Capstone Project

Bank Marketing Effectiveness Prediction

Ritik Gupta

Contents

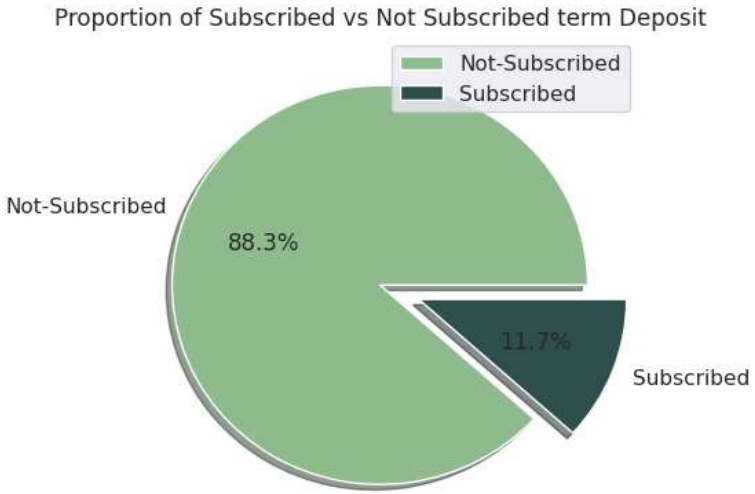
- ✓ Overview
- ✓ Data Description
- ✓ EDA (Insights)
- ✓ Feature Engineering
- ✓ Correlation & Selection
- ✓ Models Used
- ✓ Model Selection & Explainability
- ✓ Challenges
- ✓ Conclusion

Overview

- This project will enable the bank to develop a more granular understanding of its customer base, predict customer's response to its telemarketing campaign and establish a target customer profile for future marketing plans.
- The data is related with marketing campaigns (phone calls) of a Portuguese banking institution. Often, more than one contact to the same client was required, in order to access if the bank term deposit would be subscribed or not.
- Analyzing customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviors and identify which type of customers is more likely to make term deposits.

- **Main Objectives:-** predict customer's responses to future marketing campaigns & increase the effectiveness of the bank's telemarketing campaign
- The classification goal is to predict if the client will subscribe a term deposit '1' (yes) or '0' (no) of (variable y).

Data Description



> Dataset shape –

45211 Records and 17 Attributes

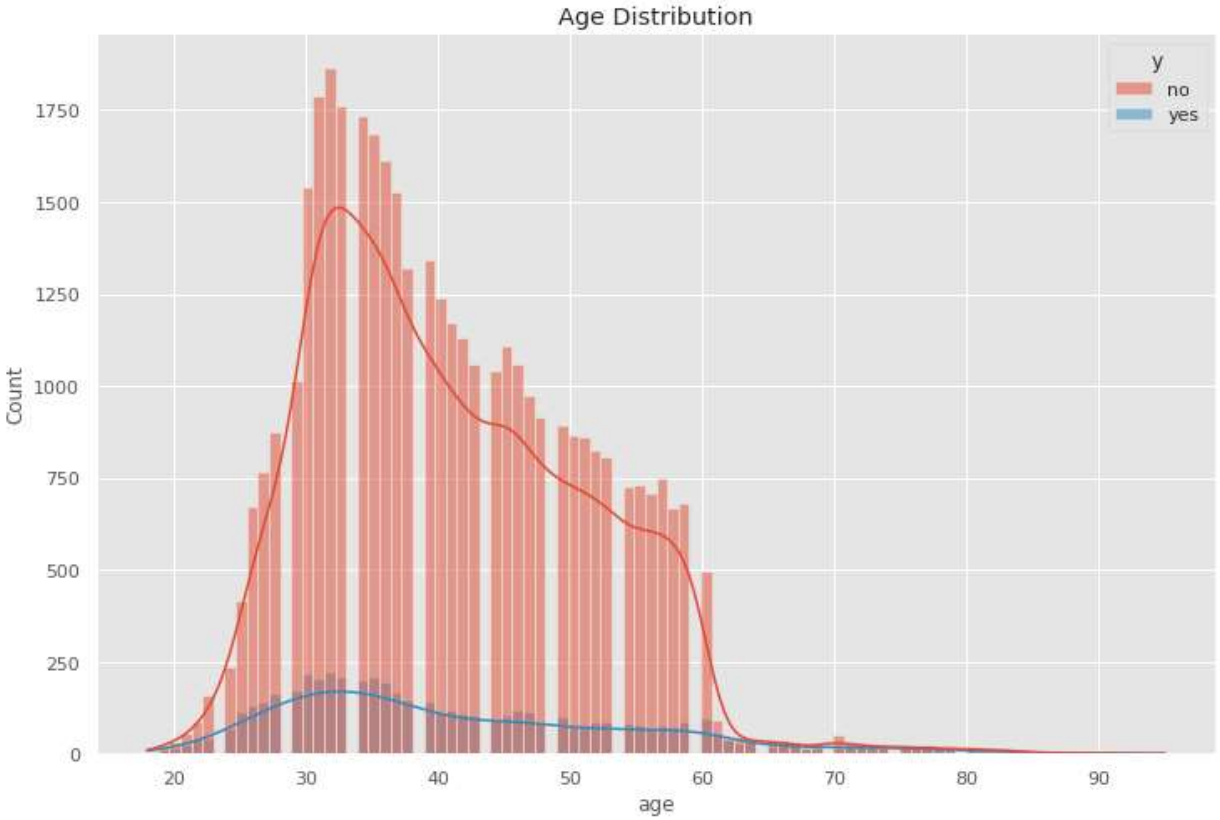
> Given Features-

- ❑ Age
- ❑ Job
- ❑ Marital
- ❑ Education
- ❑ Default
- ❑ Balance
- ❑ Housing
- ❑ Loan
- ❑ Contact
- ❑ Day
- ❑ Month
- ❑ Duration
- ❑ Campaign
- ❑ Pdays
- ❑ Previous
- ❑ Poutcome

Target Feature:

- ❑ outcome (Y/N)

Exploratory Data Analysis

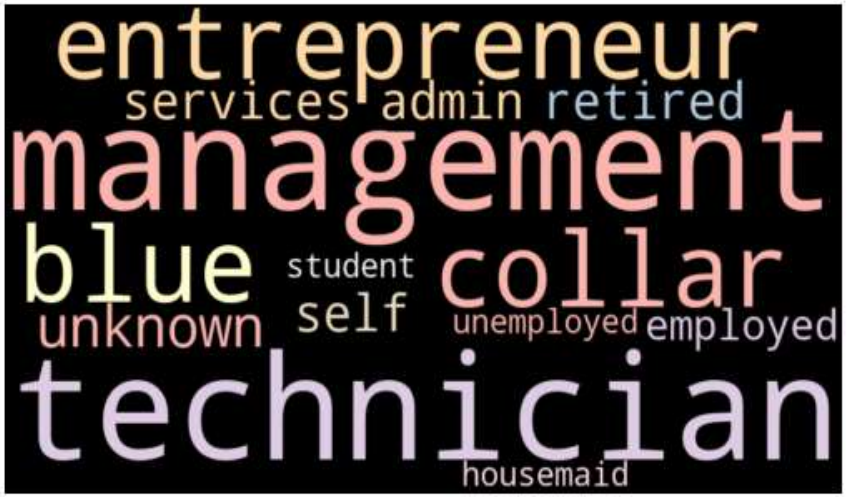


Bivariate analysis of Age with target variable

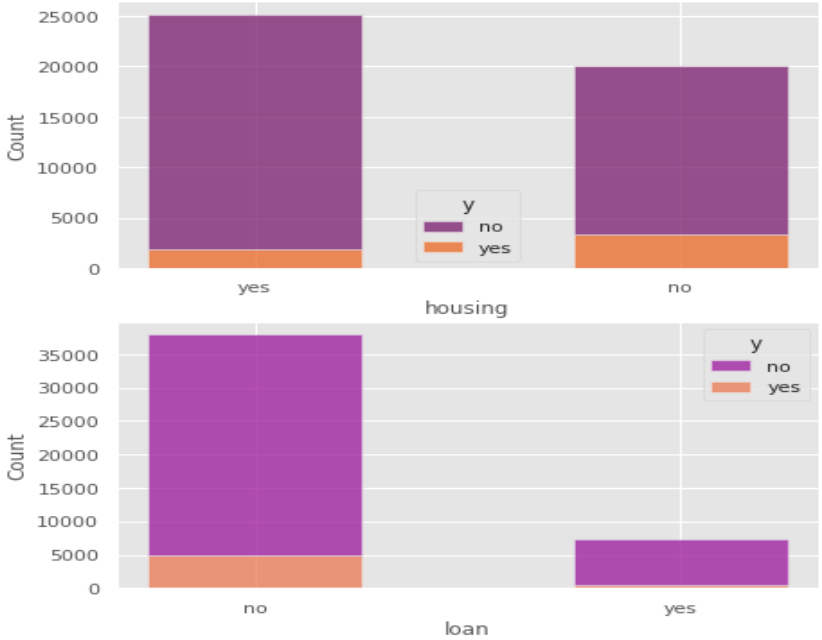
- *Marketing team is focusing from age 25 to 55.*
- *Old age clients along with young one's are also likely to subscribe.*
- *Given the subscription rate of clients more than age 60, the contact rate is very low.*
- *Contact rate should be increased between age 20-25 and above 60*

Exploratory Data Analysis

Job's contacted

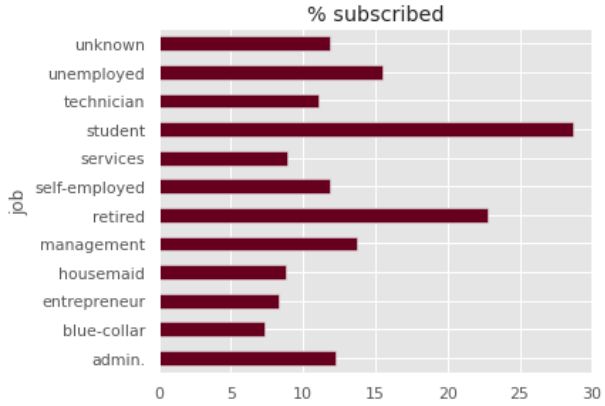


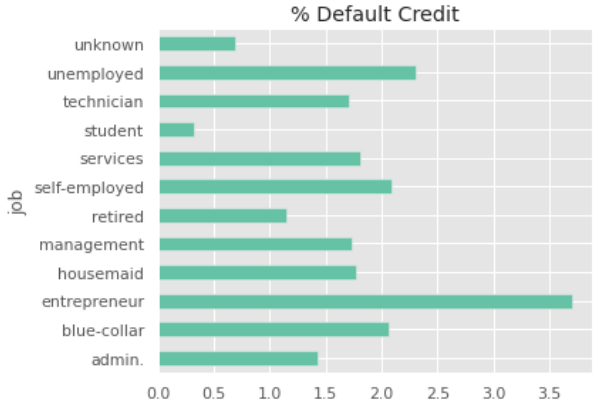
Housing & Personal Loan w.r.t Subscriptions

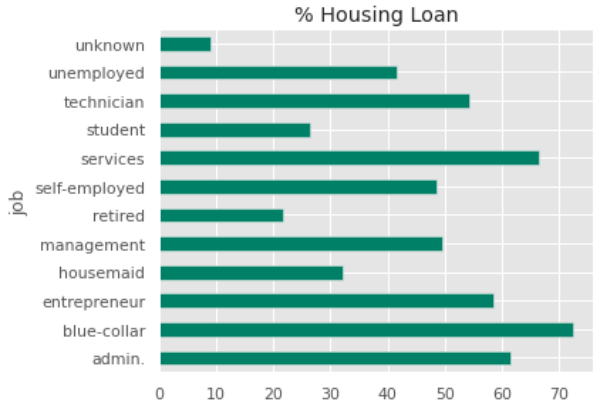


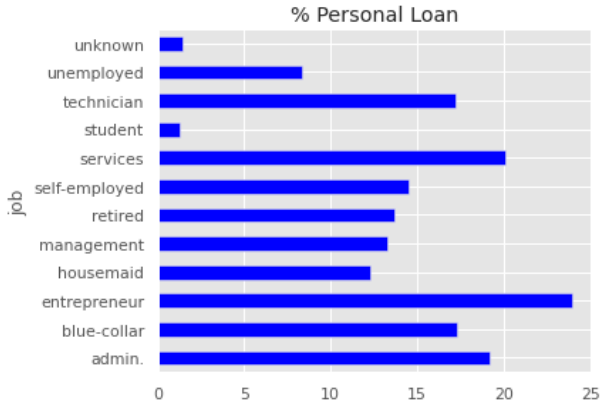
- *'Management', 'Technician' are the most contacted clients. and 'Student', 'Retired' are the least contacted.*
- *When a client has no housing loan then chances of subscription rises.*
- *when a client has a personal loan, likelihood of subscription is almost 0*

Exploratory Data Analysis





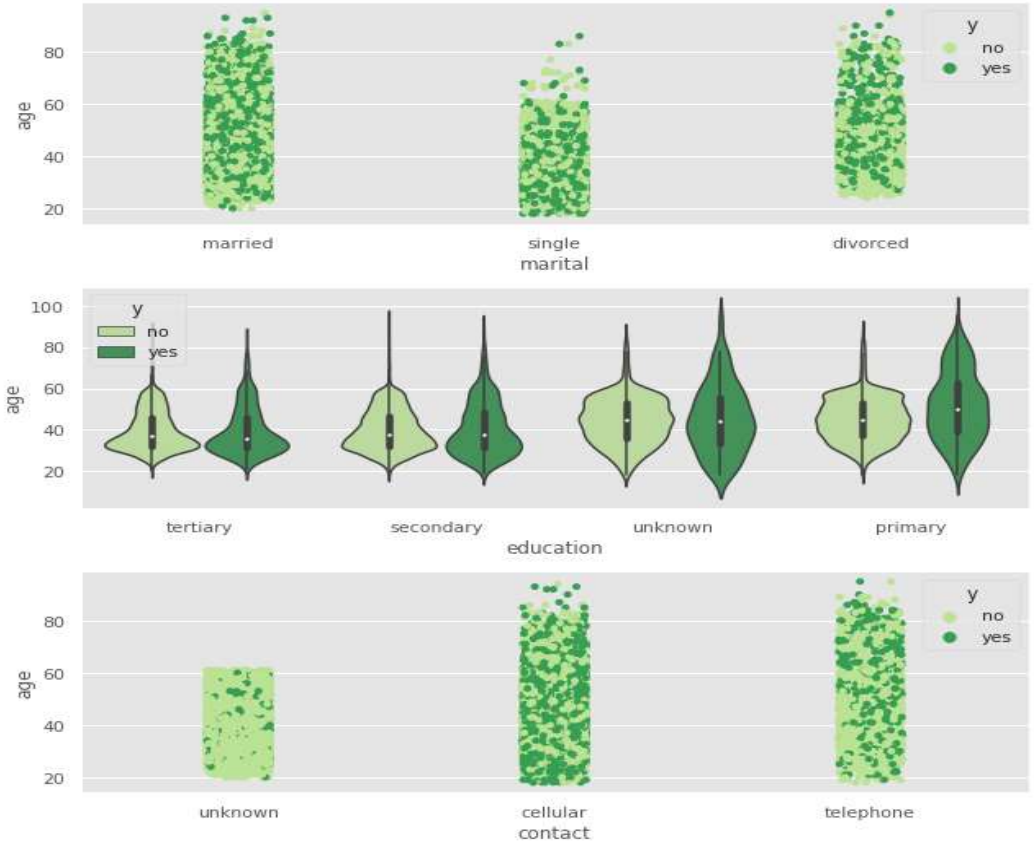




- *'Student' and 'Retired' profession has 2 highest in percentage of subscription and Blue-collar are one of the lowest.*
- *Entrepreneurs has highest Default credits followed by unemployed and self-employed.*
- *Blue-collar profession has highest housing loan rate followed by Services, admin., technician and Entrepreneurs.*
- *Entrepreneurs has high personal loan rate followed by Services, technician, admin and Blue-collar.*

Job with respect to Subscription, Default credit, House loan and Personal loan

Exploratory Data Analysis

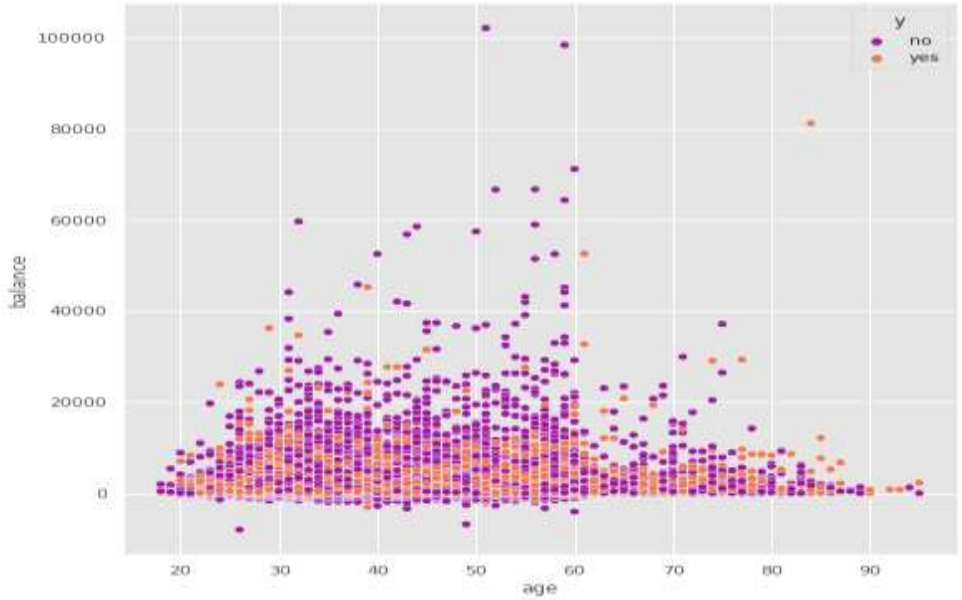


Multivariate analysis of Marital status, education, contact to age w.r.t subscription

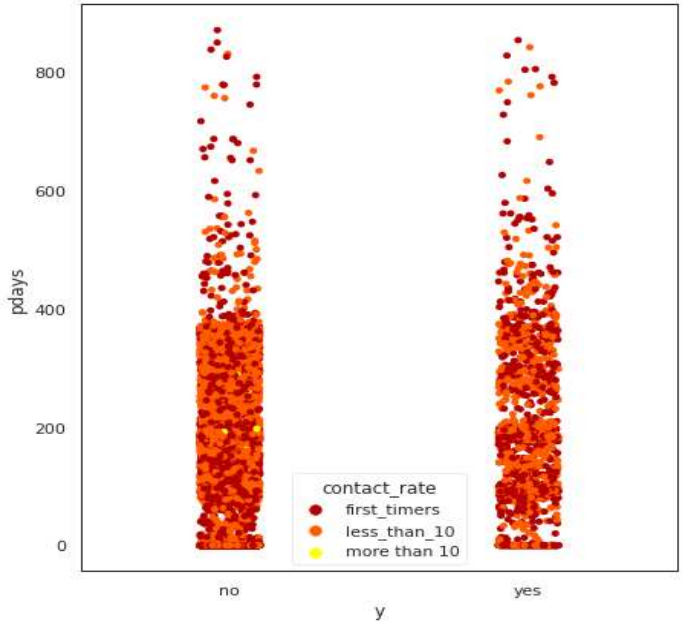
- *'Married' and 'singles' are more likely for subscription than 'divorced'.*
- *Unknown and primary educated clients have random pattern of subscription, but secondary and tertiary educated are more likely to subscribe.*
- *Cellular contact mode is the best mode, and have the highest subscription rate.*

Exploratory Data Analysis

Relation between Balance and age of client w.r.t. to subscriptions



pdays on the subscription rate w.r.t contact rate

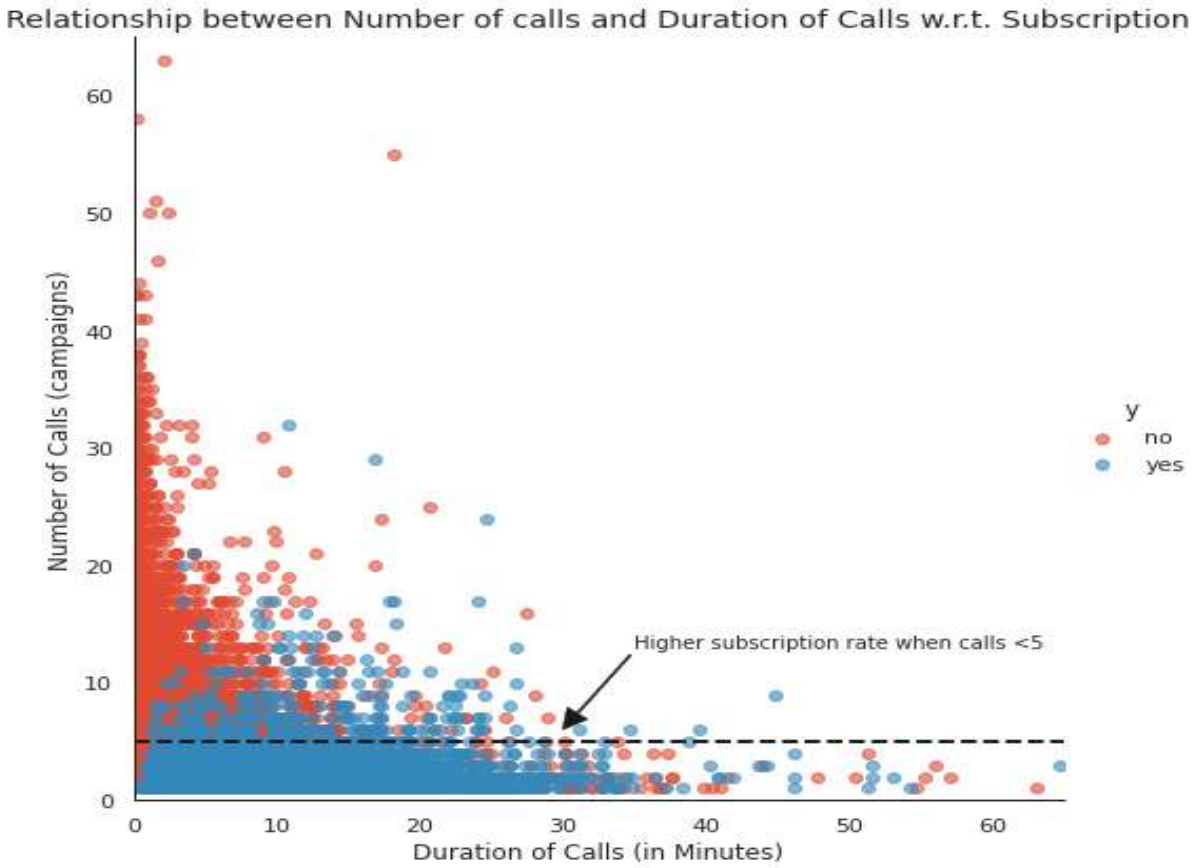


- *clients whose balance is under 15000 euros are more likely to subscribe, Also when age is above 60 with low balance but they are still likely to subscribe.*
- *We see quite high subscription rate when pdays are from 10 to 200 days.*
- *Most of the contacts are either first time or called less than 10 times.*

Exploratory Data Analysis

- *Bank contacted most of clients between May(5) and July(7).*
- *contacted rate is lower in march, September, October, December.*
- *subscription rates in, march, September, October, and December are over 40%.*
- *Data reflects some inappropriate timings for campaign, and to improve this, we should focus on months which shows us higher subscription rate.*

Exploratory Data Analysis

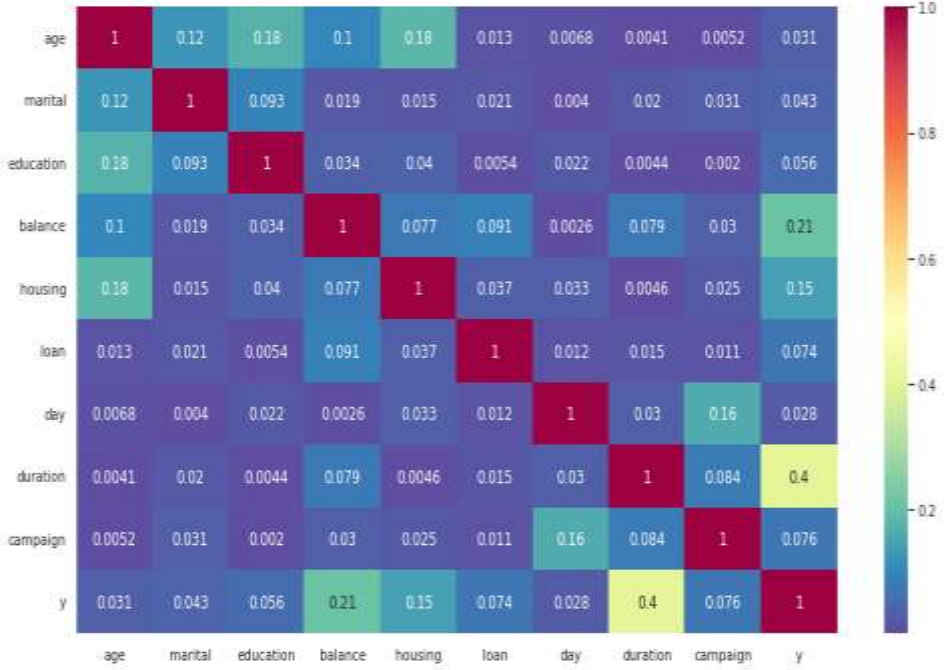


- *Increasing more number of calls made negative impact to clients.*
- *Subscription rate is high when calls are less than 5 times.*
- *Duration also made an high impact in subscriptions, so team has to make good conversation with client, that will increase the likelihood of subscribing.*

Feature Engineering

	Details
week_number	Derived from ‘day’ feature.
Previous	Converted numbers of calls made into 4 categories “first time caller”, “less_than_5”, and “more_than_5”
month	Created quarter category from month number
marital	Target encoded divorced as 0, single as 1, and married as 2
education	Encoded primary, secondary and tertiary with respective numbers.
Others	Encoded “housing”, “loan” and target variables with 1 and 0, for yes and no.

Correlation & Selection



Least 5 important feature from ANNOVA F-value for Feature selection

18	job_technician	4.267812
31	month_Q_3	1.505021
29	week_number_week_4	0.410626
15	job_self-employed	0.170817
13	job_other	0.035567

➢ *We have to drop:-*

- ❖ *Week_number_week_4*
- ❖ *Job_self_employed*
- ❖ *Job_other*

❖ *"Duration" attribute highly affects the output target, yet, the duration is not known before a call is performed. Also, after the end of the call "y" is obviously known & Our goal is to build realistic predictive model, Thus, we will not use this for our models.*

ML Models Used

- Random Forest
- Naive Bayes Classifier

SMOTE-tomek Oversampling technique

- K Neighbor Classifier
- XG Boost
- Support Vector Classifier
- Neural Network Classifier (multilevel perceptron)

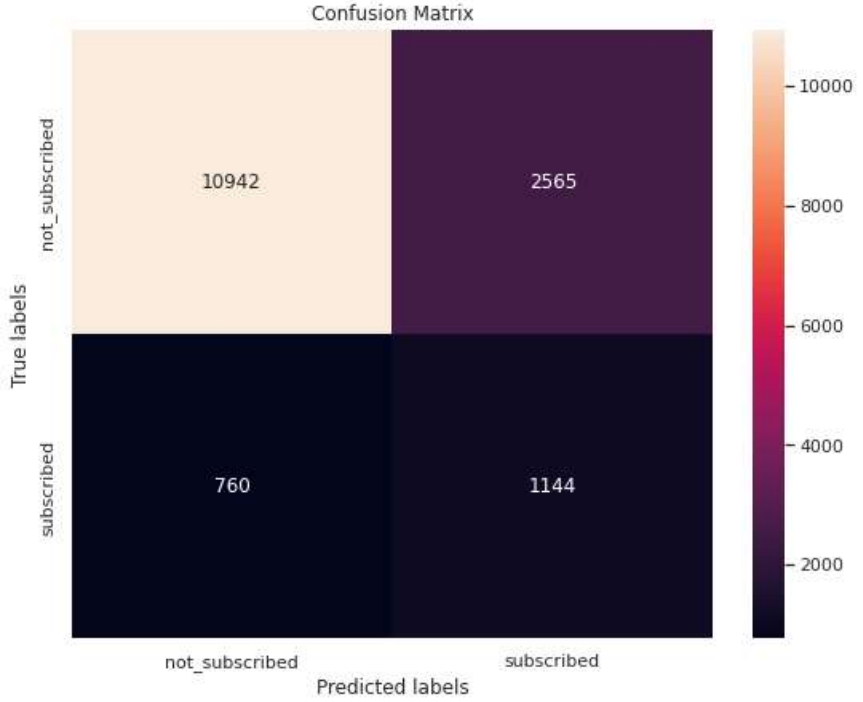
Nearmiss Undersampling technique

Models used

Oversampling

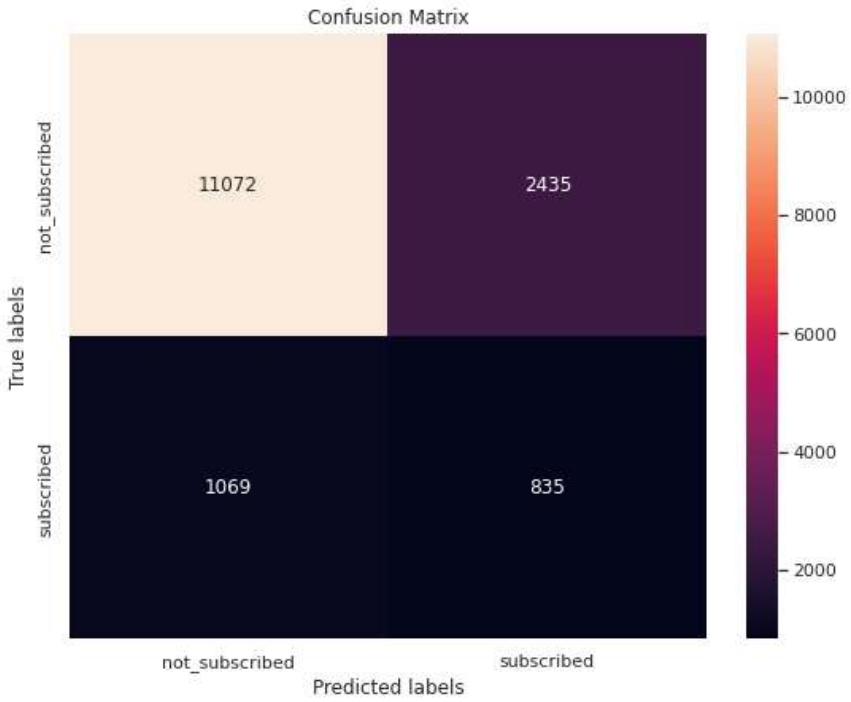
Random Forest

Roc_Auc = 0.70



Naive Bayes Classifier-

Roc_Auc = 0.629

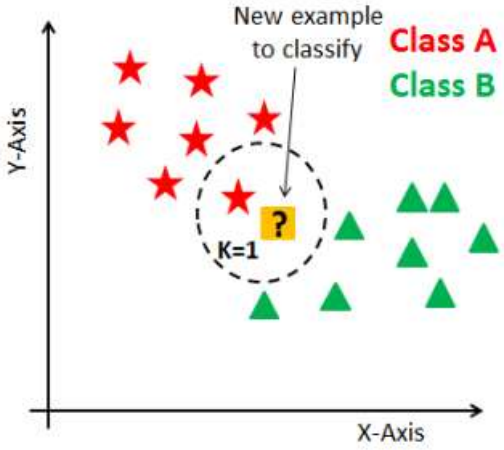


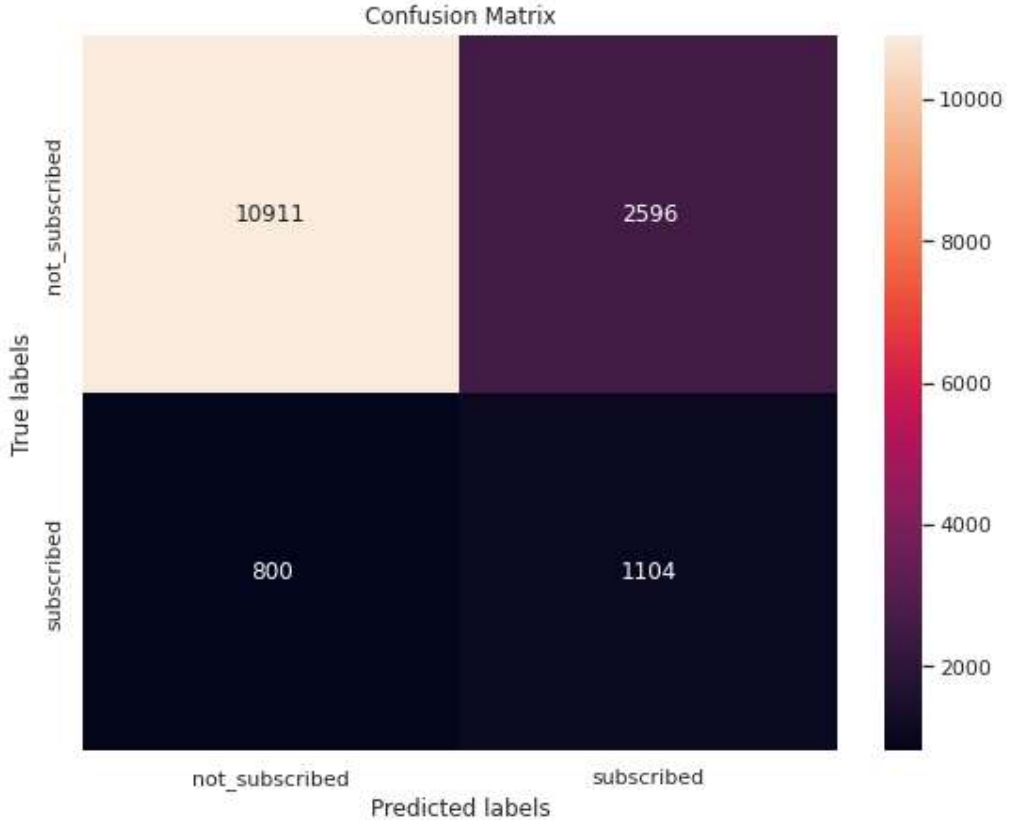
Models used

Undersampling

K Nearest Neighbor

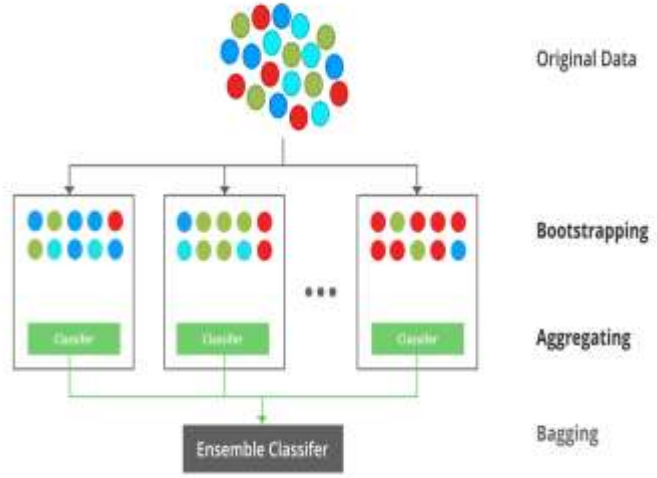
➢Roc_Auc = 0.693

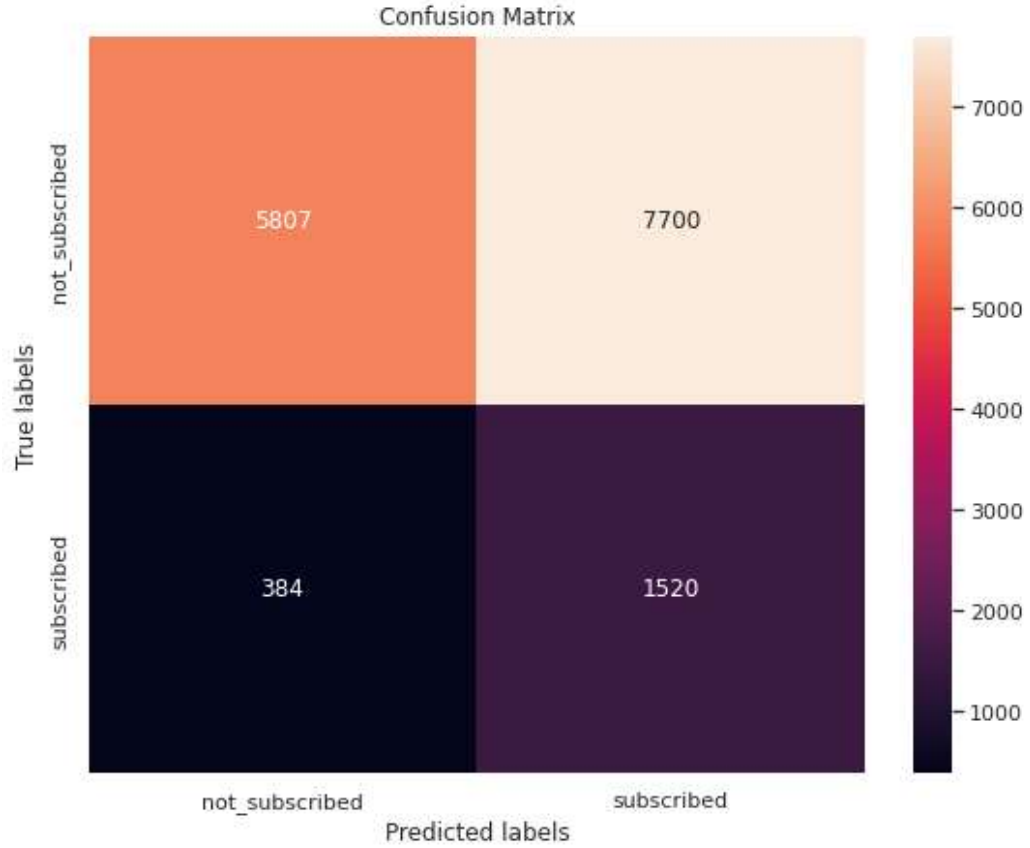




XG Boost classifier

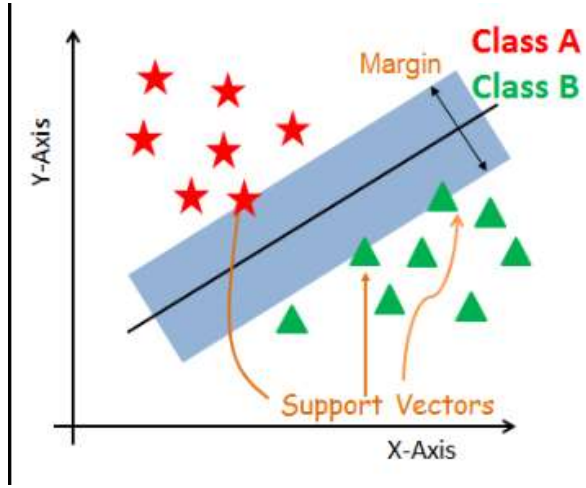
- Roc_Auc = 0.614

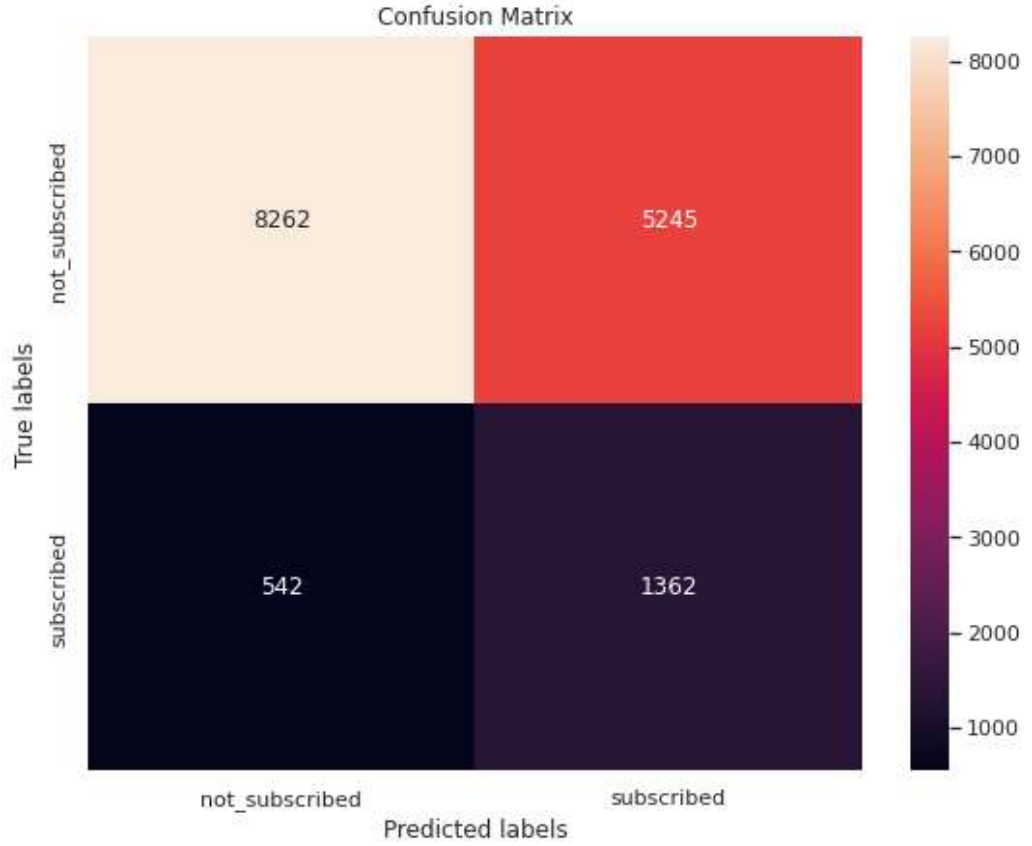




Support Vector Classifier

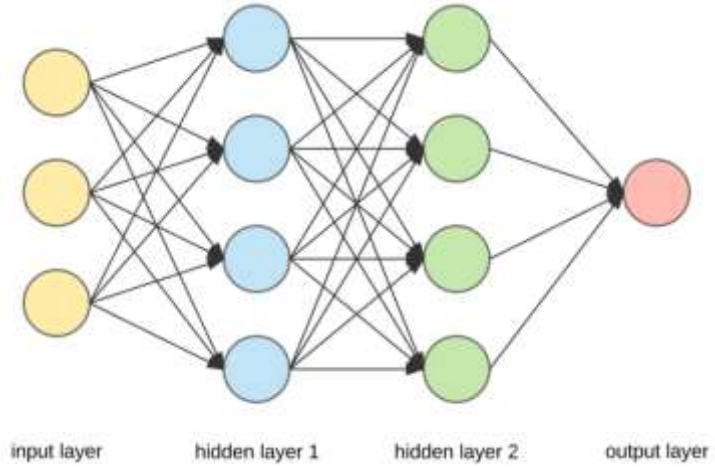
➢ Roc_Auc = 0.663

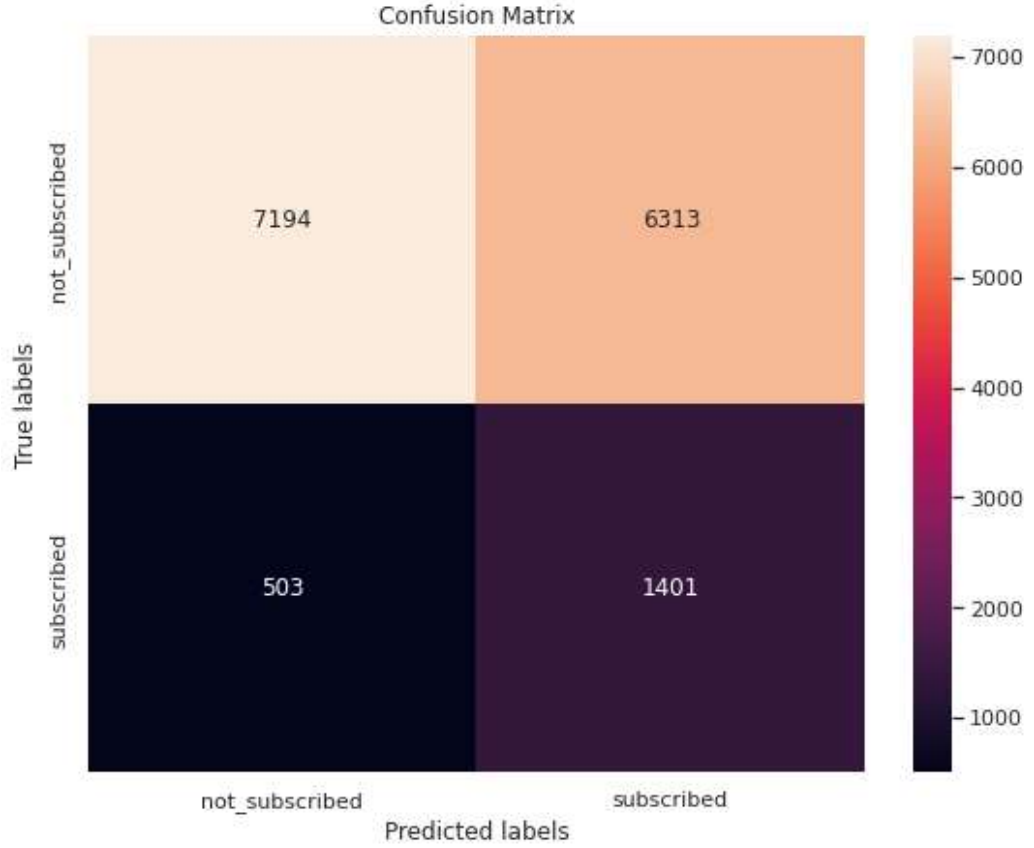




Neural Network Classifier

- Roc_Auc = 0.634





Model Comparision

	auc-roc	F1_score	Recall	Matt_Corr_Coef
Random Forest	0.708463	0.408851	0.611345	0.318589
Naive Bayes	0.625688	0.316105	0.443277	0.198735

When trained on Oversampled data

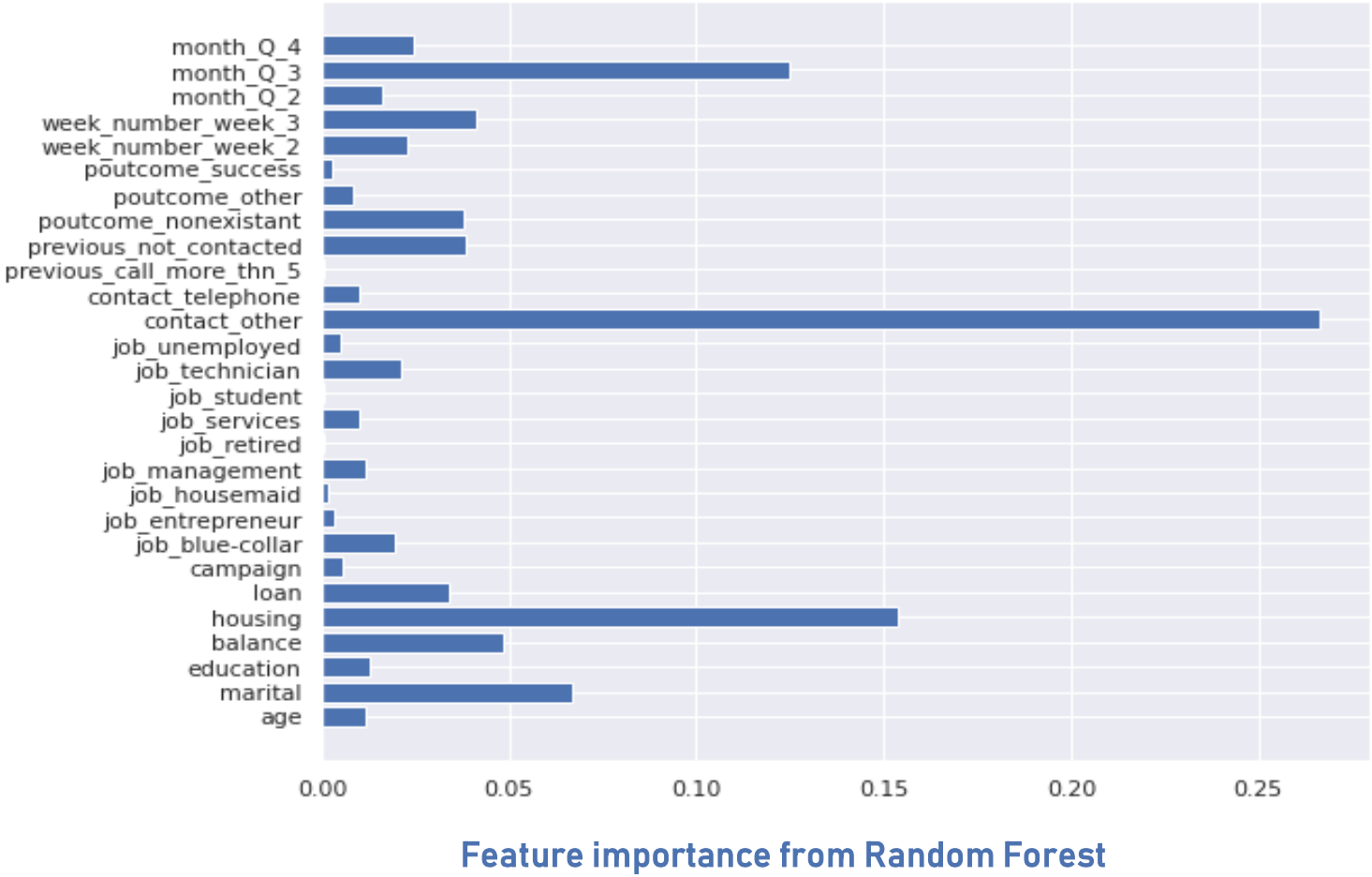
When trained on Undersampled data

	auc-roc	F1_score	Recall	Matt_Corr_Coef
K Neighbor Classifier	0.693818	0.394004	0.579832	0.298633
XG Boost	0.614122	0.273283	0.798319	0.153203
Support Vector Machine	0.663509	0.320056	0.715336	0.217442
Neural Network	0.634216	0.291329	0.735819	0.176663

Model Selection

- Due to imbalance nature of data, accuracy is not a good option to choose, and generally, Roc_Auc score is used in these cases.
- Since our data has not any correlated features, we have to decide from False positives or False negatives.
- As per business use-case for model, we cannot leave any True positive, so we need to reduce False Negatives i.e choosing highest Recall.
- Highest Recall is given by XG Boost algorithm from undersampled data.

Model Explainability and Importance



Challenges

- Imbalanced class data is the biggest challenge in this problem and techniques use to handle these situations has their own advantages and disadvantages.
- Using Nearmiss undersampling technique, our models may have missed some useful information, but still XG boost able provide 80% Recall.
- “Duration” Feature which is giving us highest correlation is biased and can’t be used in realistic predictive model, and by removing that from input, and using over and undersamplig techniques, model’s overall performance is reduced.
- Outliers create big trouble, and while removing outliers, we lost data including the minority class which is already low in this case.
- Hyperparameter tuning is computationally expensive process, and to fasten our search we may have missed some combinations.

Conclusion

- *Young and Old clients are more likely to subscribe this term deposit than any other age group.*
- *Clients whose balance is under 15000 euros are more likely to subscribe. Clients with more balance may tends to choose risky investment options.*
- *Increasing number of calls will not make customer to subscribe, but calls made <5 times has high subscription rate..*
- *March, September, October, December, has low contact rate, but has high subscription rate, this shows irregularity in contact timings.*
- *when a client has no personal loan then he is more likely to subscribe.*

ML Model:-

- *Random Forest gave us best overall performance with 'roc_auc' score of 70% .*
- *Business use-case suggest us to choose model which gives us highest Recall 80% which is XG Boost. Also, Tree based models have provided god results than other models.*
- *Neural Network can be more improved using more layers and complicacy.*
- *Thresholds can also be tuned (default is 0.5) using predictions that also help us to reduce False Positive or False Negatives*


Feedback