Capstone Project
Bank Marketing Prediction (Supervised ML - Classification)

Bank Marketing Effective Prediction

Problem Statement

Make predictions

01

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

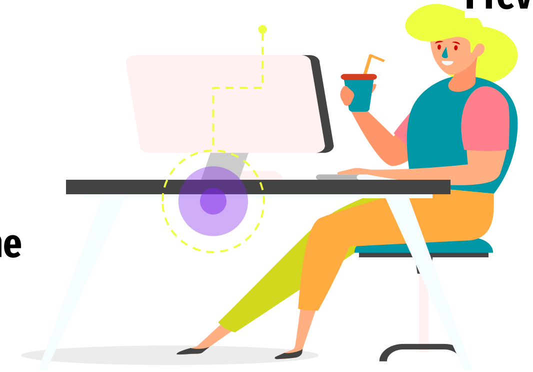
Model Development

02

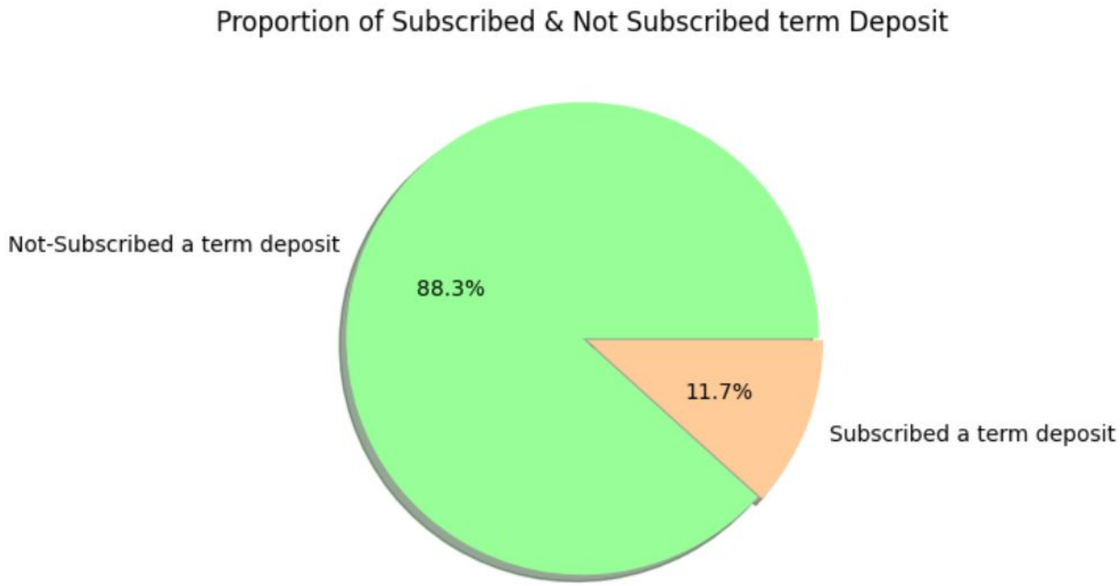
Develop a **Supervised Machine Learning Model** using **Classification.**

Features

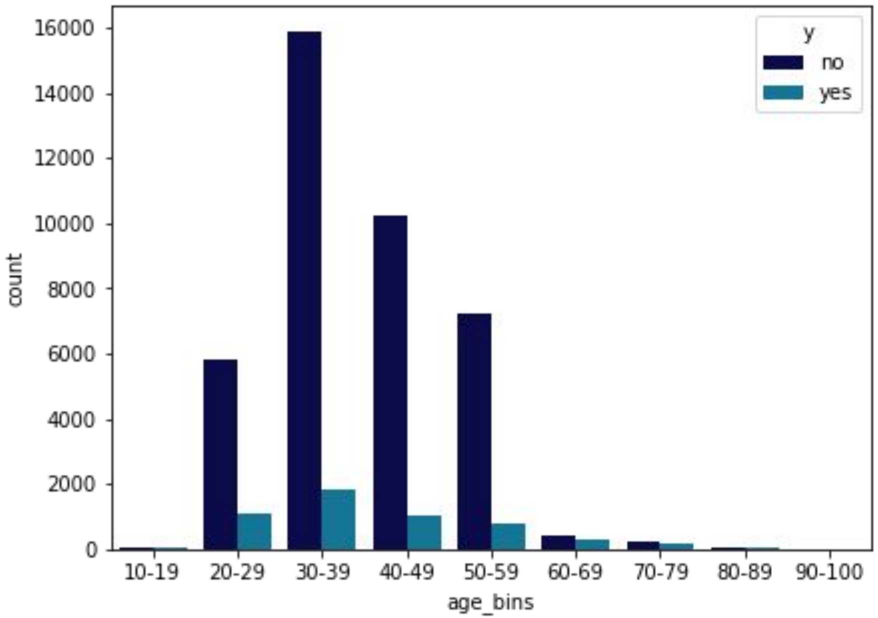
01 Age
02 Job -type of a job
03 Marital -Marital Status
04 Education
05 Default- Has credit in default?
06 Housing - Has housing Loan?
07 Loan - Has a personal loan
08 Contact - Communication type
09 Month- Last contact month of the year
10 Day_of_week -Last day of week
11 Duration -Last contact duration
12 RH5- humidity in bathroom

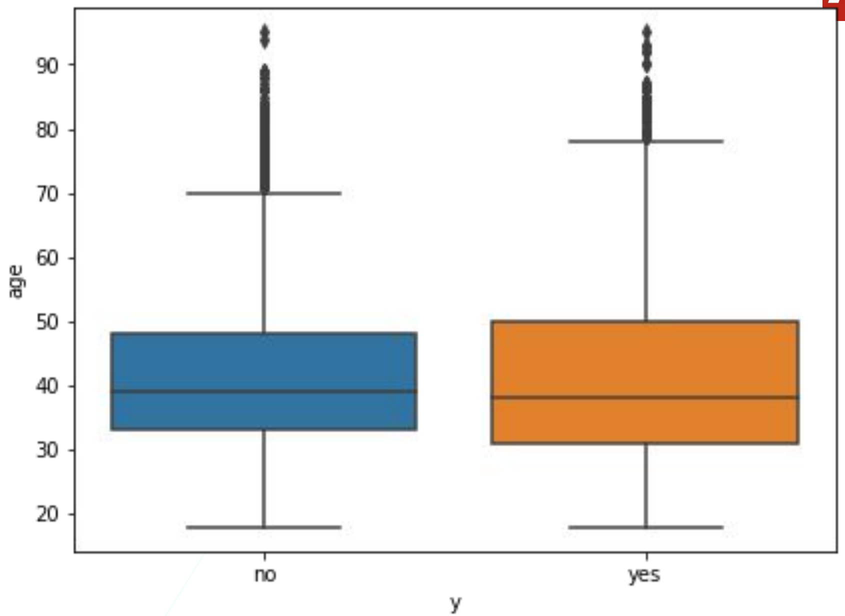


Campaign-No. Of contacts performed during this campaign 13
Pdays- No. of days 14
Previous-No. Of contacts performed before this campaign 15
P-outcome -Outcome of previous marketing campaign 16
Target variable(y) -Has client subscribed to term deposit 17



We can see from the above plot that the dataset is imbalanced, where the number of the subscribed class is close to 8 times the number of Not-subscribed class

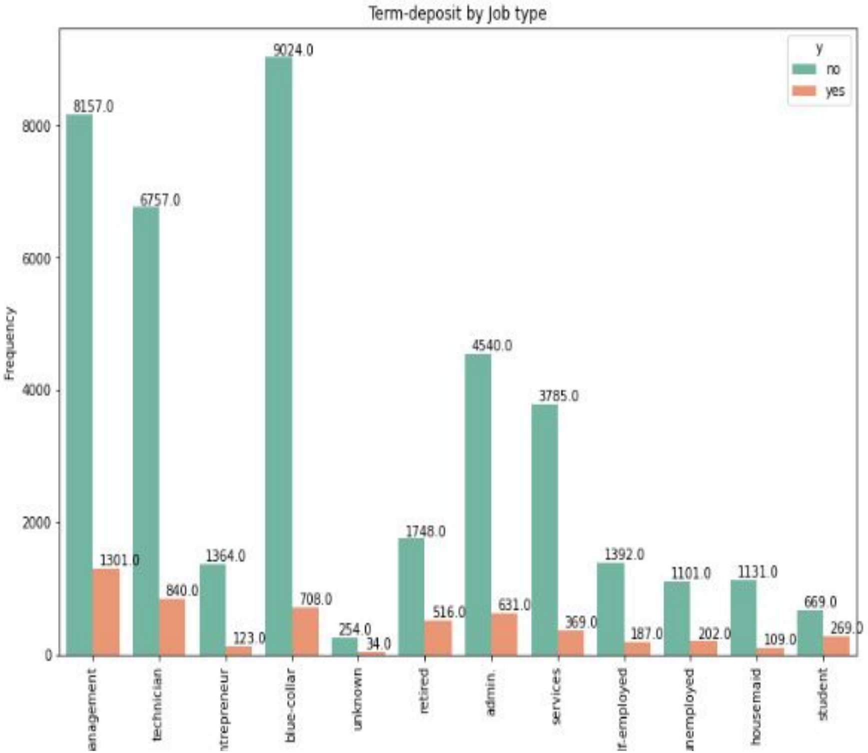




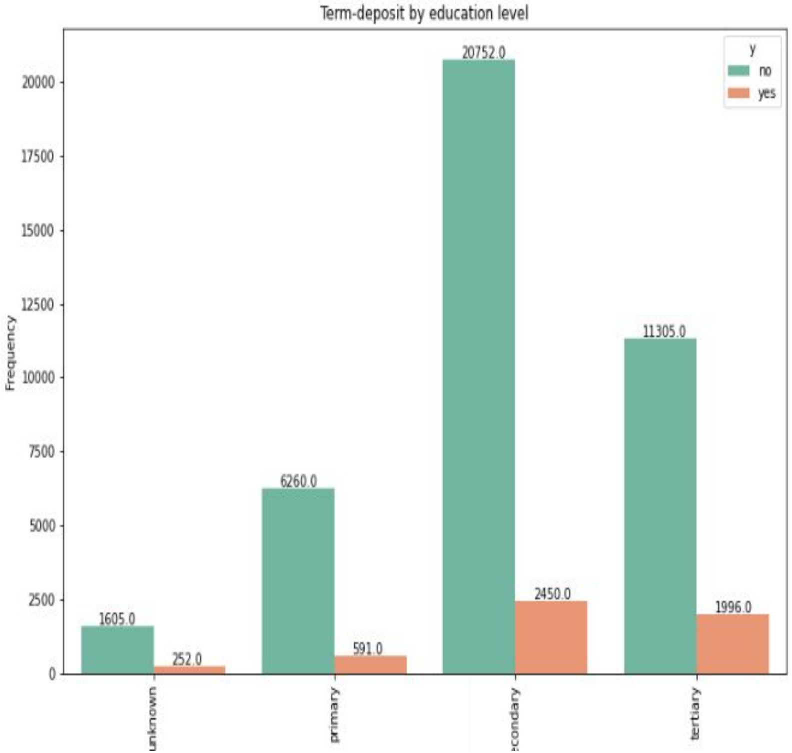
Majority of the customers are of the age group 30-39. Followed by 40-49 and 50-59

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

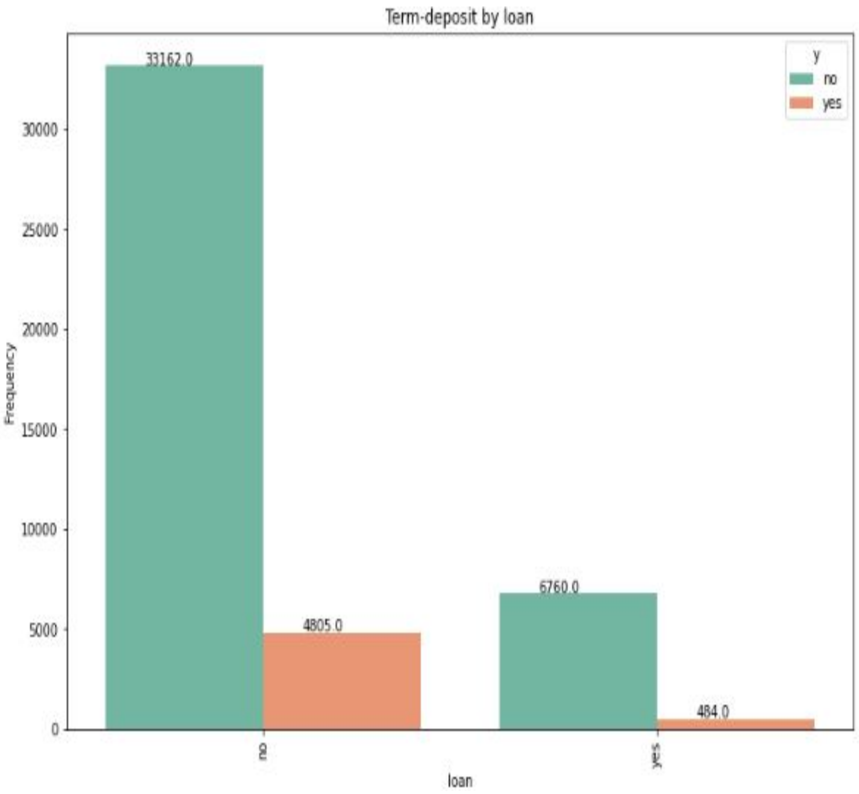


The majority of the customers who were contacted by the bank have blue-collar jobs but most of the term deposits have been taken by Management professionals

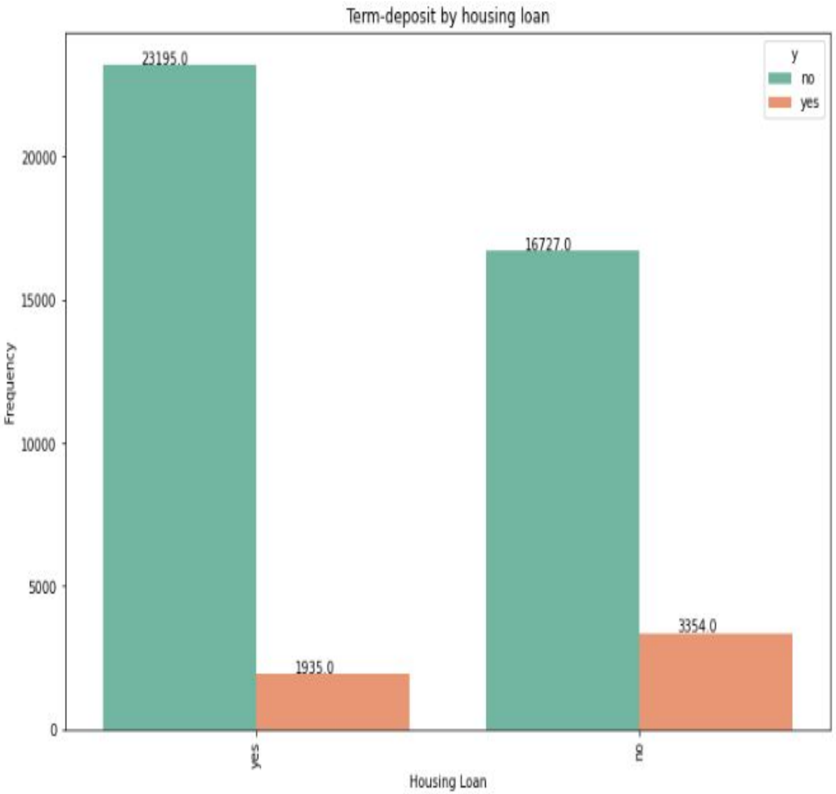


The majority of the customers have secondary education as the highest level of education

EDA

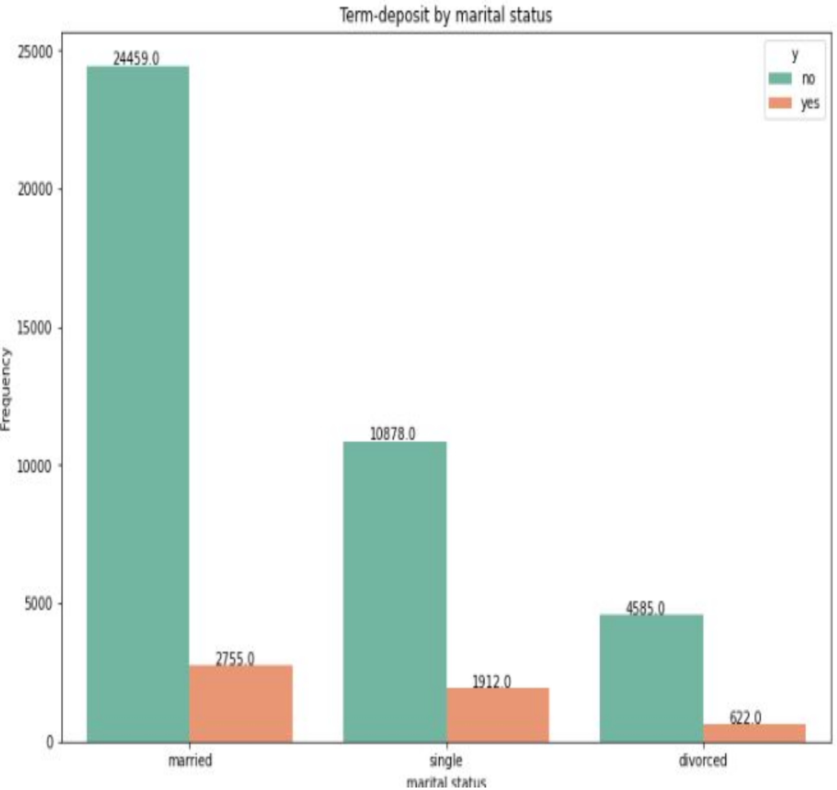


Very few people have taken a loan and these are more likely to take a term deposit.

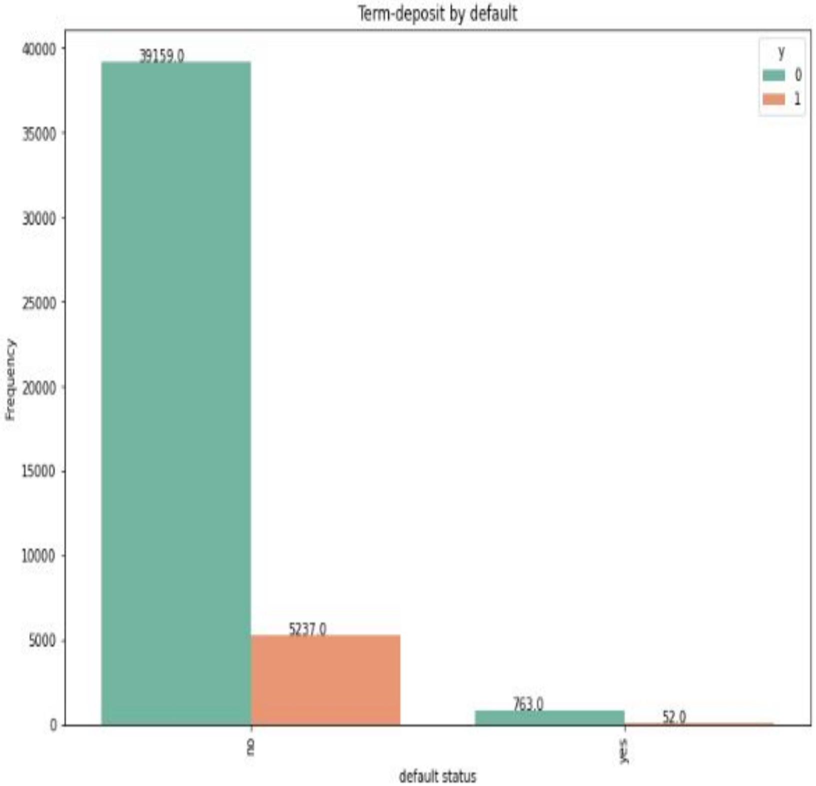


The majority of the customers have a housing loan. But those who do not have a housing loan are more likely to subscribe to a term deposit.

EDA

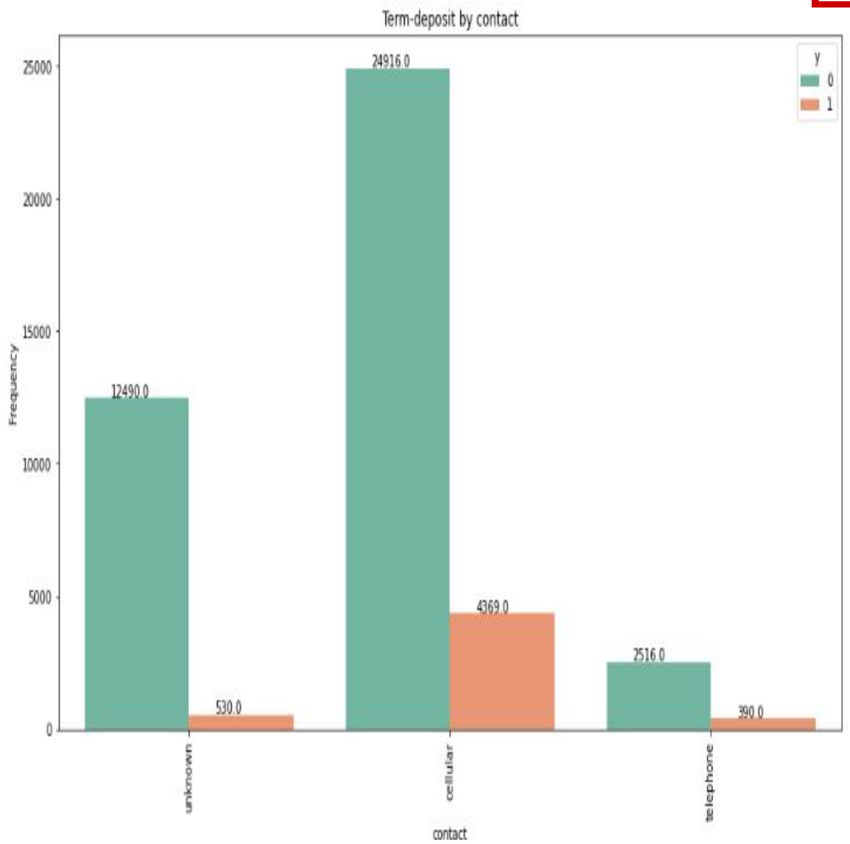


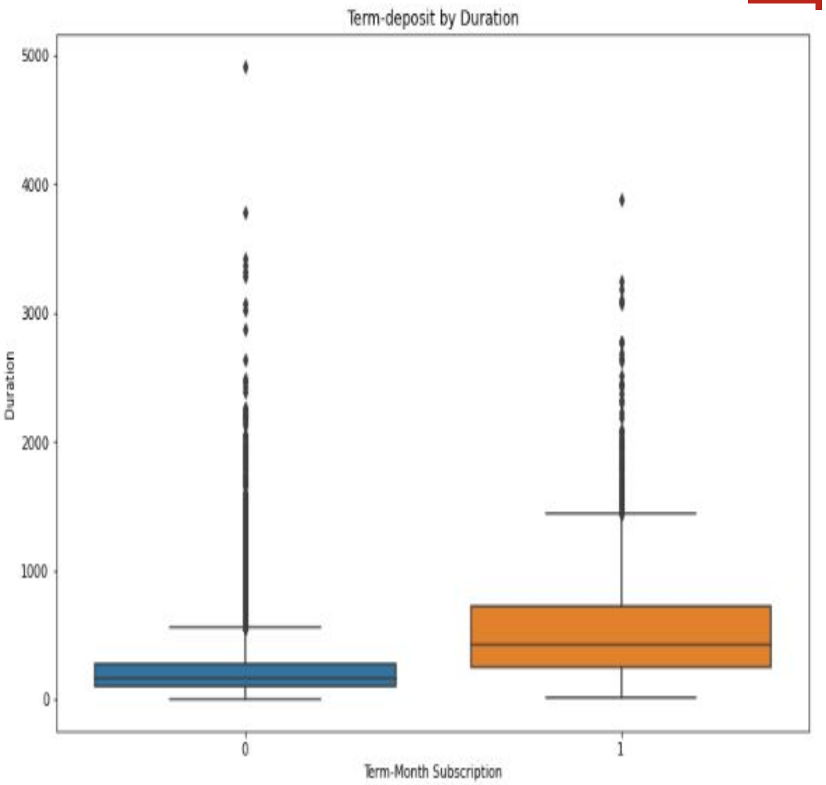
The Majority of the customers are married, followed by Single and Divorced.



People with default status as ‘no’ are the most who have been contacted by the bank for the deposits.

EDA

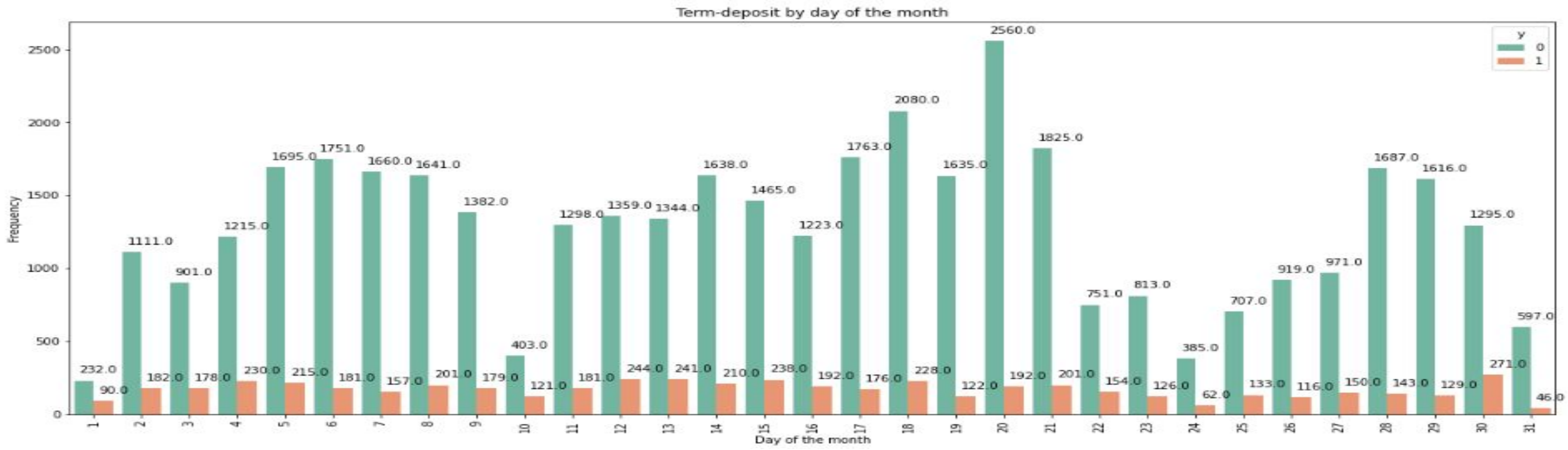


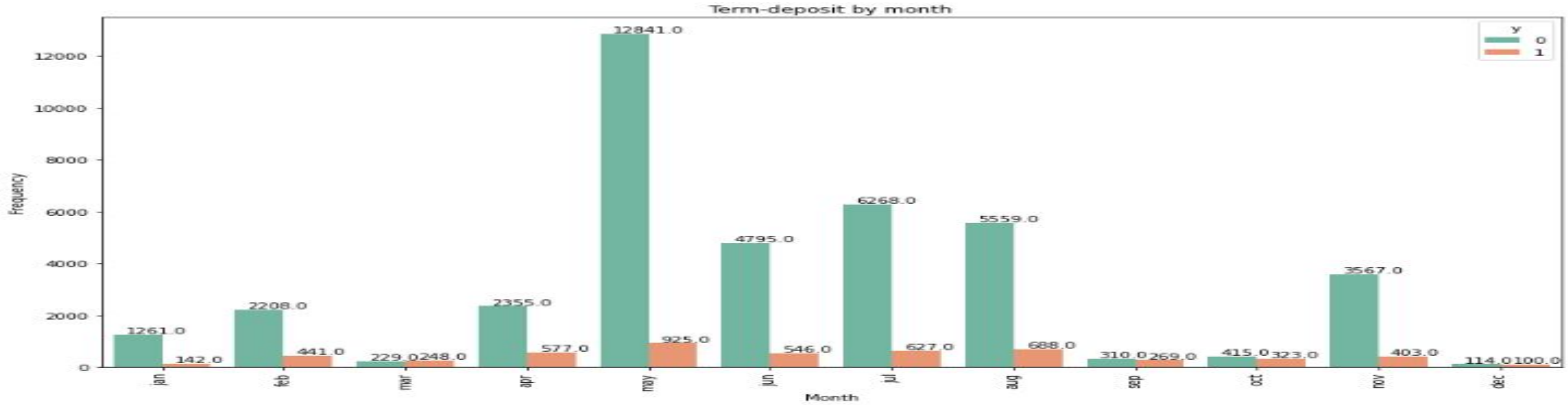


Customers are contacted more by cellular rather than telephone.

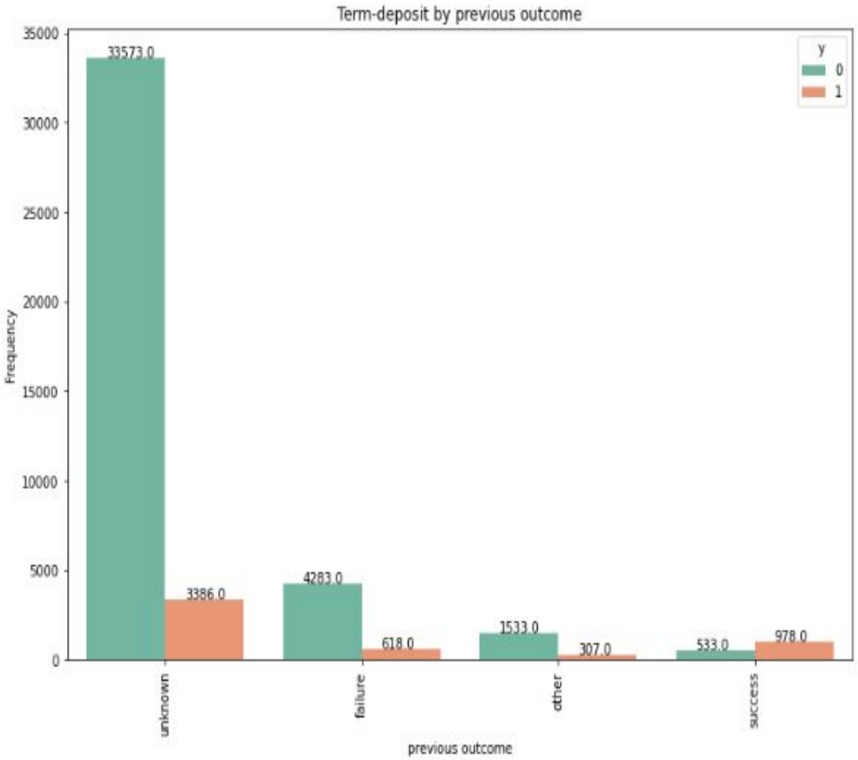
Customers who subscribed to term deposits have a relatively higher call duration.

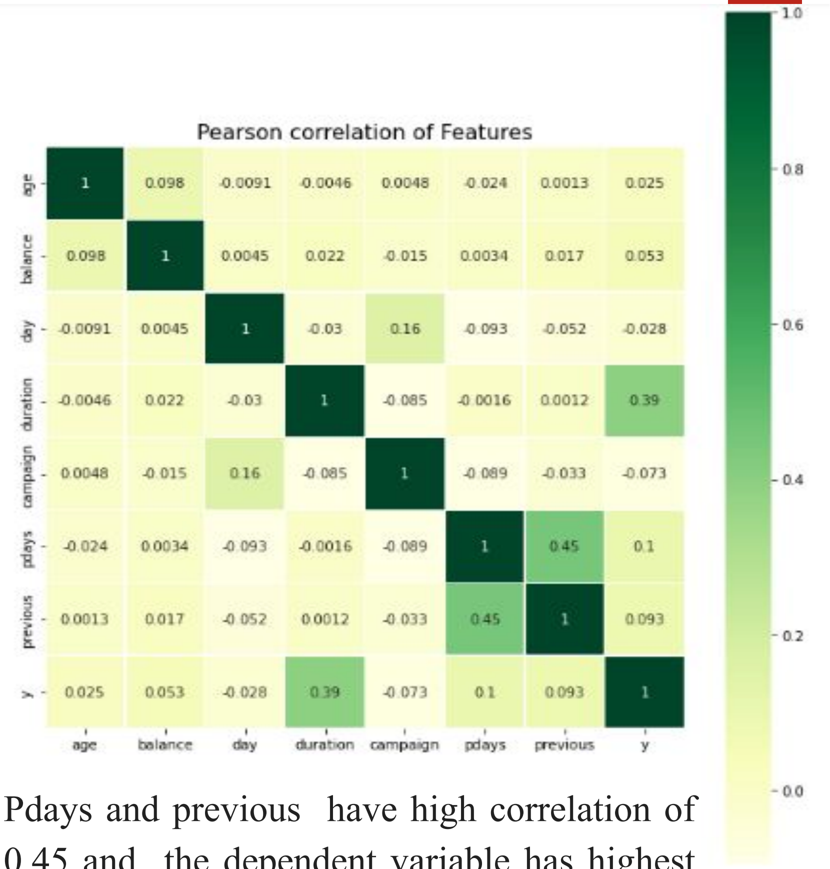
EDA


Term-deposit by day of the month
Frequency
Day of the month
y
0
1
0
500
1000
1500
2000
2500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
232.0 90.0 1111.0 182.0 901.0 178.0 1215.0 230.0 1695.0 215.0 1751.0 181.0 1660.0 157.0 1641.0 201.0 1382.0 179.0 403.0 121.0 1298.0 181.0 1359.0 244.0 1344.0 241.0 1638.0 210.0 1465.0 238.0 1223.0 192.0 1763.0 176.0 2080.0 228.0 1635.0 122.0 2560.0 192.0 1825.0 201.0 751.0 154.0 813.0 126.0 385.0 62.0 707.0 133.0 919.0 116.0 971.0 150.0 1687.0 143.0 1616.0 129.0 1295.0 271.0 597.0 46.0


Term-deposit by month
Frequency
Month
y
0
1
0
2000
4000
6000
8000
10000
12000
jan feb mar apr may jun jul aug sep oct nov dec
1261.0 142.0 2208.0 441.0 229.0 248.0 2355.0 577.0 12841.0 925.0 4795.0 546.0 6268.0 627.0 5559.0 688.0 310.0 269.0 415.0 323.0 3567.0 403.0 114.0 100.0

EDA

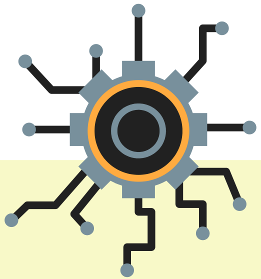




The outcome of the previous campaign is unknown for most of the customers but 64% of customers who had a successful outcome in the previous campaign did subscribe for a term deposit.

Pdays and previous have high correlation of 0.45 and the dependent variable has highest correlation with duration .

Feature Engineering



Feature Engineering

01

Dropping unknowns

Education, Job,contact, outcome and balance.

02

Label Encoding

Default, Housing, Loan and y.

03

Getting Dummies

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

Model Selection

01
Logistic Regression

02
Decision Trees

03
Random Forest Classifier

04
Gradient Boosting Classifier

05
XGBoost Classifier

06
Neural Networks Classifier

07
Naive Bayes Classifier

08
Linear SVM

Cross Validation

Definition:

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

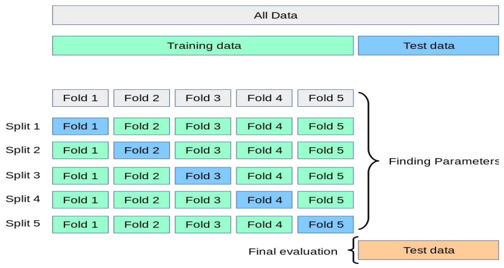


Fig A: K-Fold Cross-Validation

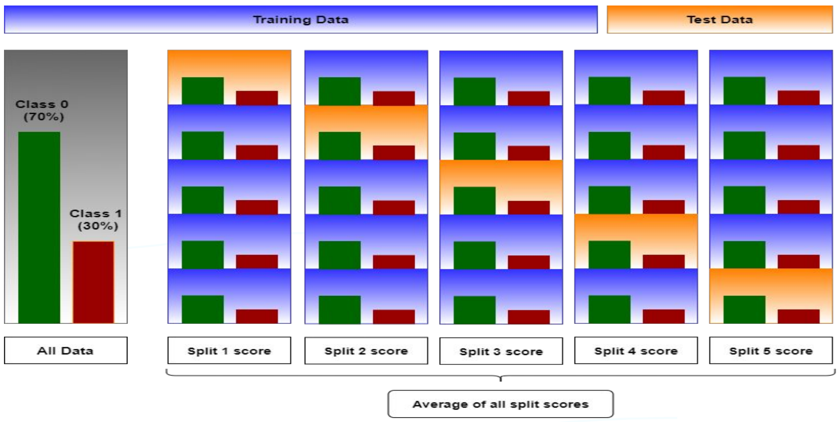
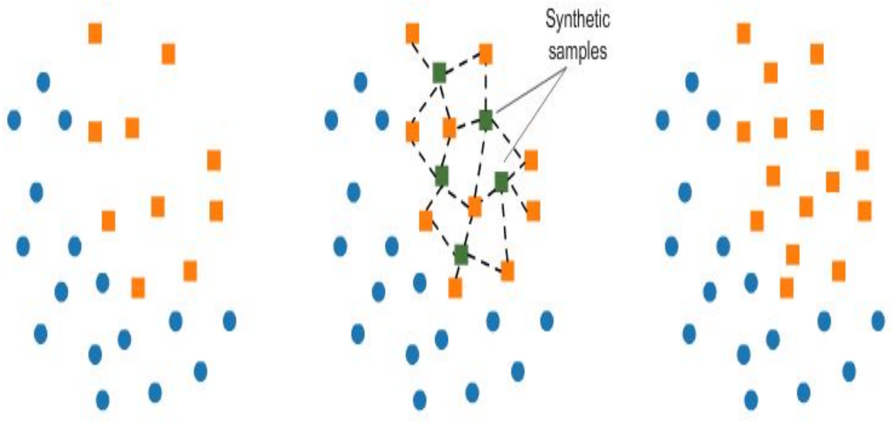


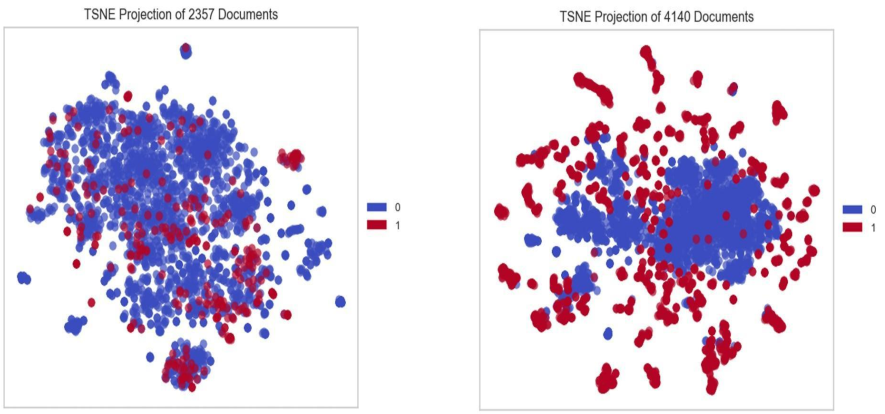
Fig B: Stratified K-Fold Cross-Validation

Stratified K-Fold Cross-Validation

Stratified K-Fold is an enhanced version of K-Fold cross-validation which is mainly used for imbalanced datasets. Just like K-fold, the whole dataset is divided into K-folds of equal size.But in this technique, each fold will have the same ratio of instances of target variable as in the whole datasets.

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.

Tree Visualization

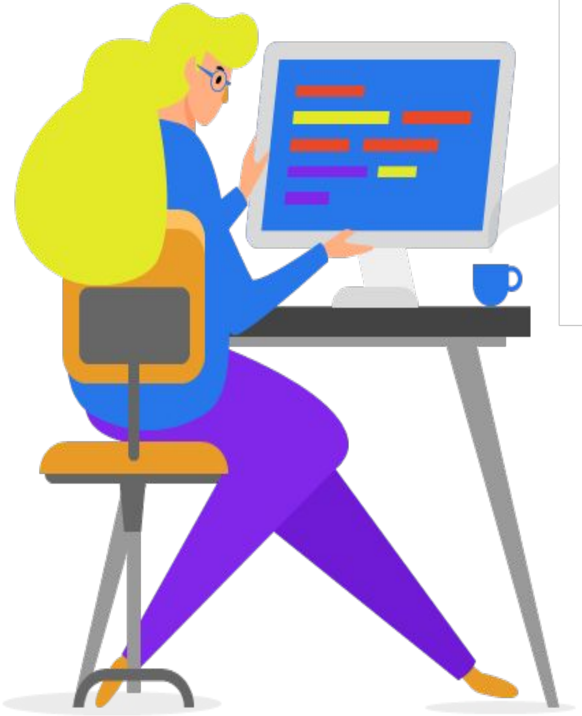

```python
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.tree import export_graphviz
import numpy as np

# Classifier
clf = GradientBoostingClassifier(random_state=0,n_estimators=100)
clf.fit(xx_train_smote,yy_train_smote)

# Get the tree number 42
sub_tree_42 = clf.estimators_[42, 0]

# Visualization
# Install graphviz: https://www.graphviz.org/download/
from pydotplus import graph_from_dot_data
from IPython.display import Image
dot_data = export_graphviz(
    sub_tree_42,
    out_file=None, filled=True, rounded=True,
    special_characters=True,
    proportion=False, impurity=True, # enable them if you want
)
graph = graph_from_dot_data(dot_data)
Image(graph.create_png())
```

Confusion matrix

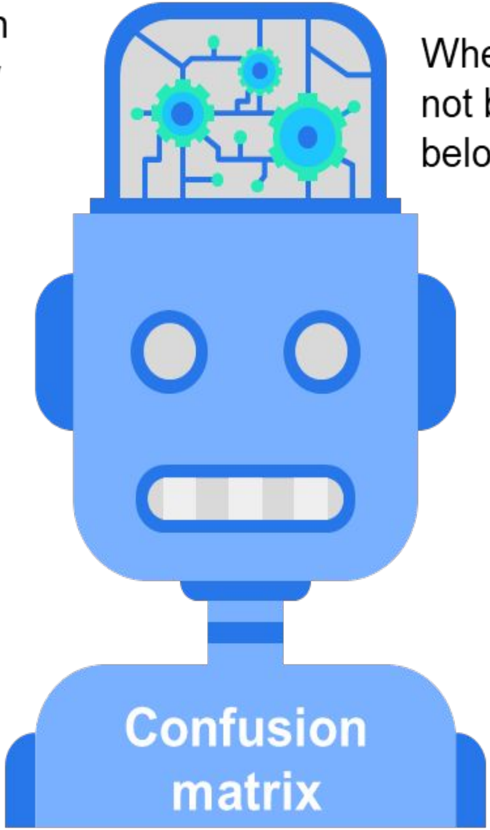


Confusion matrix terminologies

TP When you predict an observation belongs to a class and it actually does belong to that class.

TN When you predict an observation does not belong to a class and it actually does not belong to that class.

FP When you predict an observation belongs to a class and it actually does not belong to that class.



FN When you predict an observation does not belong to a class and it actually does belong to that class

FPR

Type 1 error

FP/(FP+FN)

FNR

Type 2 error

Evaluation matrix for classification

Accuracy

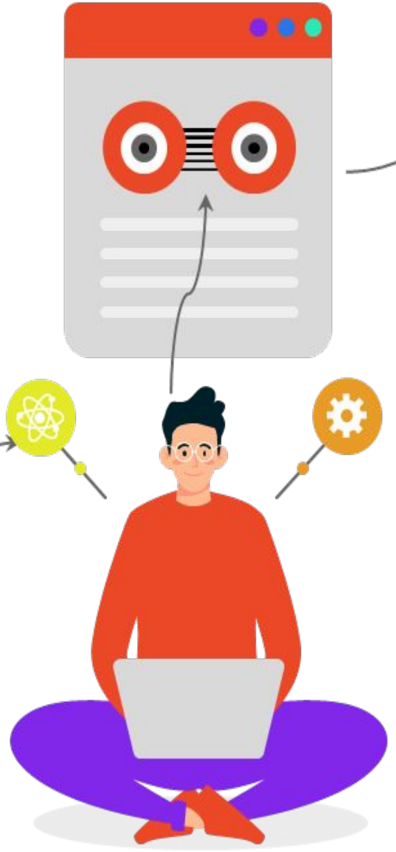
0.91

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TP}$$

F1 Score

92

2*((precision*recall)/(precision+recall))

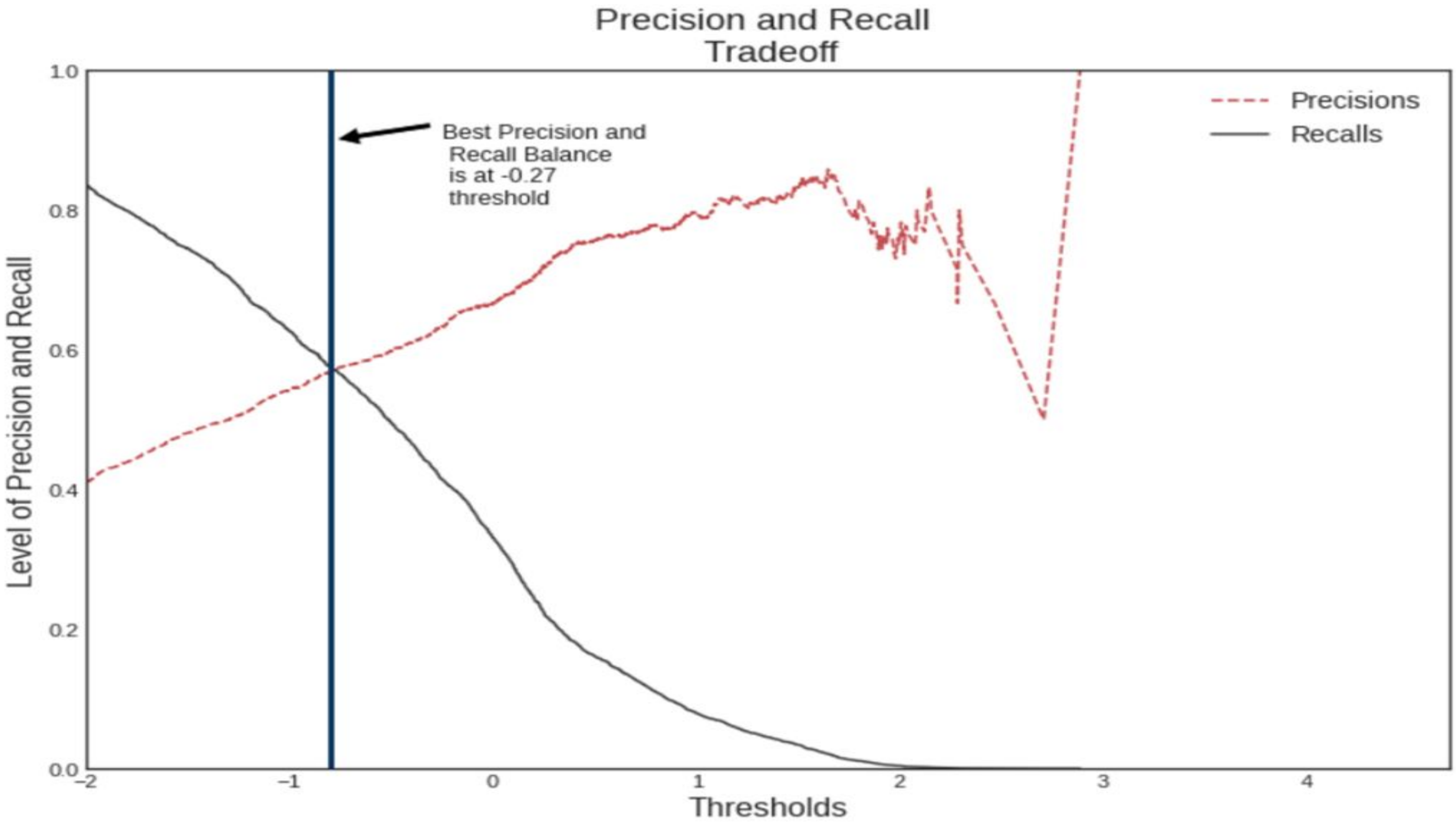


Precision

TP/(TP+FP)=91

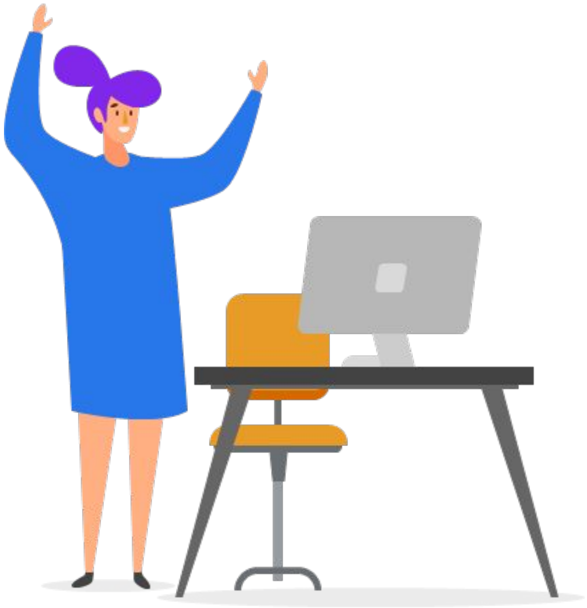
Recall

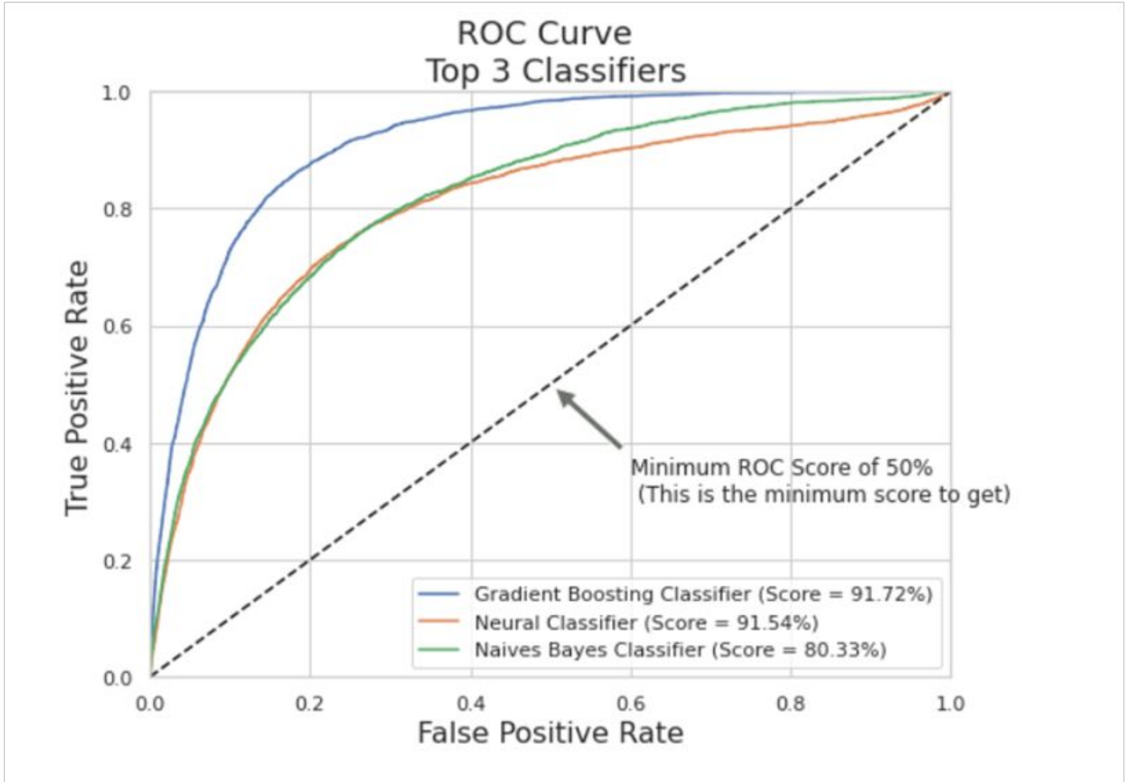
TP/(TP+FN)=92



Precision and Recall
Tradeoff
Best Precision and
Recall Balance
is at -0.27
threshold
Precisions
Recalls
Level of Precision and Recall
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
3
4

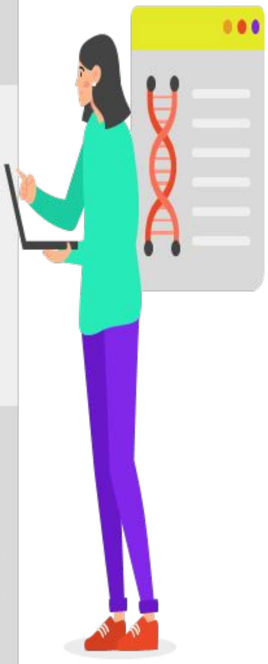
ROC AUC curve and Model performance comparison





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