PREDICTION OF SUBSCRIPTION FOR A TERM DEPOSIT

TEAM MEMBERS

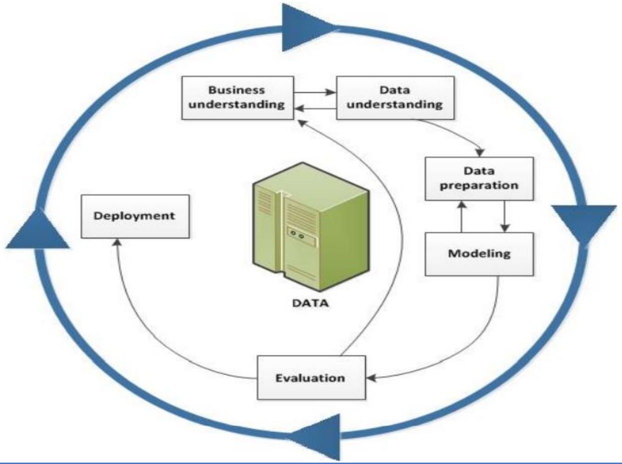
1. **Swapnil Pattanshetty**
2. **Rani Priya**
3. **Alzaman Siddiqui**
4. **Kilani Teja**
5. **Vimal Kanth**

Mentor – Anjana Agrawal

GREAT LAKES INSTITUTE OF MANAGEMENT, CHENNAI

PROBLEM STATEMENT

- In Banking Sector the older marketing options have contributed minimal in increasing the business of banks. European Banks were under pressure to increase their financial assets so they wanted to use technology to come up with a solution that increases the efficiency by making fewer calls but improves the success rate.
- In this project We will be using a Bank Marketing Campaign of a Portuguese Bank Institution happened over phone calls
- Finding out the characteristics that are helping Bank to make customers successfully subscribe for Term deposits, which helps in increasing campaign efficiently and selecting high value customers.



The Direct marketing campaign has a significant impact on the business in the following ways (Applications):

- Helps to build relationships with new customers.
- Test the appeal of your product or service (in this case a term deposit).
- Gives an insight on which marketing approaches reach your target market.
- Provide customers with compelling content they can share with potential customers.
- As an overall impact it leads to increase in sales of the product (term deposit).

OBJECTIVES

- What is the main marketing campaign factor that can increase the customer’s decision to subscribe to a term deposit?
- How accurate can we be in predicting the customer's decision to subscribe to a term deposit?
- Business interpretation of the different models using Visualisation
- Business evaluation to convince that our model predicts the best

DATA INFORMATION

1) Bank client

Sr No.	Variable	Datatype
1	Age	int64
2	Job	object
3	Marital	object
4	Education	object
5	Default	object
6	Housing	object
7	Loan	object

2) Related with the last contact of the current campaign

Sr No.	Variable	Datatype
1	Contact	object
2	Month	object
3	Day_of_week	object
4	Duration	int64

- Features – 21
- Rows – 41188
- Categorical Features – 10
- Numeric Features – 10
- Y is target (Term Deposit)

Output variable (desired target- Term Deposit)

Sr No.	Variable	Datatype
1	Y	object

3) Other attributes

Sr No.	Variable	Datatype
1	Campaign	int64
2	Pdays	int64
3	Previous	int64
4	Poutcome	object

4) Social and economic context attributes

Sr No.	Variable	Datatype
1	Emp.var.rate	float64
2	Cons.price.idx	float64
3	Cons.conf.idx	float64
4	Euribor3m	float64
5	Nr.employed	float64

DATA PREPROCESSING

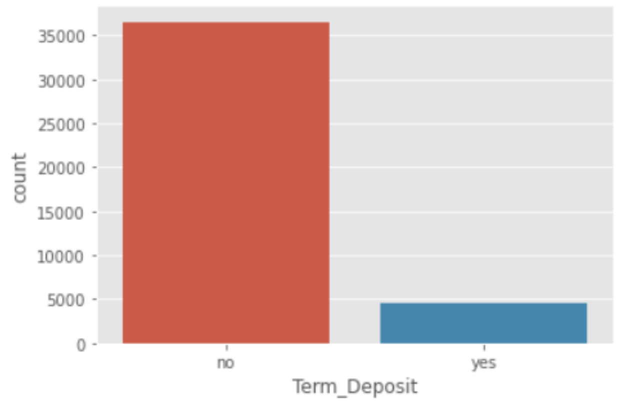
Null Value Treatment

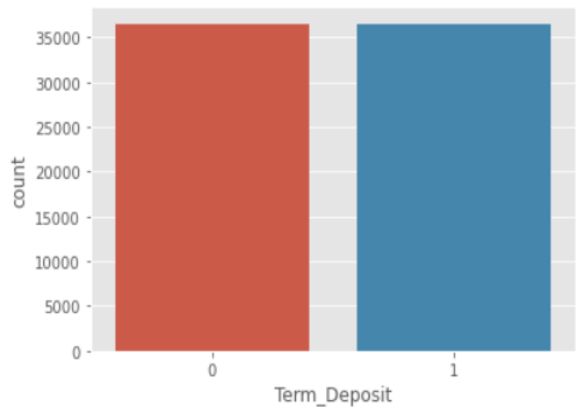
Attribute	Null Value Percentage
housing	0.0240361
Default	0.208725
Martial_status	0.000024
loan	0.0240361

- The Data technically doesn't have Null values rather camouflaged as "unknown" values
- To treat these unknown values we use mode imputation on categorical variables
- The default variable has more than 20% unknown values,
- So we are going to consider it as a separate entity called as Non-Existant and notice how it is affecting the column while model building

Handling the Imbalanced Data

Our Target variable consists of imbalanced data which can affect our model building So we are using SMOTE technique to add more synthetic rows.






```
Y["Term_Deposit"].value_counts()*100/Y["Term_Deposit"].count(
0    88.734583
1    11.265417
```

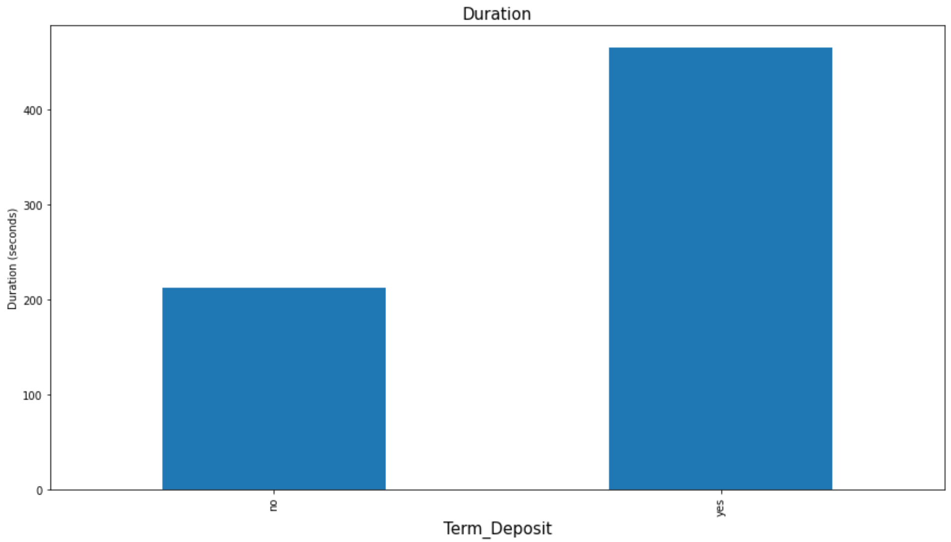
```
Y1["Term_Deposit"].value_counts()*100/Y1["Term_Deposit"].count()
0    50.0
1    50.0
```

Before Smote

After Smote

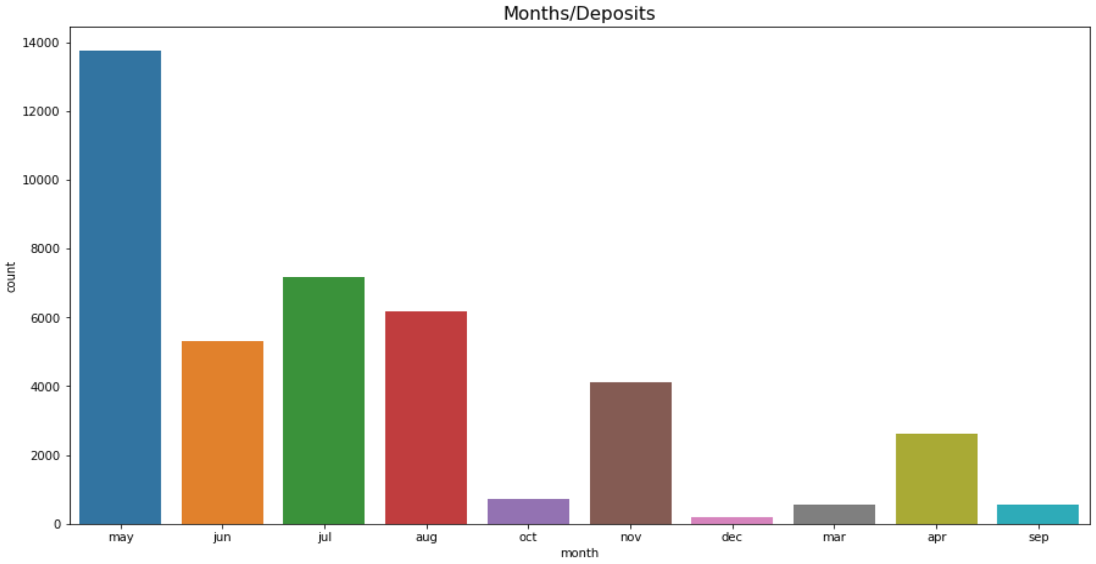
Business Questions:

1 - What is the average duration (in seconds) of the call for those who did not make a term deposit (0) ? And for those who made term deposits (1)?



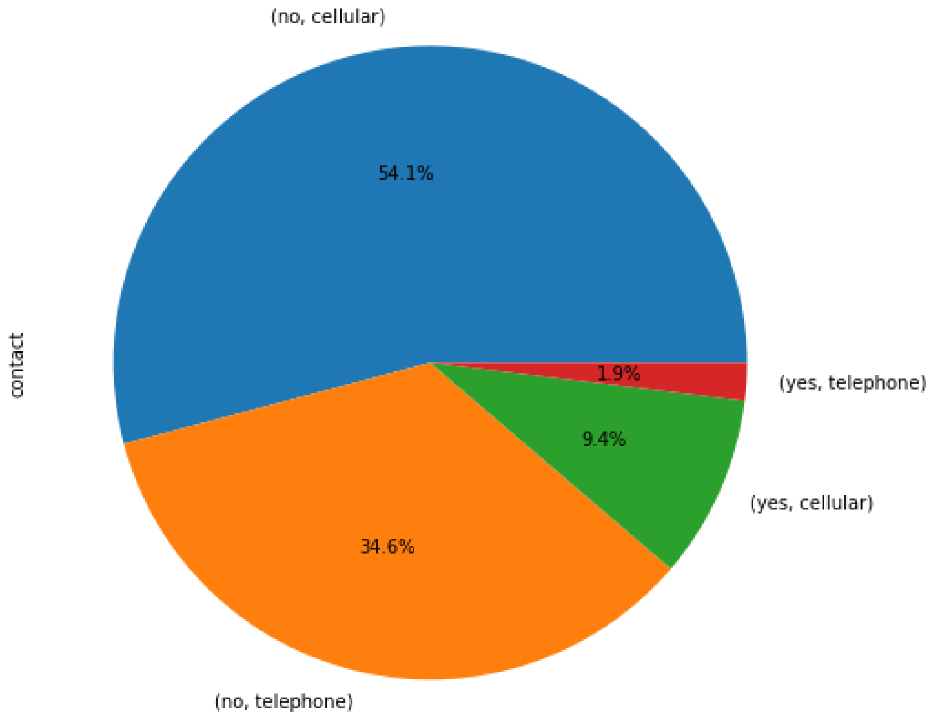
Duration :For those who made term deposits (1), the average time was 465 seconds.
For those who did not make
a term deposit (0), the average time was 213 seconds.
It means that,
for a customer to make a term deposit,
more time is needed to convince him/her.

2 - In which month do customers usually make the most deposits?



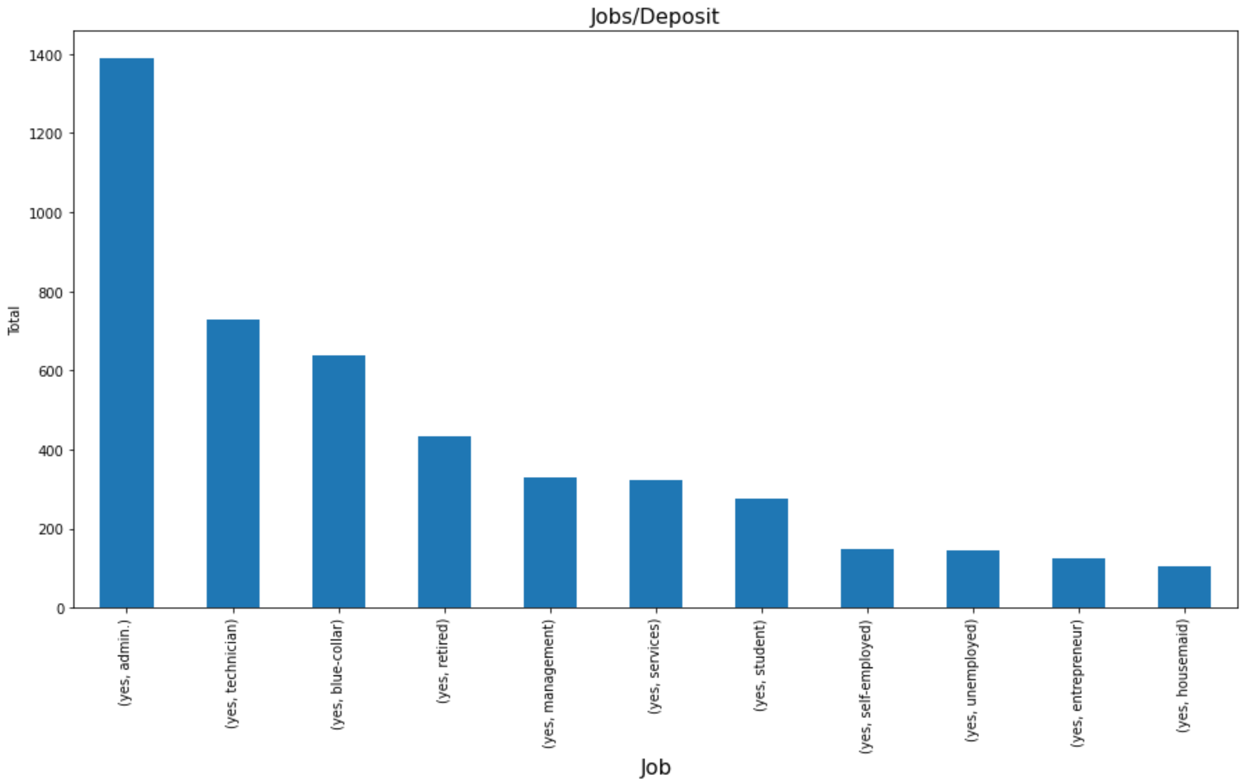
Month: In month variable most of the subscription happened during month of May but during September, October, March with respect to the highest conversation ratio almost contacted person taken subscription.

3 - Among those who made bank deposits, what was the main form of contact?



contact : The main form of contact is the cellular. Few customers who made term bank deposits were contacted by telephone.

4 - What type of job is most common among those who made bank deposits?



Job: people with admin jobs have been contacted more by the bank.

People with unknown jobs are very few.

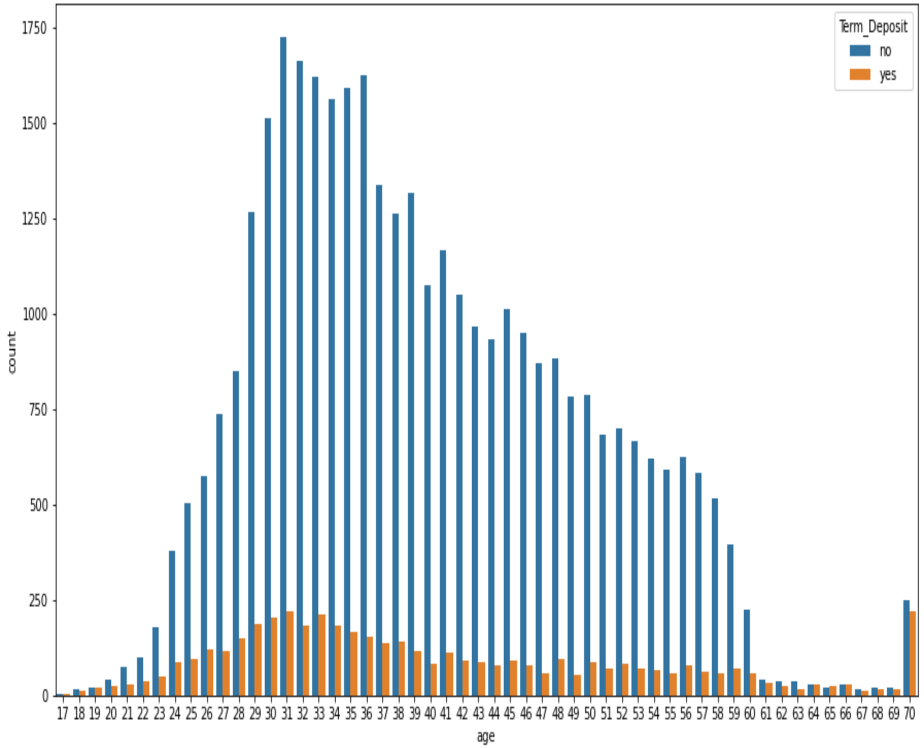
As we can see people with admin jobs subscribed the most.

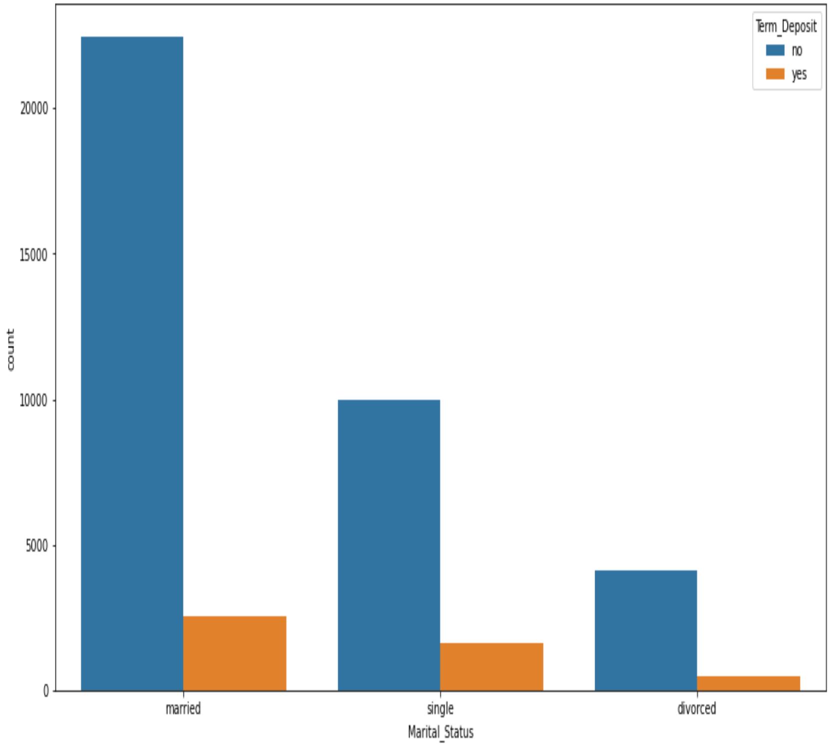
Implications

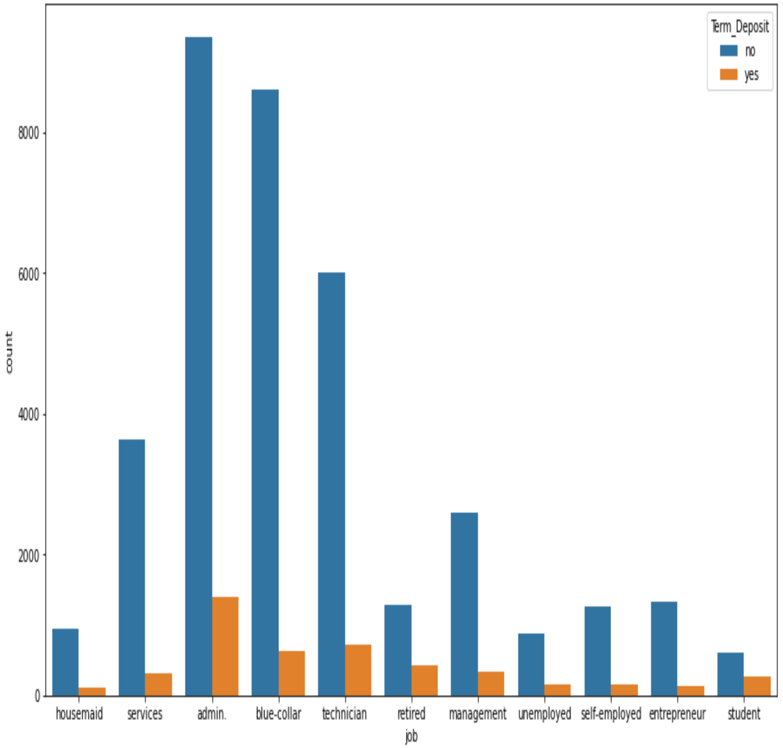
Age: Most of the data we got was from the ages between 25-50 but the people who has subscribed to term deposit are the older people who are above 55 so the conclusion for this attribute is to target older generation people more

Marital Status: Marital status had less impact on target
variable which we have noticed in the feature importance,
so this attribute need not be considered in making important decisions
on whom to target.

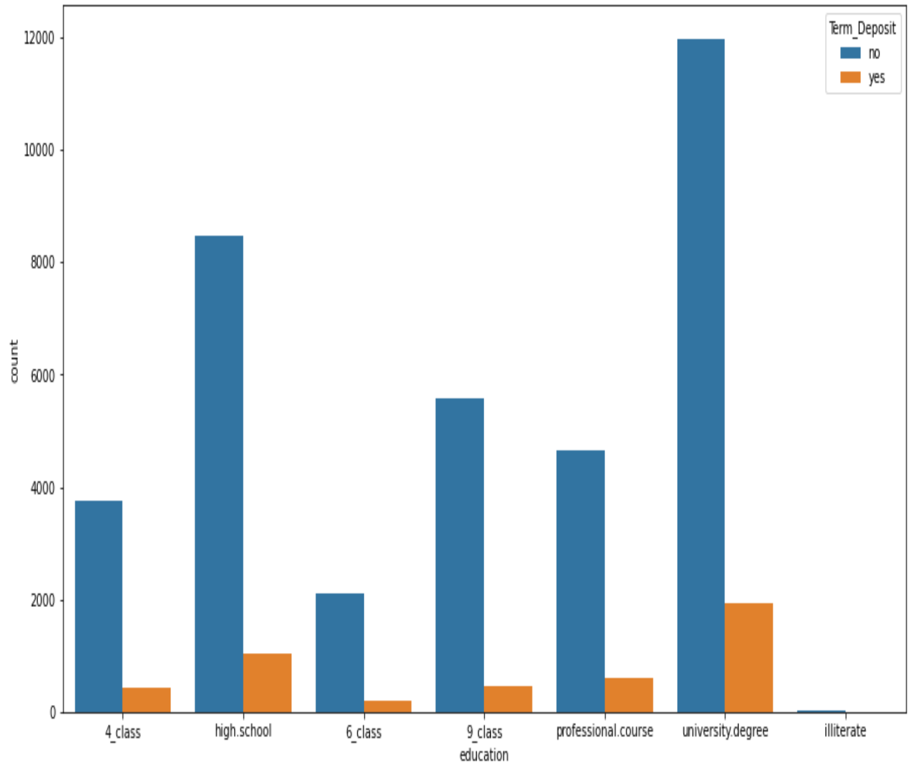
Jobs: The people who have been target the most are admin, blue-collar and technicians and these people have subscribed more as well but other occupations such as who are students even though they have been targeted less but the conversion rate for students is better than other attributes so need to target more people who are students



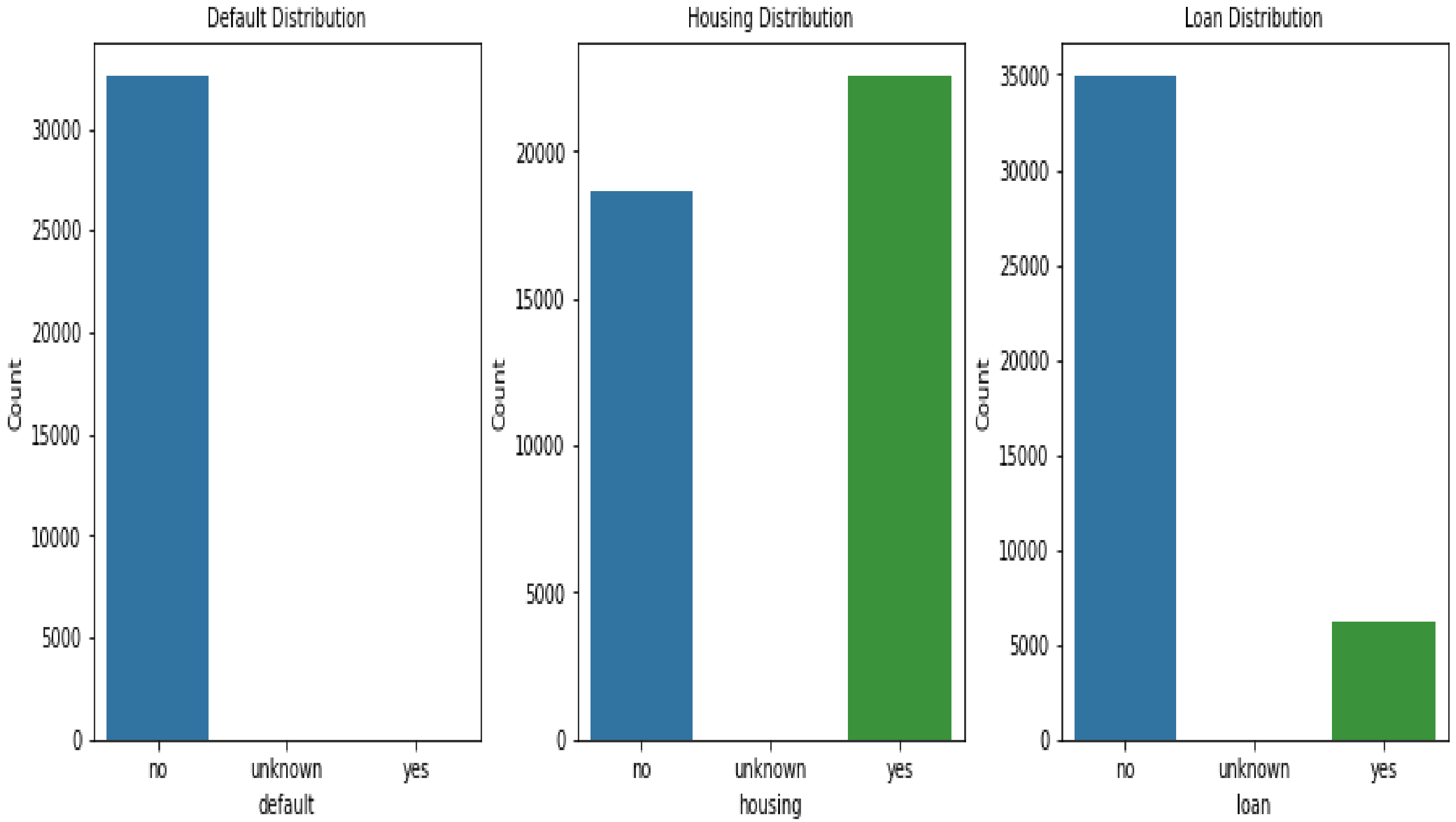




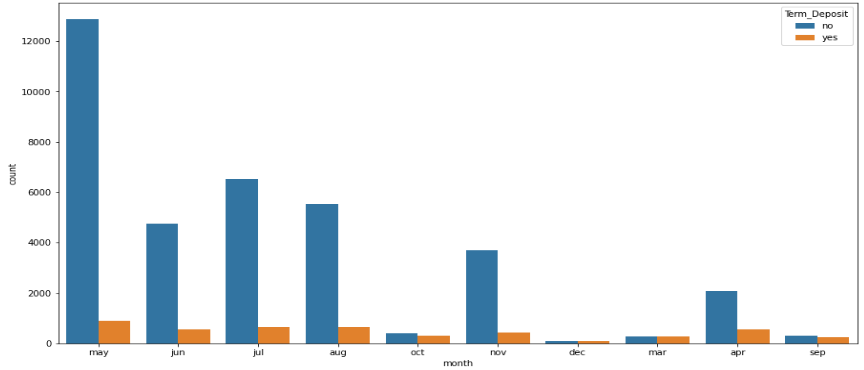
Education: People with university degree and high school have been targeted the most but we can see that the people who subscribe to term deposit are mostly from university degree and people who are pursuing a professional course since they are more mature to understand the importance of term deposit.

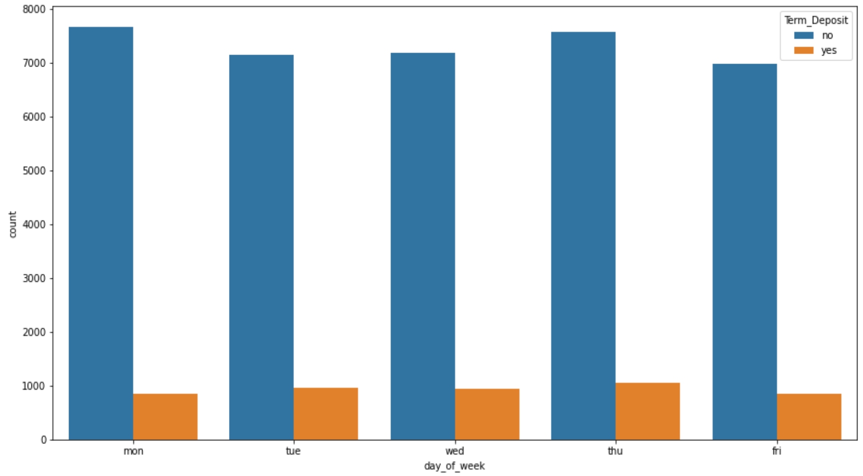


Housing , Loan and Default: From these 3 attributes we can notice that people who are financially independent and less burdened by loan, or housings tend to subscribe to more term deposit, so we can target accordingly to more financially independent people.

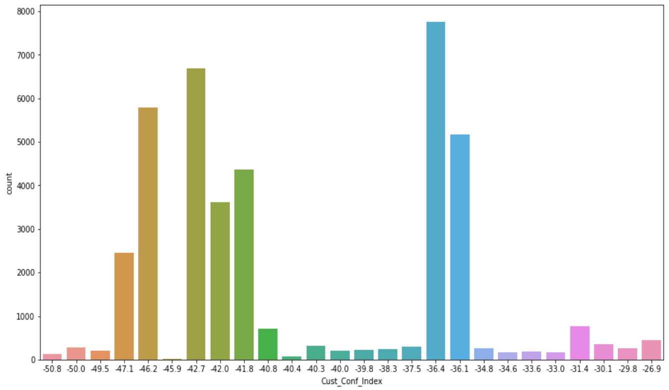


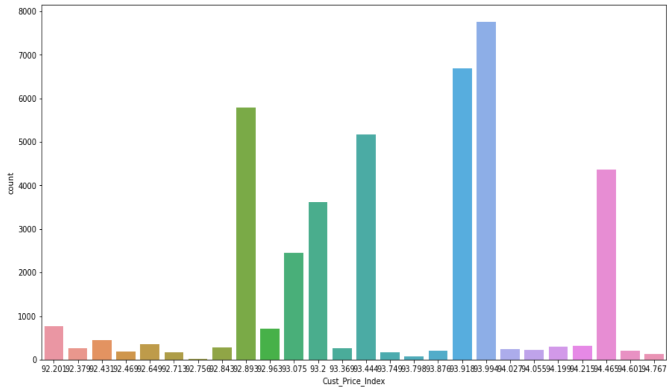
Month and Day of week: May is the month where the people have been targeted along with some other months June July, this indicates that people have been targeted more after the beginning of new fiscal year but results show that the conversion rate has been the most during the end of the year starting from September to December there is a high chance of subscribing to term deposit, The day of the week doesn’t matter much interms of subscriptions and they have been targeted equally.

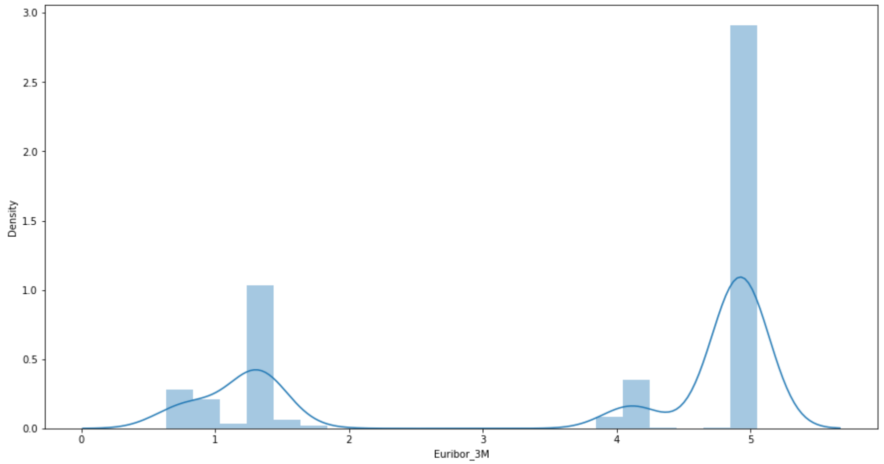




Social and Economic Attributes: These attributes such as Customer Confidence index, Customer price index, Euribor_3M are some of the most important attributes which are significant in deciding whether the person will subscribe to term deposit or not, so the people with higher numbers in these attributes should be targeted more and these attributes are significant in model building as well.

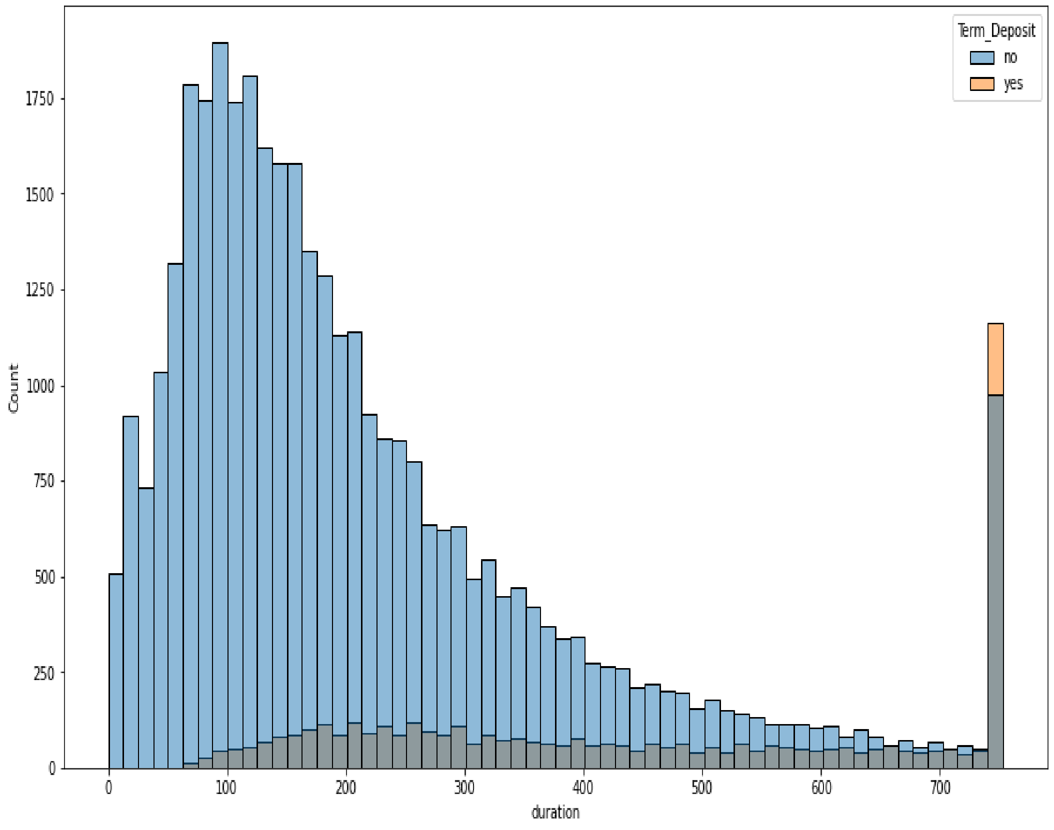


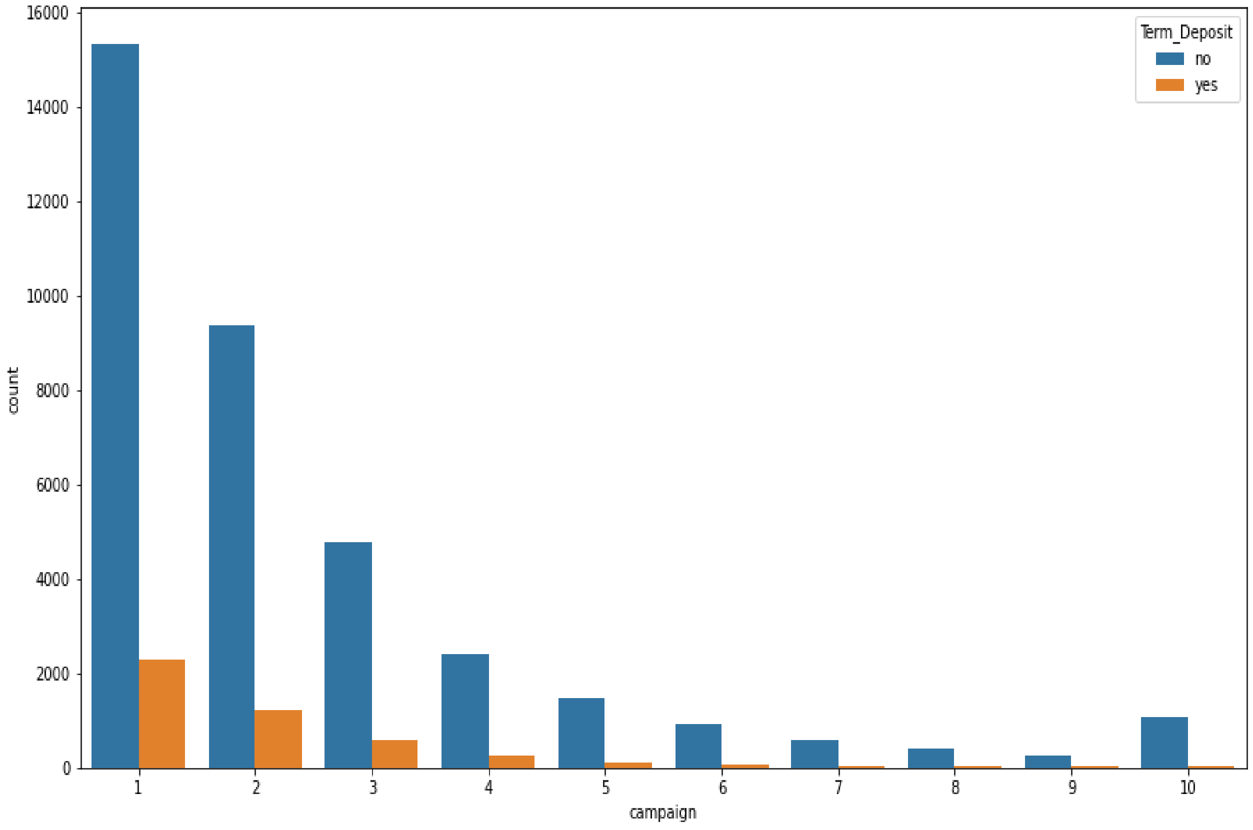




Duration: duration is the most important attribute of all the attributes, this tells us that people who have been engaged more with the customer executives are tend to subscribe more to the term deposit, so we can conclude that people should be more engaged one on one to influence them into subscribing to term deposit more.

Campaign: The campaign attributes which are related to previous campaign success rate are not that important from the existing data plus these lack data points to inference more from them, so we can conclude that the success of previous outcome doesn't decide the success of current campaign.





Statistical Analysis

Chi-Square Test

	p-value
job	6.808200e-200
Marital_Status	5.915954e-27
education	8.692046e-39
default	5.161958e-89
housing	2.547931e-02
loan	3.763354e-01
contact	1.525986e-189
month	0.000000e+00
day_of_week	2.958482e-05
poutcome	0.000000e+00

Test for Normality, Variance & Skew

	Shapiro	levene	skew
Age	0.0	1.096398e-150	0.784697
Duration	0.0	0.000000e+00	3.263141
Campaign	0.0	1.438210e-29	4.762507
Prev_Contacted_Duration	0.0	0.000000e+00	-4.922190
Prev_count	0.0	0.000000e+00	3.832042
emp.var.rate	0.0	3.164752e-07	-0.724096
Cust_Price_Index	0.0	2.023126e-53	-0.230888
Cust_Conf_Index	0.0	3.747063e-227	0.303180
Euribor_3M	0.0	1.086785e-06	-0.709188
No_employed	0.0	6.363730e-225	-1.044262

Mann-whitney U Test

	p-value
age	1.608054e-02
duration	0.000000e+00
campaign	3.418527e-38
Prev_Contacted_Duration	0.000000e+00
Prev_Count	0.000000e+00
emp.var.rate	0.000000e+00
Cust_Price_Index	9.572609e-136
Cust_Conf_Index	5.901951e-17
Euribor_3M	0.000000e+00
No_employed	0.000000e+00

From the statistical tests conducted above, we assessed the significance of features with respect to the Target variable, Term_Deposit

Odds for each variable

	Odds
const	0.051005
age	1.089690
duration	3.335775
campaign	0.852085
Prev_Count	0.971874
job	1.107602
Marital_Status	1.051678
education	1.090711
default	0.724272
housing	1.002046
loan	0.984845
contact	1.271910
month	1.437489
day_of_week	0.988040
poutcome	1.790610

- Odds_const: The odds of customer subscribing term deposit is 0.05, considering all other variables take zero value
- Odds_duration=3.335,it implies that the odds of customer subscribing term deposit increases by factor of 3.335 due to one unit increase in duration, keeping the other variables constant
- As we can notice from the table, duration seems to have the most impact on prediction of the term deposit subscription among all variables

Model Building

The algorithms that are used for classification are:

- Decision Tree
- Random Forest Classifier
- XG Boost

Decision Tree

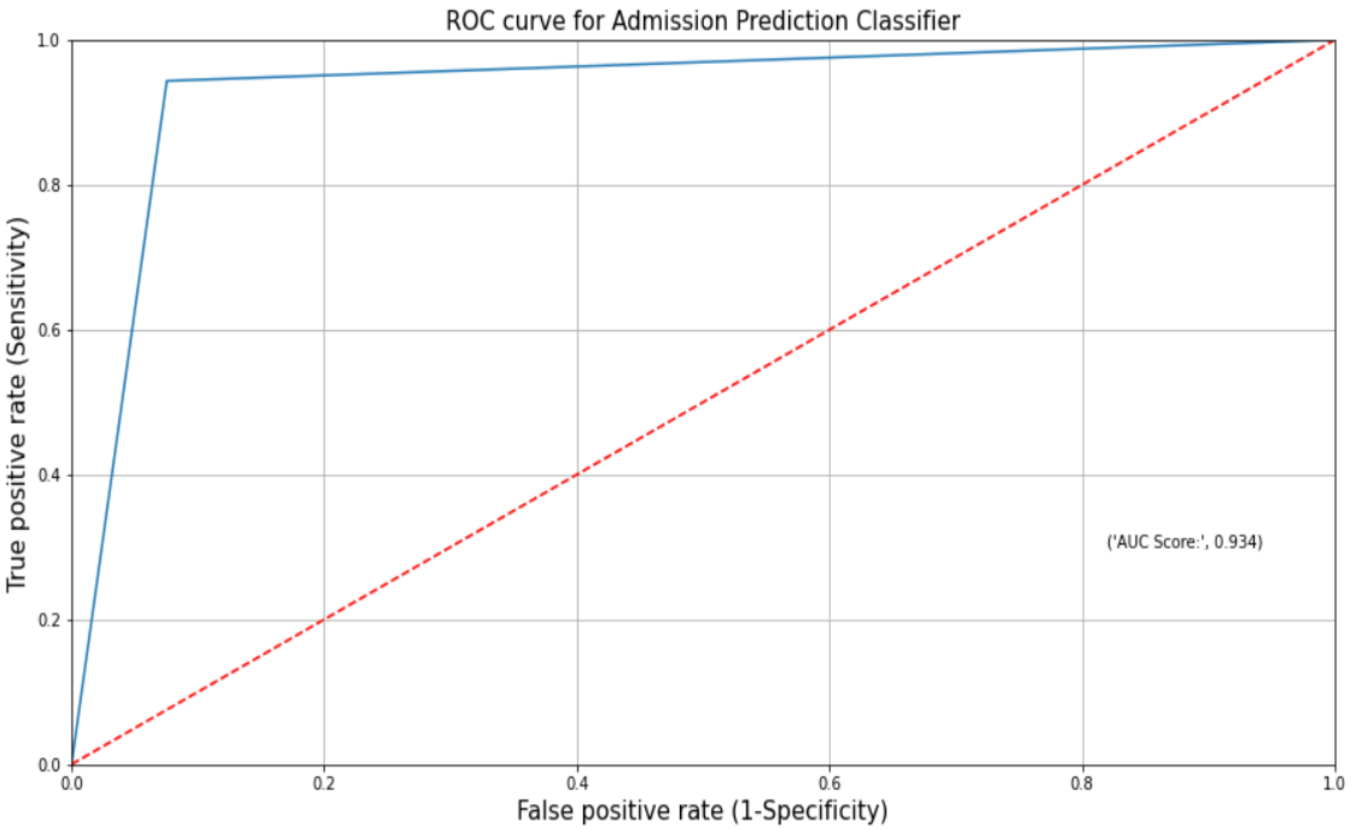
Classification report on Train set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	29216
1	1.00	1.00	1.00	29260
accuracy			1.00	58476
macro avg	1.00	1.00	1.00	58476
weighted avg	1.00	1.00	1.00	58476

Classification report on Test set

	precision	recall	f1-score	support
0	0.94	0.92	0.93	7332
1	0.93	0.94	0.93	7288
accuracy			0.93	14620
macro avg	0.93	0.93	0.93	14620
weighted avg	0.93	0.93	0.93	14620

ROC Curve



AUC score: 0.93

Accuracy Score = 0.9339261285909712

XG Boost

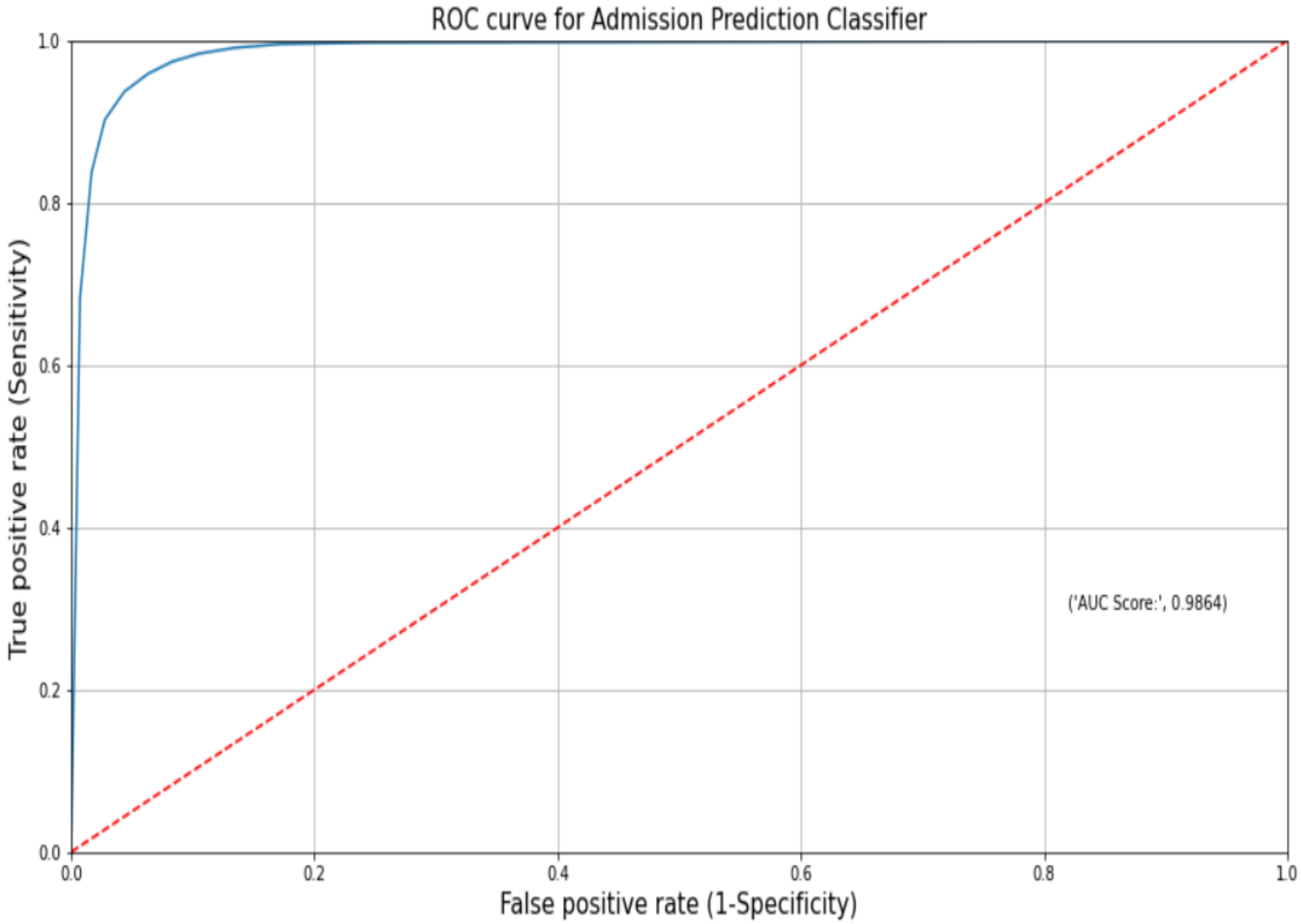
Classification report on Train set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	29216
1	1.00	1.00	1.00	29260
accuracy			1.00	58476
macro avg	1.00	1.00	1.00	58476
weighted avg	1.00	1.00	1.00	58476

Classification report on Test set

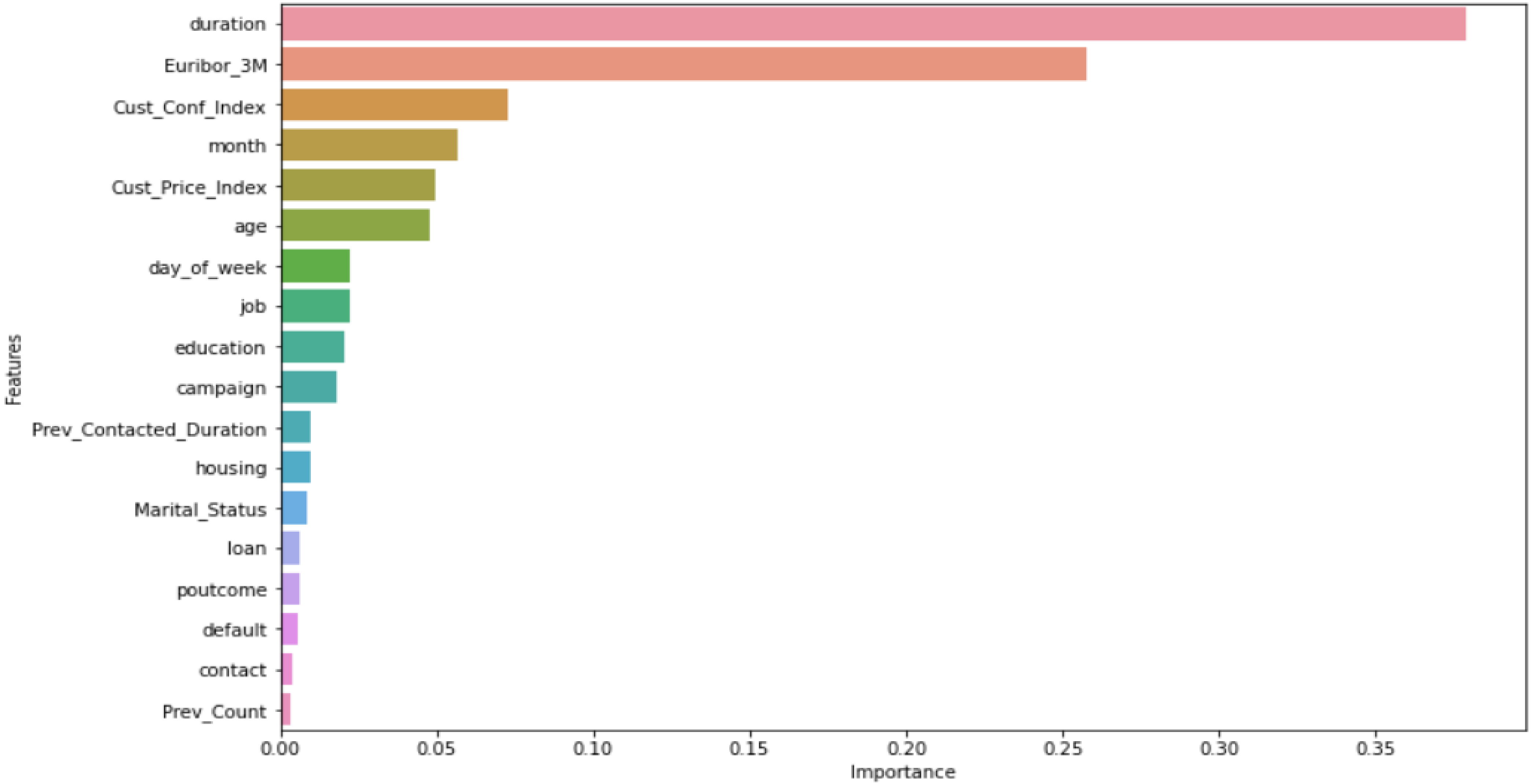
	precision	recall	f1-score	support
0	0.96	0.94	0.95	7332
1	0.94	0.96	0.95	7288
accuracy			0.95	14620
macro avg	0.95	0.95	0.95	14620
weighted avg	0.95	0.95	0.95	14620

ROC Curve



AUC score: 0.9889 **Accuracy Score = 0.9454856361149111**

Important features



Hyper Parameters : Best parameters for decision tree classifier: ('criterion': 'gin', 'max_depth: 9, 'min_samples_split': 7}

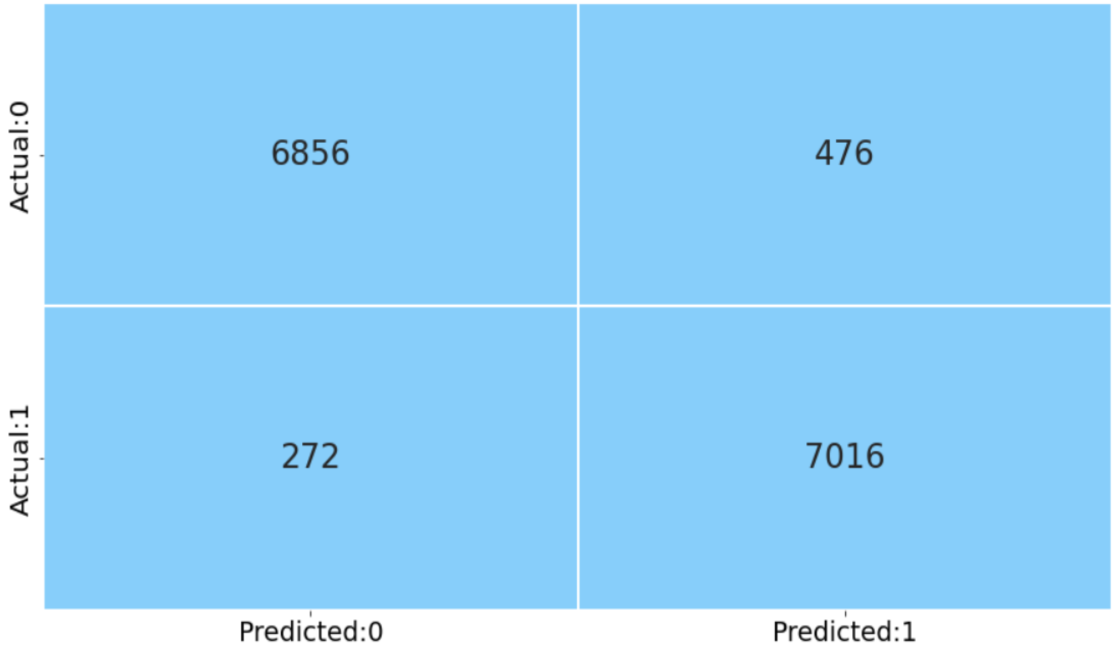
Model Evaluation after selecting important features and hyperparameter tuning.

Best parameters for XGBoost classifier: ('gamma': 0, 'learning_rate ': 0.3, 'max_depth":9}

Classification report:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	7332
1	0.94	0.96	0.95	7288
accuracy			0.95	14620
macro avg	0.95	0.95	0.95	14620
weighted avg	0.95	0.95	0.95	14620

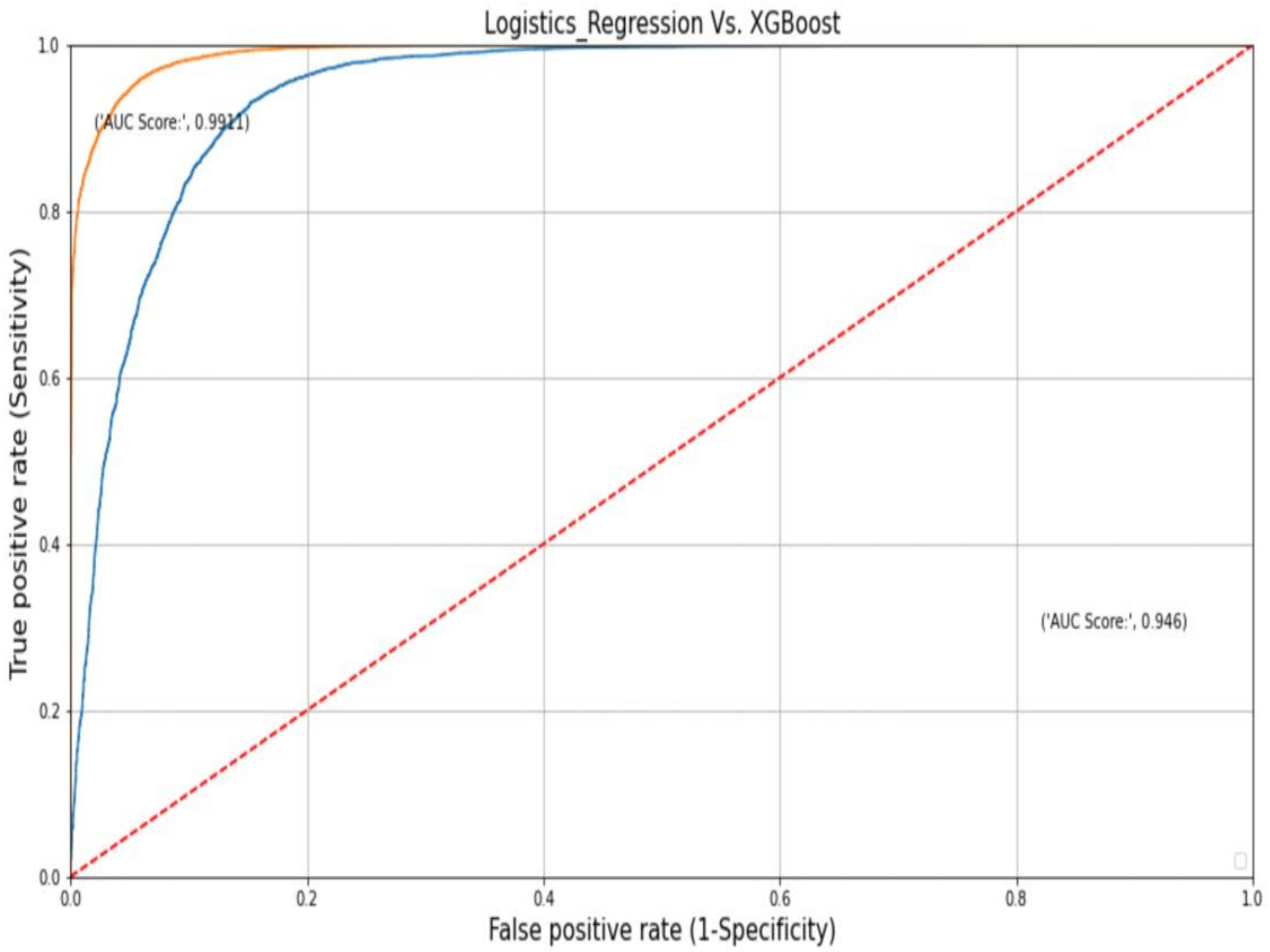
Confusion matrix:



AUC score: 0.9889

Accuracy score = 0.9488372093023256

Updated Model Performance compared to Benchmark



Overall Model Results

Model	Accuracy Score
Decision Tree	0.92
Random Forest	0.95
XG Boost	0.95

UPDATED WORK COMPARED TO BENCHMARK

Treating Outliers

Finding Important Features

Hyper Parameter Tuning

Different model building like Ensemble Methods

Selecting Best Prediction model

Evaluation

Business Interpretation

Limitations

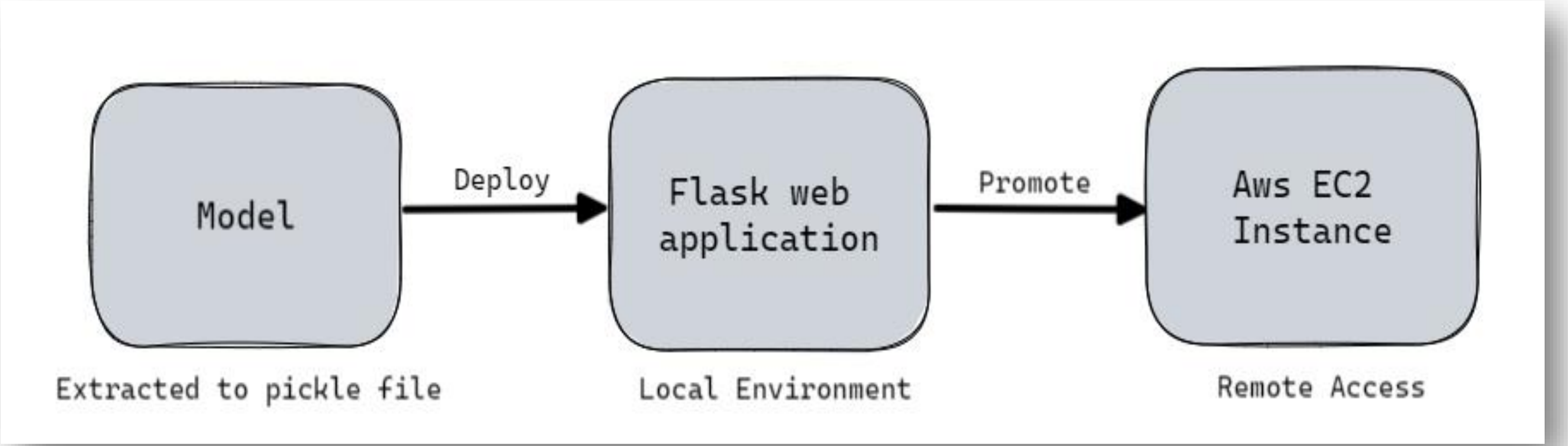
- This research is subject to limited data. Concerning the restrictions in the banking industry, The research has limited access to the data. Many banks would not allow researchers to get information on some essential aspects of the business their marketing strategies due to fear of competition. As a result, this research may not identify the actual campaign strategies used by the selected banks.
- The model which was built was mainly influenced by the social and economic attributes of a person , this may have resulted in not understanding the importance of an individual's personal attributes while predicting the subscriptions of Term deposit
- The data collected from bank may have refused to share some of information of attributes related to previous campaign so it had a lot of missing or non-existent values, Because of this the model may not be able to consider the importance of previous campaign outcomes while predicting the output.

Conclusion

In this research work, the resampling method was used in dealing with the problem of imbalanced data, and three machine learning algorithms (Decision Tree, Random Forest and XGBoost) were deployed to find out the main factor that influences customers decision to subscribe to a term deposit in the bank. Feature Importance and Hyperparameter Tuning was then used to get the most important factors that influence customer decisions and was then retrained to perform the second experiment. XGBoost had the best accuracy, precision, recall and auc-roc score out of all the models The correlation heatmap and feature importance highlighted five factors that can influencing the customer's decision and 'duration' has the highest correlation coefficient with (positive correlation) of 0.41. Which means that the longer the bank continue to advertise their product and service, the more customers can subscribe to a term deposit. Followed by social and economic attributes and month and age as well.

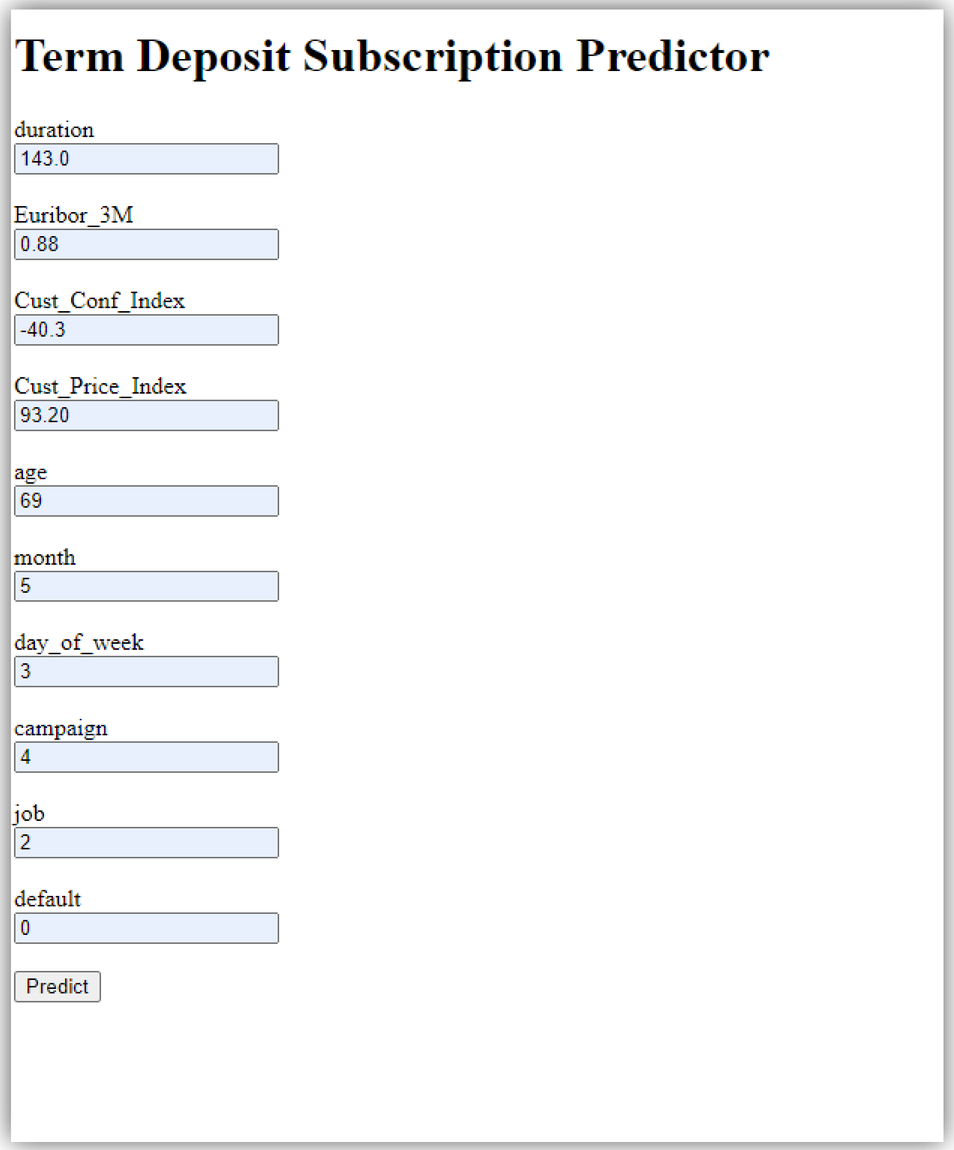
Model Deployment

- Request sent via REST API
- Cleaned & preprocessed by Feature Extractor
- Trained model is used in giving the predictions.
- Model can be deployed as a pickle file.
- Flask is used to manage HTTP requests/responses. It runs on WSGI server toolkit and Jinja2 template engine.

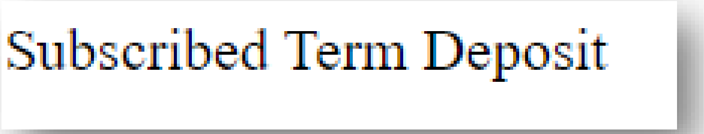


Screenshot

Link : [Term Deposit Subscription Predictor]()



On clicking predict :




Thank you!