROC-AUC Score :-

Train Data	0.99
Test Data	0.96

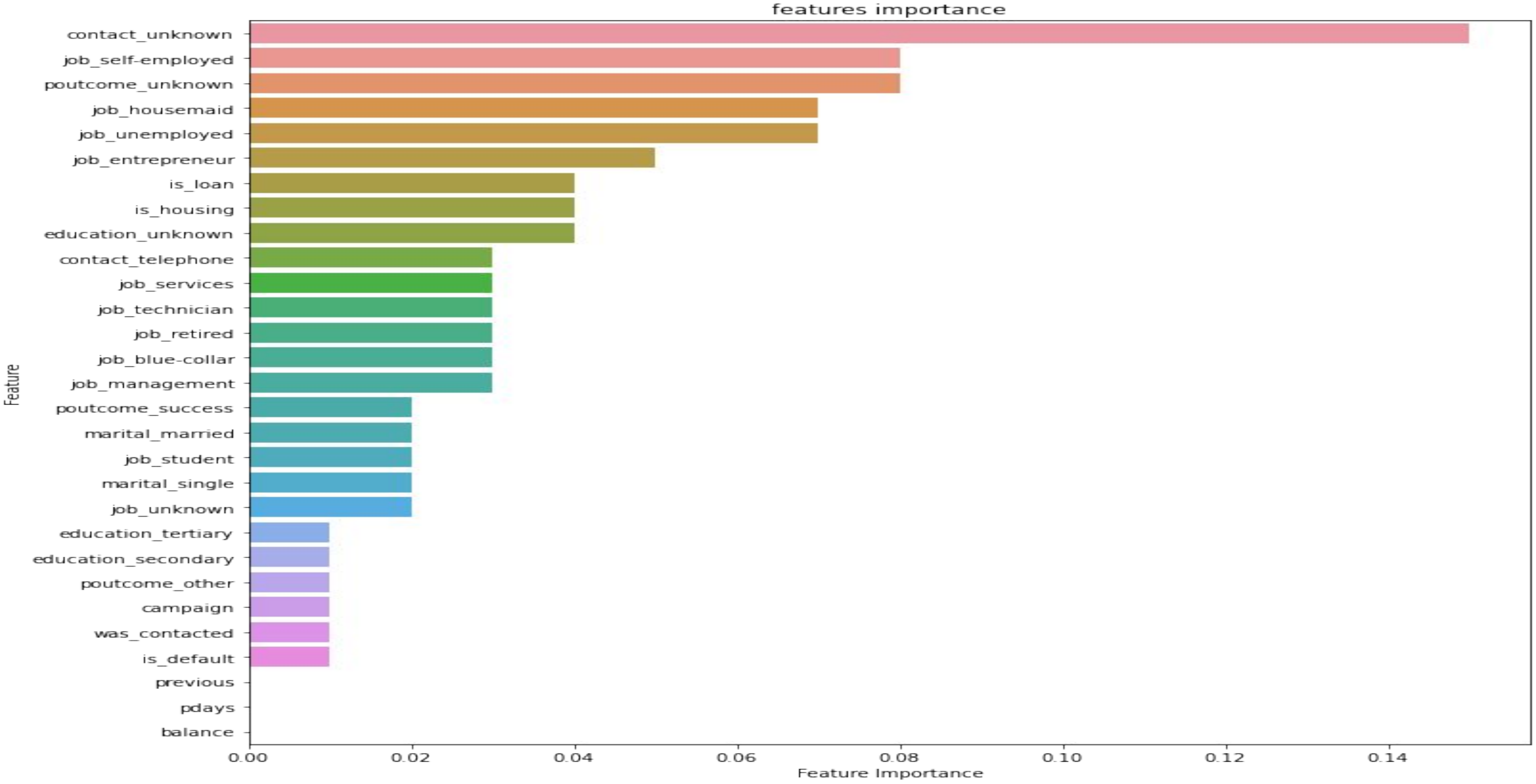
Confusion matrix of Train Data



Confusion matrix of Test Data



XGBoost Feature Importance :-



Decision Tree:-

Best Parameters

max_depth : 10

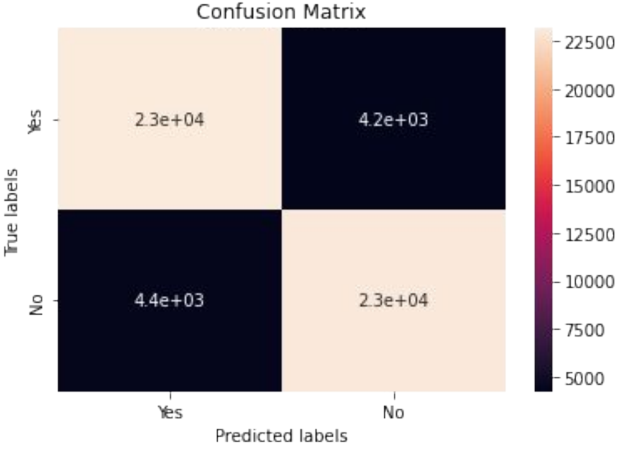
min_samples_leaf : 10

min_samples_split : 20

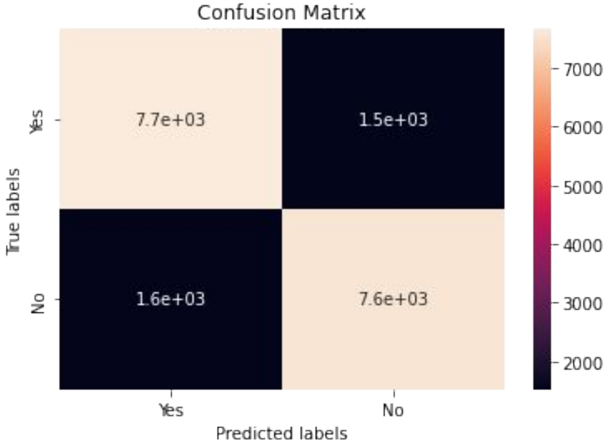
ROC-AUC Score :-

Train Data	0.92
Test Data	0.90

Confusion matrix of Train Data



Confusion matrix of Test Data



Random Forest:-

Best Parameters

max_depth : 10

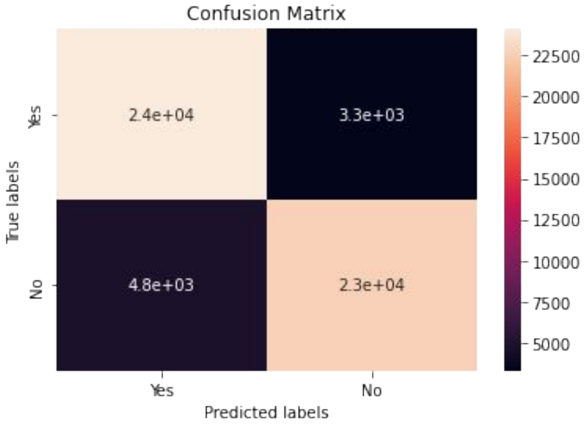
min_samples_leaf : 10

min_samples_split : 20

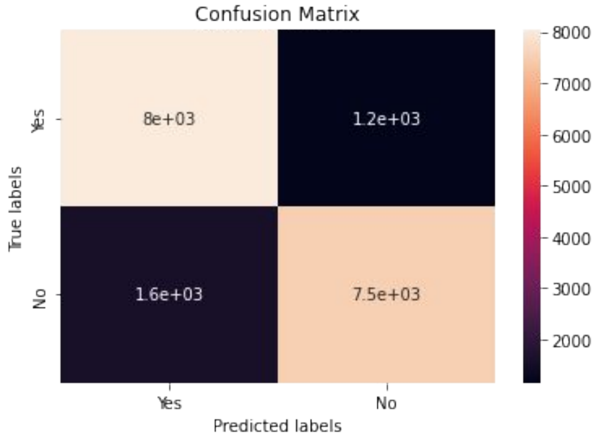
ROC-AUC Score :-

Train Data	0.93
Test Data	0.92

Confusion matrix of Train Data



Confusion matrix of Test Data



Hyperparameter Tuning Evaluation

Model	Test AUC	Test Accuracy	F1-Score	Precision
Logistic Regression	0.93	0.86	0.87	0.89
KNN	0.94	0.87	0.89	0.91
XGBoost	0.96	0.91	0.96	0.97
Decision Tree	0.90	0.83	0.84	0.85
Random Forest	0.82	0.85	0.86	0.88

Conclusion

- For age, most of the customers are in the age range of 30-40.
- For balance, above 1000$ is like to subscribe a term deposit.
- The model can help to classify the customers on the basis on which they deposit or not
- The model helps to target the right customer rather than wasting time on wrong customer
- Comparing to all algorithms **XGboost algorithm** has best accuracy score and ROC-AUC score . So it is concluded as optimal model.

THANK YOU