IIT-M Certified Advanced Programmer with Data Science Mastery Program

By

GUVI - an IIT-Madras incubated company

IITMDSA DW42DW43

Final Project report on

Predicting Term Deposit Subscription by a client

Submitted by,

Gururaj H C

DW42DW43

gururaj008@gmail.com **+8050461594**

https://www.linkedin.com/in/gururaj-hc-machine-learning-enthusiast/

Title: Predicting Term Deposit Subscription by a client

Abstract:

Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

A Term deposit is a deposit that a bank or a financial institution offers with a fixed rate (often better than just opening a deposit account) in which your money will be returned back at a specific maturity time.

Problem Statement:

Predict if a customer subscribes to a term deposits or not, when contacted by a marketing agent, by understanding the different features and performing predictive Analytics

About the dataset:

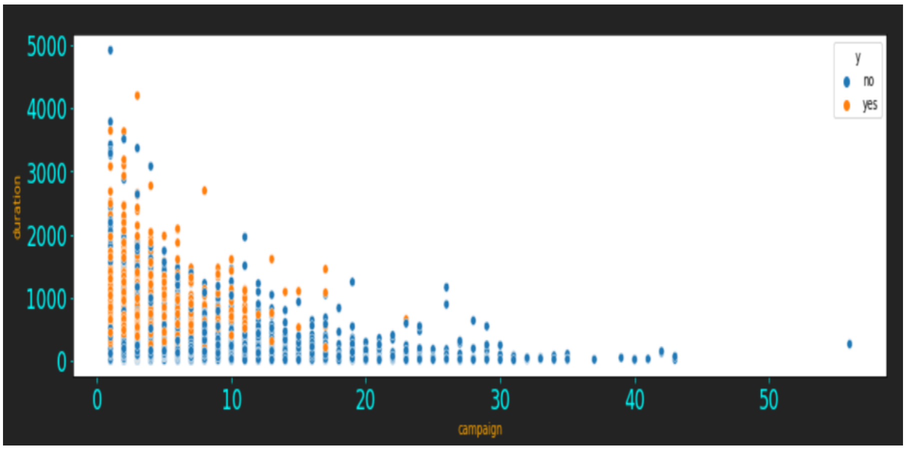
The dataset is comprising of 41188 instances with a total of 20 features and a target (‘y’). If ‘y’ is Yes then the customer subscribed for term deposit, if its no then the customer didn’t subscribed for term deposit.

Data Pre-preprocessing:

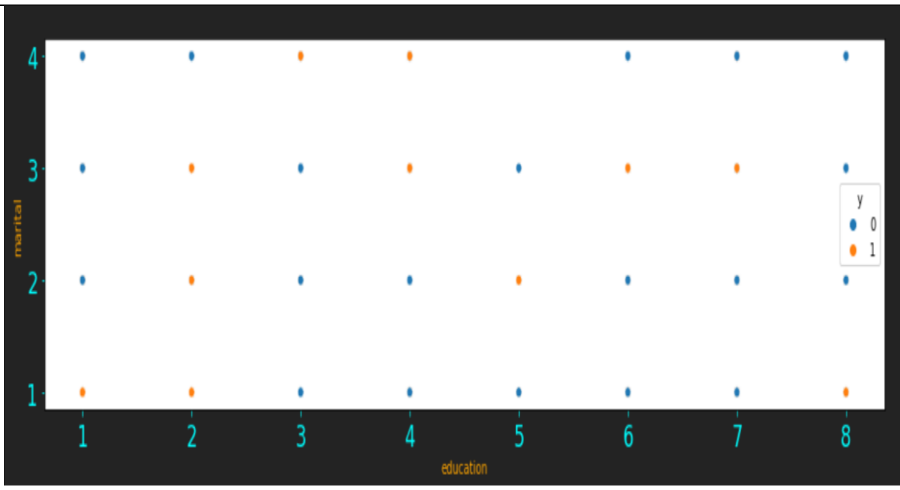
- No Null Values in the dataset hence no imputation required
- The mean value of features is on different scale for most of the features.Hence, Scaling is required
- The target has 36458-No and only 4640-Yes. Hence, the dataset is imbalanced
- Dataset can be balanced by using- Undersampling, Oversampling and SMOTE (Synthetic Minority Over-sampling Technique) techniques.

Exploratory data analysis

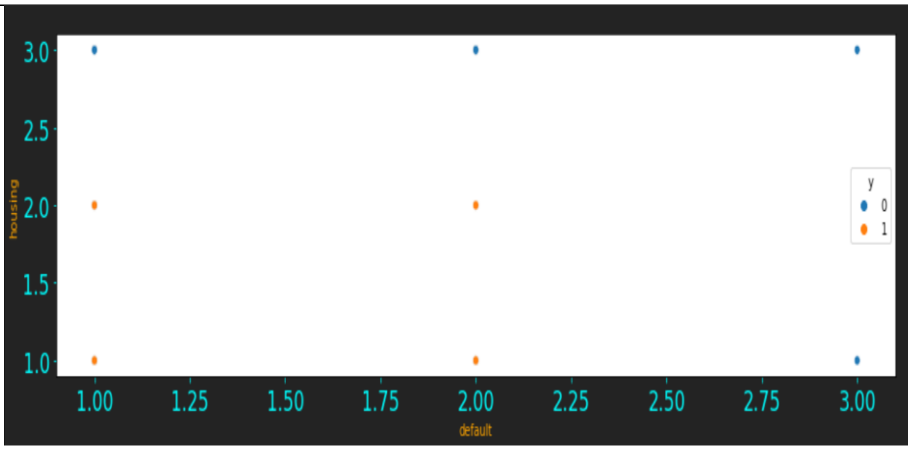
Relationship between the important features as a pair to explore their impact on target



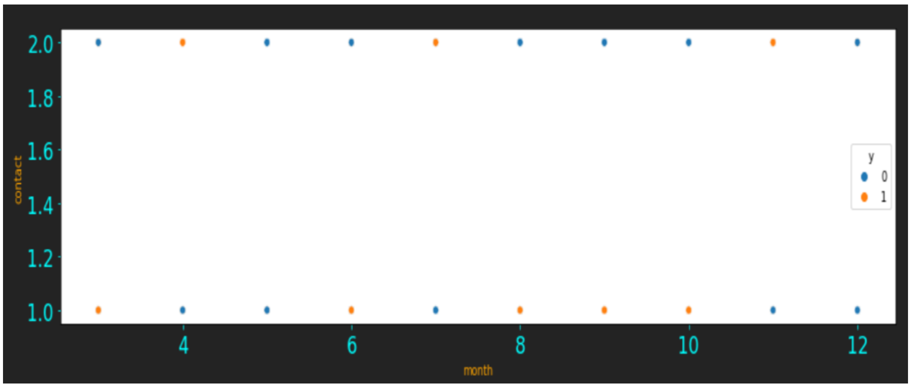
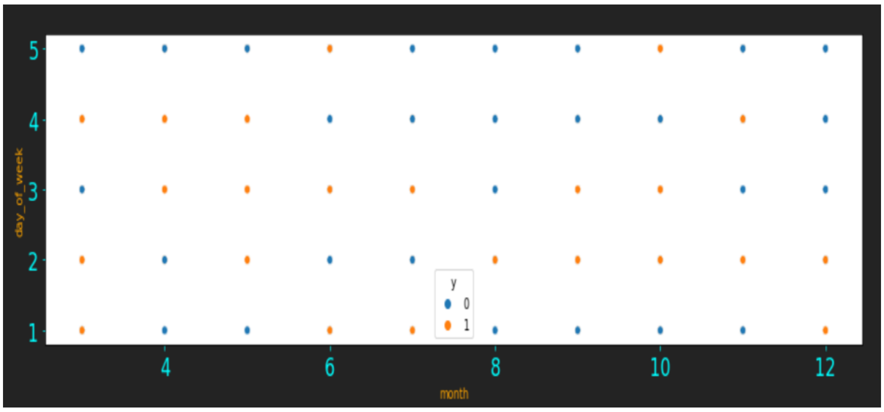
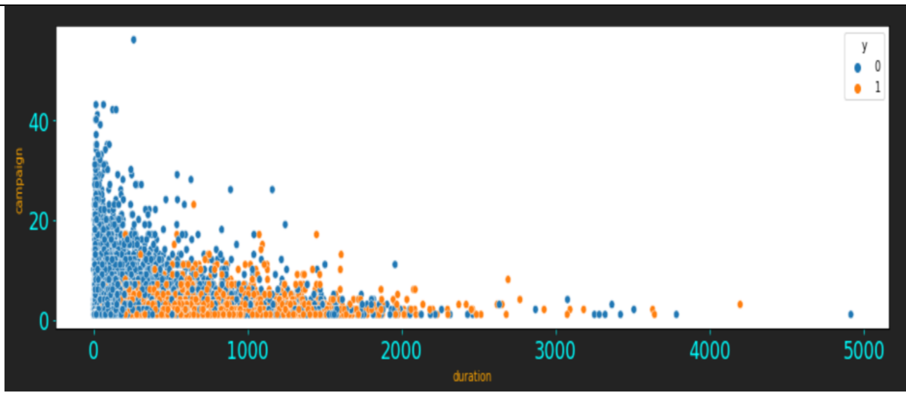
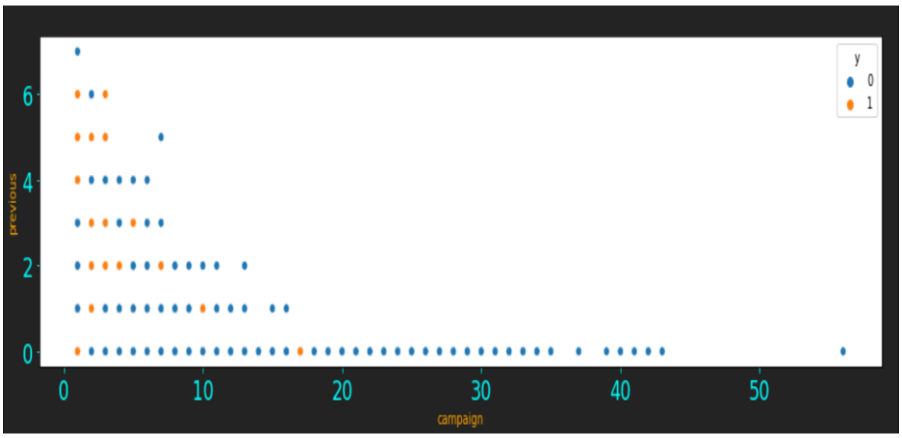
The optimum number of contacts per campaign is less than 5 and duration of the call less than 3000 secs for getting the Term deposit, beyond these numbers customers going for term deposit decreases rapidly

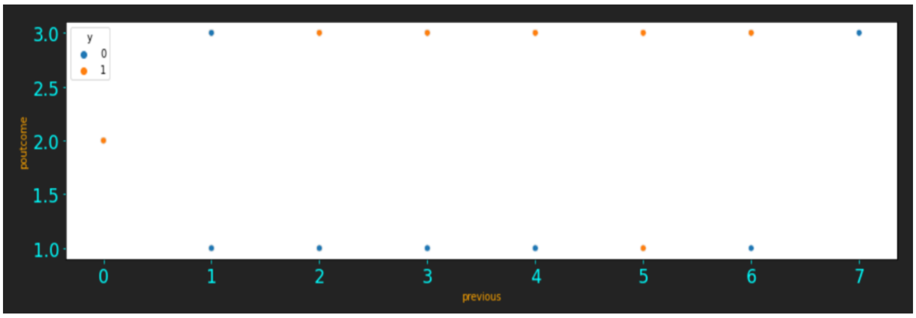


Those who are single and have education qualification of : 'basic.6y', 'high.school', 'professional.course', 'university.degree' are more likely to go for tem deposit



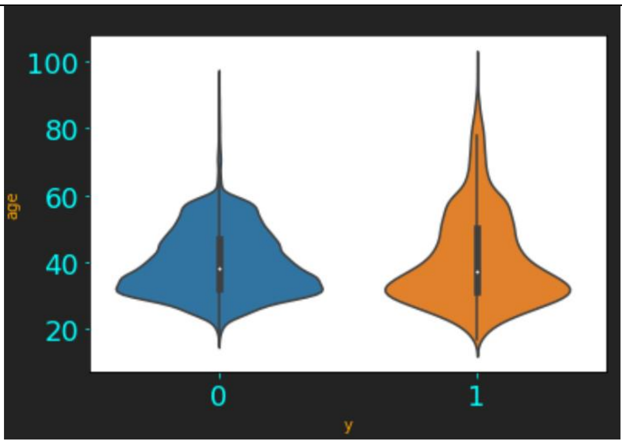
The customers whose default status is no and unknown are more likely to go for term-deposit

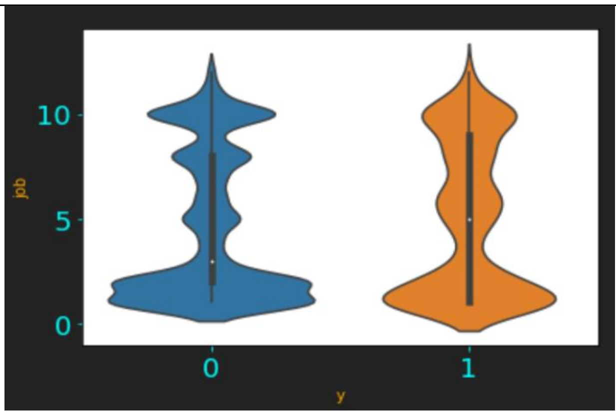
	The customers contacted on the months of March, June, August, September, October through cellular phone are more likely to go for Term deposit
	Most fruitful days for communication are Tuesday and Wednesday, regardless of the month communicated
	Successful campaigns are those where the number of contacts are less than 15 and duration of the call is 200-1400
	Customers went for term deposit when the number of contacts made during the campaign were less than 5 and when they were contacted more than twice prior to the campaign

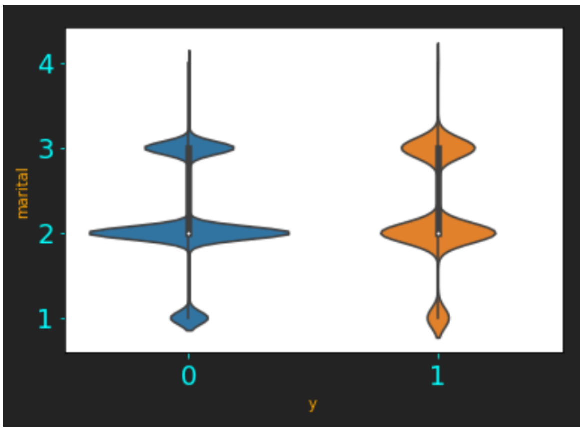
<table>
<tr>
<td>


</td>
<td>Previous marketing campaign succeeded when the customers were contacted 2 to 6 times</td>
</tr>
</table>

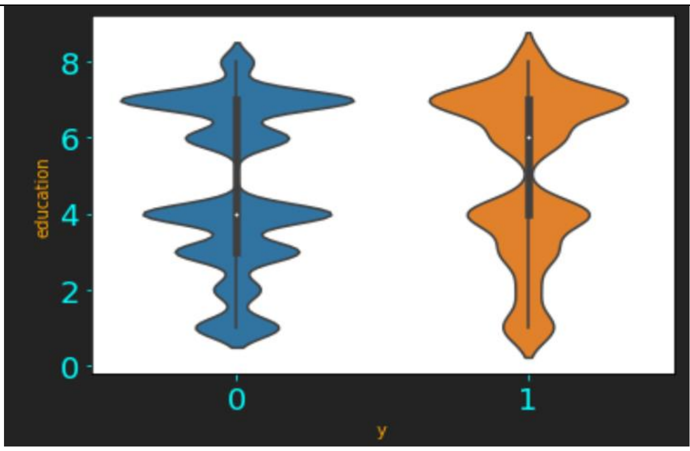
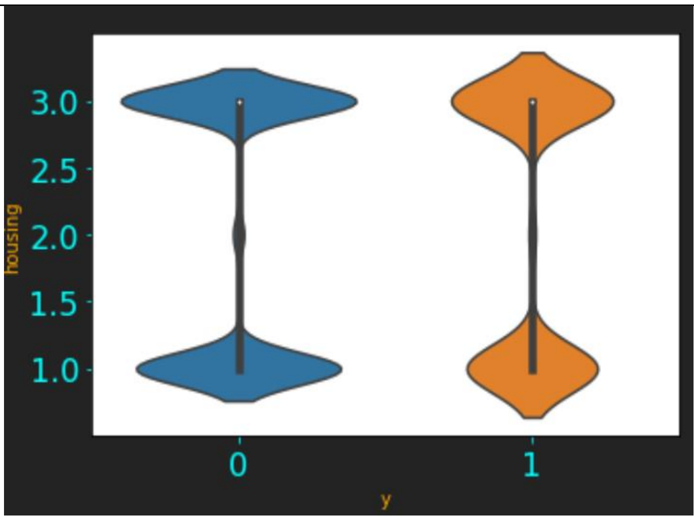
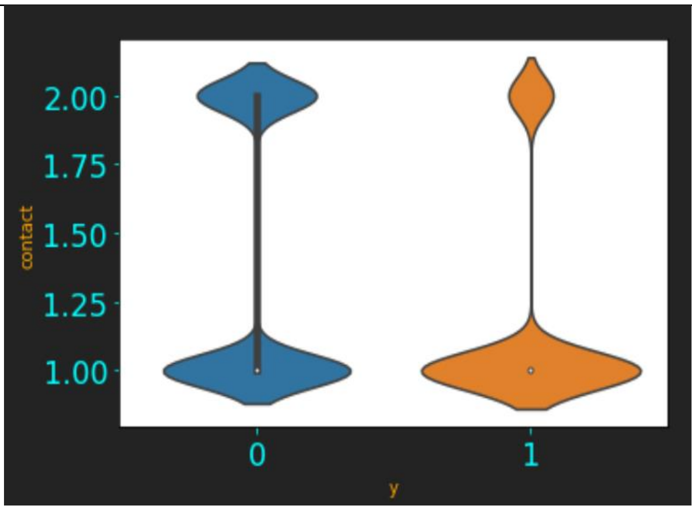
Relationship between the important features and target

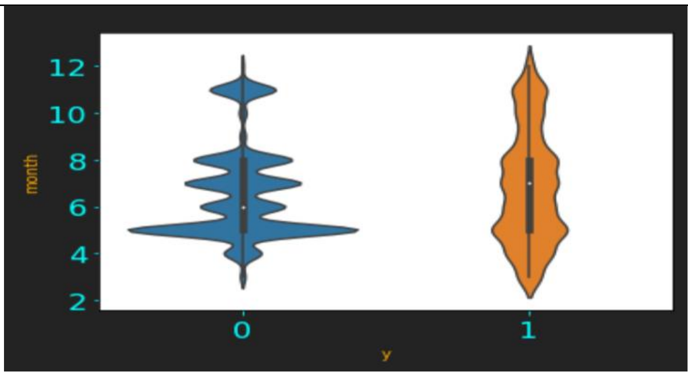
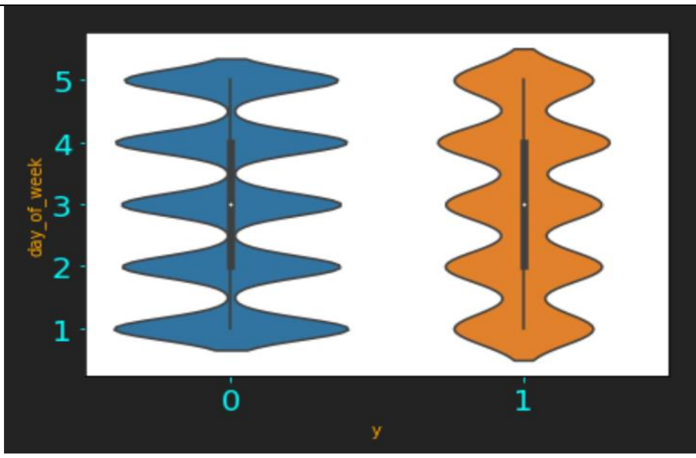
<table>
<tr>
<td>


</td>
<td>Most people are in thier higher 30s and those aged 60 and above are more likely to subscribe to term deposit</td>
</tr>
<tr>
<td>


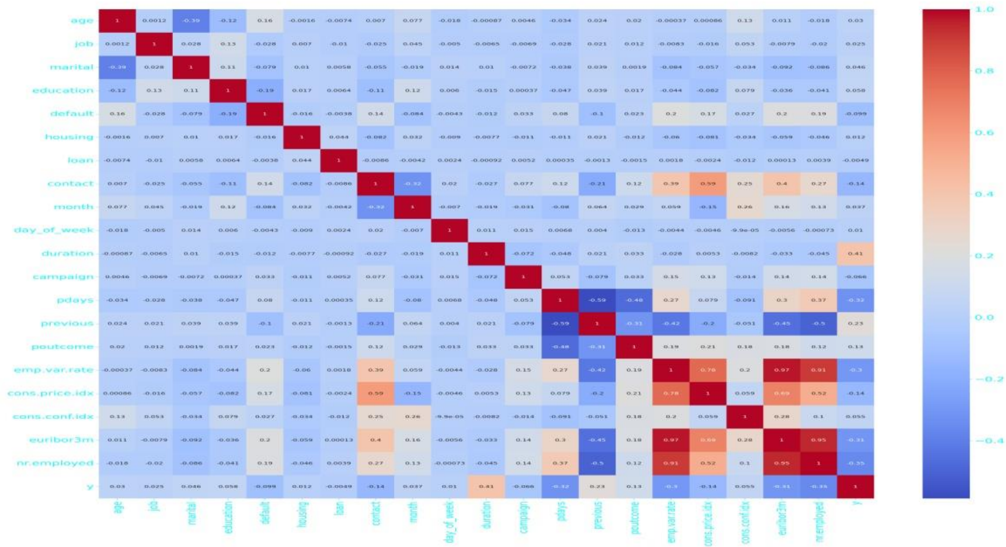
</td>
<td>Most people are either admins or in blue collar job and those who are retired or self empolyed are more likely to invest in term deposit</td>
</tr>
<tr>
<td>


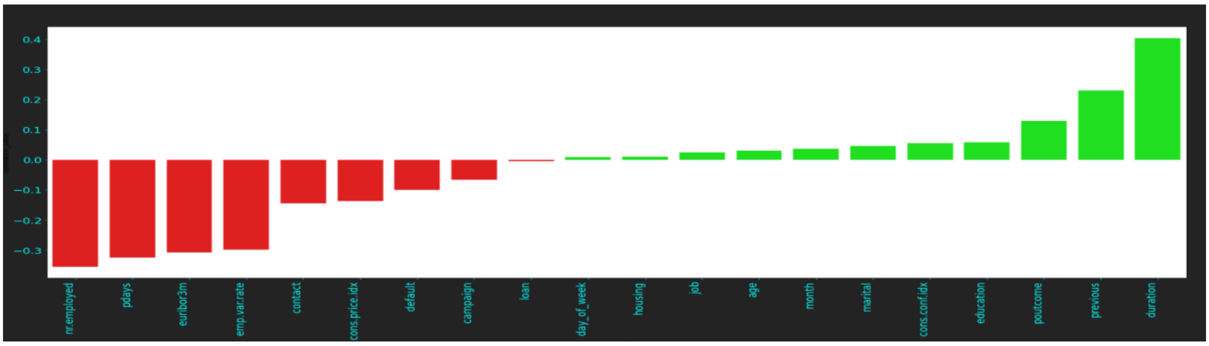
</td>
<td>Married individuals are more inclined towards not investing in term deposit</td>
</tr>
</table>

	Education of an individual doesnt seem have any influence on term deposit
	People with housing loan exhibit less interest in term deposit
	Those who received a call on their cellular phone are more likely to go for Term deposit

	In May most people didn't go for term deposit, while during Jan, Feb, Mar, Sep and Oct they did
	Among those who bought term deposit, Calls made on Monday and Friday were not much fruitful

Heatmap showing the relationship between features and target using df.corr () function





As seen above few features have negative correlation while others are positively correlated with the target. However, none of them show strong correlation.

Features with the negative correlation with the target

Features having negative correlation with the target

	parameter	correlation_value
0	nr.employed	-0.354678
1	pdays	-0.324914
2	euribor3m	-0.307771
3	emp.var.rate	-0.298334
4	contact	-0.144773
5	cons.price.idx	-0.136211
6	default	-0.099352
7	campaign	-0.066357
8	loan	-0.004909

Features with positive correlation on the target

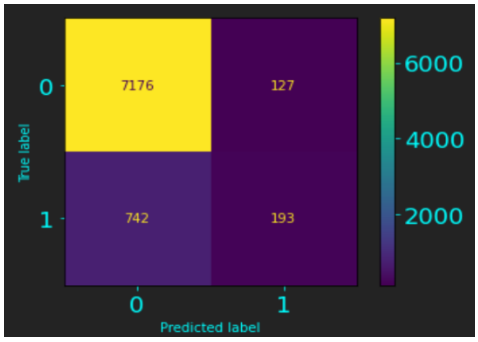
Features having positive correlation with the target

	parameter	correlation_value
0	day_of_week	0.010051
1	housing	0.011552
2	job	0.025122
3	age	0.030399
4	month	0.037187
5	marital	0.046203
6	cons.conf.idx	0.054878
7	education	0.057799
8	poutcome	0.129789
9	previous	0.230181
10	duration	0.405274

Before proceeding with feature scaling and hyper parameter tuning, let us get the F1 score of the original dataset to set a benchmark first. As seen below the original dataset with 80:20 split has an F1 score of 0.94 for not subscribing term deposit and 0.31 for subscribing term deposit with an overall accuracy of 0.89.

Classification report of original dataset with test size of 20%

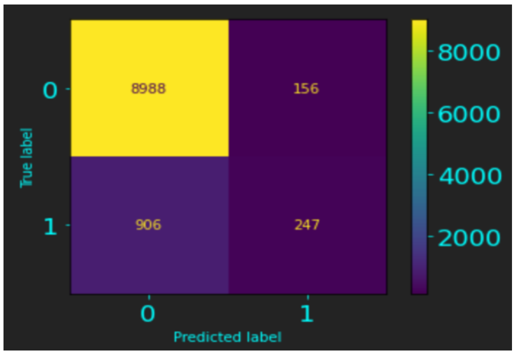
	precision	recall	f1-score	support
0	0.91	0.98	0.94	7303
1	0.60	0.21	0.31	935
accuracy			0.89	8238
macro avg	0.75	0.59	0.63	8238
weighted avg	0.87	0.89	0.87	8238



As seen below the **original dataset** with 75:25 split has an F1 score of 0.94 for not subscribing term deposit and 0.61 for subscribing term deposit with an overall accuracy of 0.90.

Classification report of original dataset with test size of 25%

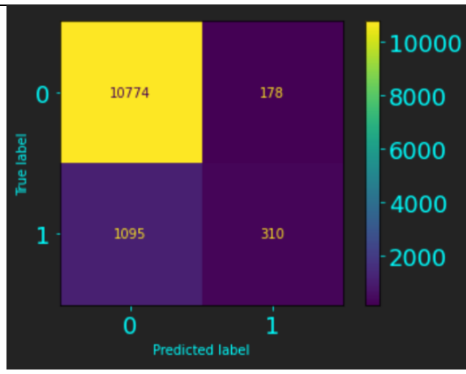
	precision	recall	f1-score	support
0	0.91	0.98	0.94	9144
1	0.61	0.21	0.32	1153
accuracy			0.90	10297
macro avg	0.76	0.60	0.63	10297
weighted avg	0.88	0.90	0.87	10297



As seen below the **original dataset** with 70:30 split has an F1 score of 0.94 for not subscribing term deposit and 0.64 for subscribing term deposit with an overall accuracy of 0.90.

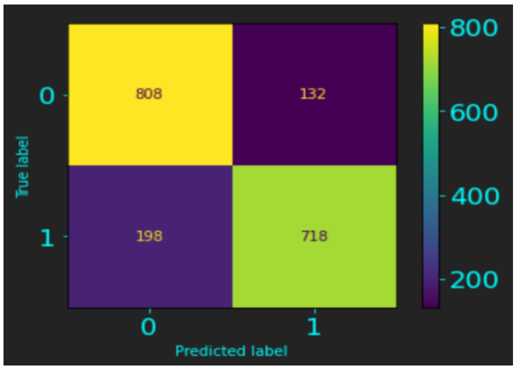
Classification report of original dataset with test size of 30%

	precision	recall	f1-score	support
0	0.91	0.98	0.94	10952
1	0.64	0.22	0.33	1405
accuracy			0.90	12357
macro avg	0.77	0.60	0.64	12357
weighted avg	0.88	0.90	0.87	12357



Classification report of Under-sampled dataset

	precision	recall	f1-score	support
0	0.80	0.86	0.83	940
1	0.84	0.78	0.81	916
accuracy			0.82	1856
macro avg	0.82	0.82	0.82	1856
weighted avg	0.82	0.82	0.82	1856

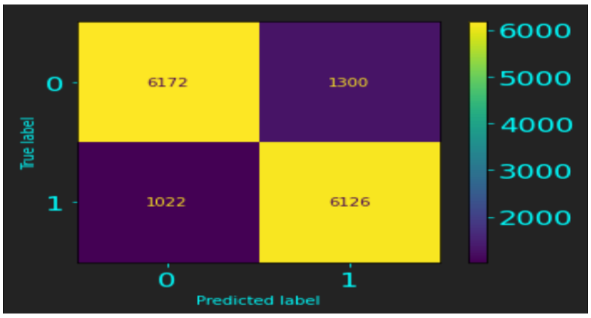


As seen above Under sampled dataset has an F1 score of 0.83 for not subscribing term deposit and 0.81 for subscribing term deposit with an overall accuracy of 0.82.

The over sampled dataset has an F1 score of 0.83 for not subscribing term deposit and 0.81 for subscribing term deposit with an overall accuracy of 0.82.

Classification report of Over-sampled dataset

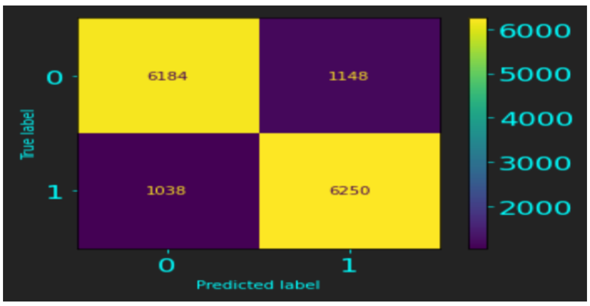
	precision	recall	f1-score	support
0	0.86	0.83	0.84	7472
1	0.82	0.86	0.84	7148
accuracy			0.84	14620
macro avg	0.84	0.84	0.84	14620
weighted avg	0.84	0.84	0.84	14620



SMOTE performed better than under-sampling and over-sampling ,

Classification report for the SMOTE dataset

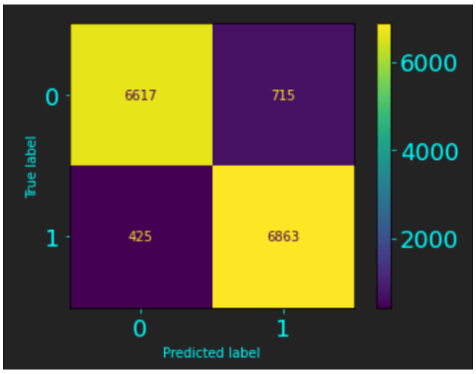
	precision	recall	f1-score	support
0	0.85	0.84	0.85	9137
1	0.84	0.86	0.85	9137
accuracy			0.85	18274
macro avg	0.85	0.85	0.85	18274
weighted avg	0.85	0.85	0.85	18274



Scaling is improving the F1-Score by about 7%,

Classification report of SMOTE-scaled dataset

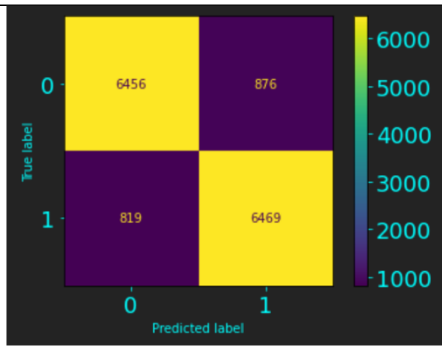
	precision	recall	f1-score	support
0	0.94	0.90	0.92	7332
1	0.91	0.94	0.92	7288
accuracy			0.92	14620
macro avg	0.92	0.92	0.92	14620
weighted avg	0.92	0.92	0.92	14620



Now using **Logistic Regression** with tuned hyper-parameters,

Classification report of SMOTE scaled dataset - Logistic Regression

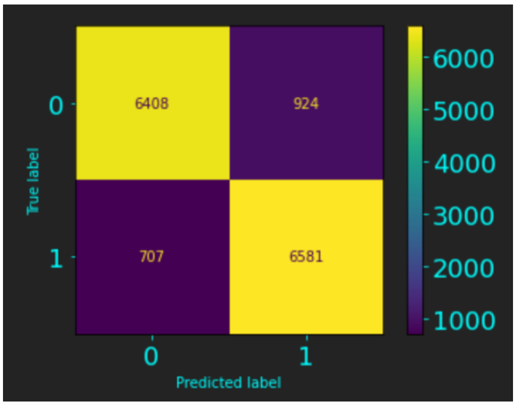
	precision	recall	f1-score	support
0	0.89	0.88	0.88	7332
1	0.88	0.89	0.88	7288
accuracy			0.88	14620
macro avg	0.88	0.88	0.88	14620
weighted avg	0.88	0.88	0.88	14620



Now using **Support Vector Machine Classifier** with tuned hyper-parameters,

Classification report of SMOTE scaled dataset - Support Vector Classifier

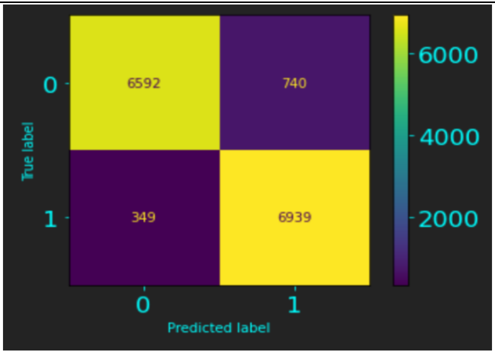
	precision	recall	f1-score	support
0	0.90	0.87	0.89	7332
1	0.88	0.90	0.89	7288
accuracy			0.89	14620
macro avg	0.89	0.89	0.89	14620
weighted avg	0.89	0.89	0.89	14620



Now using KNN Classifier with tuned hyper-parameters,

Classification report of SMOTE scaled dataset - K Nearest Neighbor

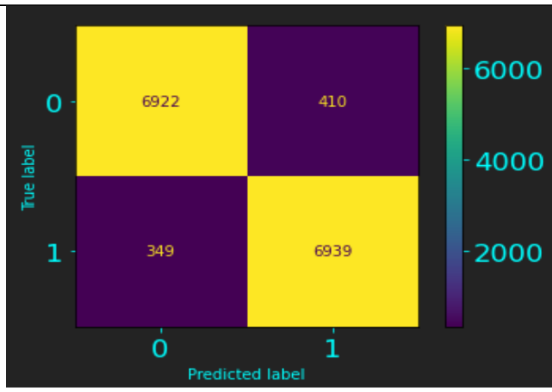
	precision	recall	f1-score	support
0	0.95	0.90	0.92	7332
1	0.90	0.95	0.93	7288
accuracy			0.93	14620
macro avg	0.93	0.93	0.93	14620
weighted avg	0.93	0.93	0.93	14620



Now using Bagging Classifier with tuned hyper-parameters,

Classification report of SMOTE scaled dataset - Bagging Classifier

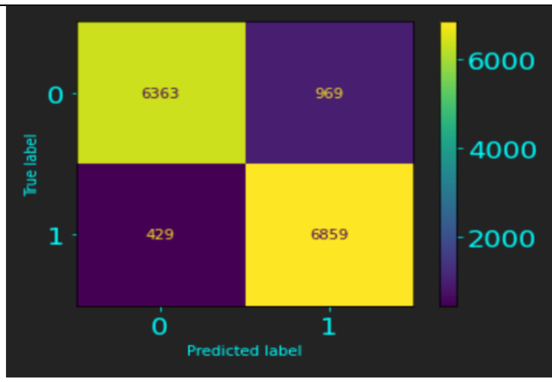
	precision	recall	f1-score	support
0	0.95	0.95	0.95	7332
1	0.95	0.95	0.95	7288
accuracy			0.95	14620
macro avg	0.95	0.95	0.95	14620
weighted avg	0.95	0.95	0.95	14620



Now using Gradient boosting Classifier with tuned hyper-parameters,

Classification report of SMOTE scaled dataset - Gradient Boosting Classifier

	precision	recall	f1-score	support
0	0.94	0.87	0.90	7332
1	0.88	0.94	0.91	7288
accuracy			0.90	14620
macro avg	0.91	0.90	0.90	14620
weighted avg	0.91	0.90	0.90	14620



Now using Random Forest Classifier with tuned hyper-parameters,

```
Classification report of SMOTE scaled dataset - Random Forest Classifier
              precision    recall  f1-score   support

           0       0.92      0.82      0.87      7332
           1       0.84      0.92      0.88      7288

    accuracy                           0.87     14620
   macro avg       0.88      0.87      0.87     14620
weighted avg       0.88      0.87      0.87     14620
```

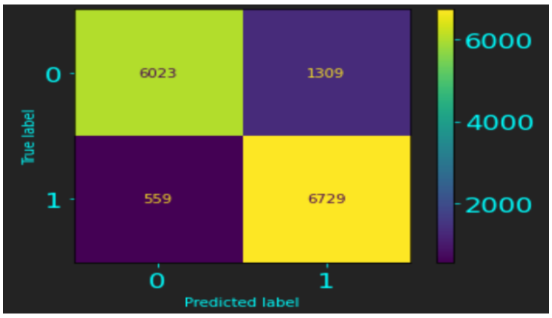


Chart showing the performance of various models on the dataset

	Dataset/Algo	False_Positive	False_Negative	Accuracy	Precision	Recall	Specificity	F1_Score
0	SMOTE Scaled dataset - Bagging Classifier	410	349	0.95	0.94	0.95	0.94	0.94
1	SMOTE Scaled dataset - K Nearest Neighbor	740	349	0.93	0.90	0.95	0.90	0.92
2	SMOTE-Scaled dataset	715	425	0.92	0.91	0.94	0.90	0.92
3	SMOTE Scaled dataset - Gradient Boosting Class...	969	429	0.90	0.88	0.94	0.87	0.91
4	SMOTE Scaled dataset - Support Vector Classifier	924	707	0.89	0.88	0.90	0.87	0.89
5	SMOTE Scaled dataset - Logistic Regression	876	819	0.88	0.88	0.89	0.88	0.88
6	SMOTE Scaled dataset - Random Forest Classifier	1309	559	0.87	0.84	0.92	0.82	0.88
7	Original dataset after SMOTE	1148	1038	0.85	0.84	0.86	0.84	0.85
8	Over-sampled dataset	1300	1022	0.84	0.82	0.86	0.83	0.84
9	Under-sampled dataset	132	198	0.82	0.84	0.78	0.86	0.81
10	Original dataset with 70:30 split	178	1095	0.90	0.64	0.22	0.98	0.33
11	Original dataset with 75:25 split	156	906	0.90	0.61	0.21	0.98	0.31
12	Original dataset with 80:20 split	127	742	0.89	0.60	0.21	0.98	0.31

Considering only Support Vector Classifier model

Support Vector Classifier with SMOTE scaled dataset with default hyper parameters performed better than any other model with SVC.

	Dataset/Algo	False_Positive	False_Negative	Accuracy	Precision	Recall	Specificity	F1_Score
0	SMOTE-Scaled dataset	715	425	0.92	0.91	0.94	0.90	0.92
1	SMOTE Scaled dataset - Support Vector Classifier	924	707	0.89	0.88	0.90	0.87	0.89
2	Original dataset after SMOTE	1148	1038	0.85	0.84	0.86	0.84	0.85
3	Over-sampled dataset	1300	1022	0.84	0.82	0.86	0.83	0.84
4	Under-sampled dataset	132	198	0.82	0.84	0.78	0.86	0.81
5	Original dataset with 70:30 split	178	1095	0.90	0.64	0.22	0.98	0.33
6	Original dataset with 75:25 split	156	906	0.90	0.61	0.21	0.98	0.31
7	Original dataset with 80:20 split	127	742	0.89	0.60	0.21	0.98	0.31

The verdict: Bagging classifier outperformed every other model including SVC, with a F1 Score of 0.94 and an accuracy of 0.95. Since, it is a problem statement pertaining to banking sector False Negative is more important than False Positive because you don't want to miss out on a potential customer owing to a bad prediction. Here too bagging classifier does better than any other model with least False Negatives.