Capstone Project – 3

Bank Marketing Effectiveness Prediction

Presentation Outline :

- Business Objective
- Exploring the Dataset
- Exploratory Data Analysis
- Data Cleaning and Oversampling
- Machine Learning Model Evaluation: Classification
- Model Evaluation Results
- Feature Importance
- Proposed Solution
- Conclusion

Business Objective

AI

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') 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)

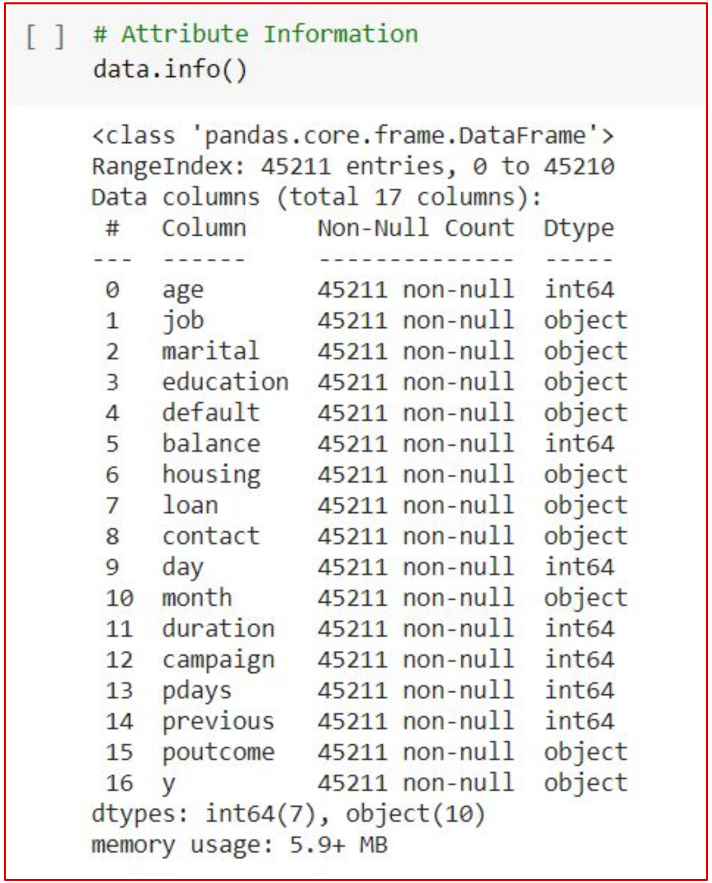
Numerical Features

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

Desired target

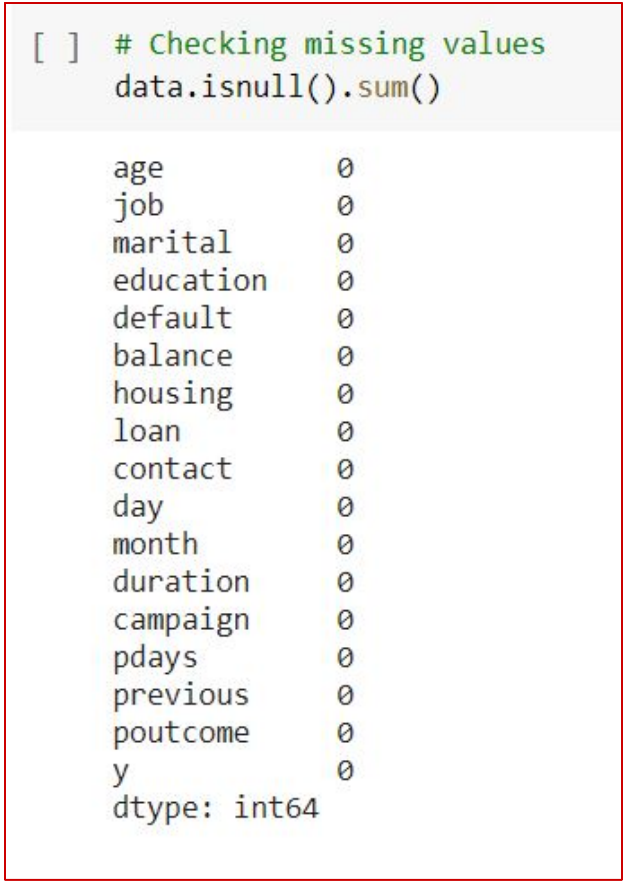
- y - has the client subscribed a term deposit? (binary: 'yes','no')

Attribute Information : Dtype & Null values



```
[ ]  # Attribute Information
     data.info()
```

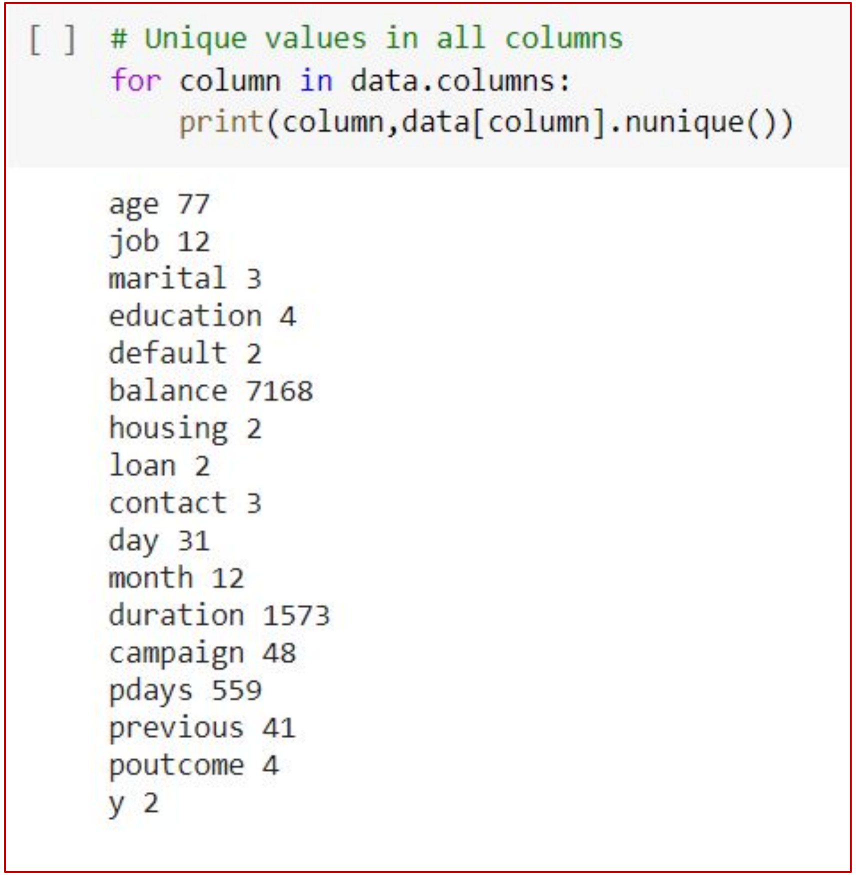
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   age        45211 non-null  int64 
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64 
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64 
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64 
 12  campaign   45211 non-null  int64 
 13  pdays      45211 non-null  int64 
 14  previous   45211 non-null  int64 
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```



```
[ ]  # Checking missing values
     data.isnull().sum()
```

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
dtype: int64
```

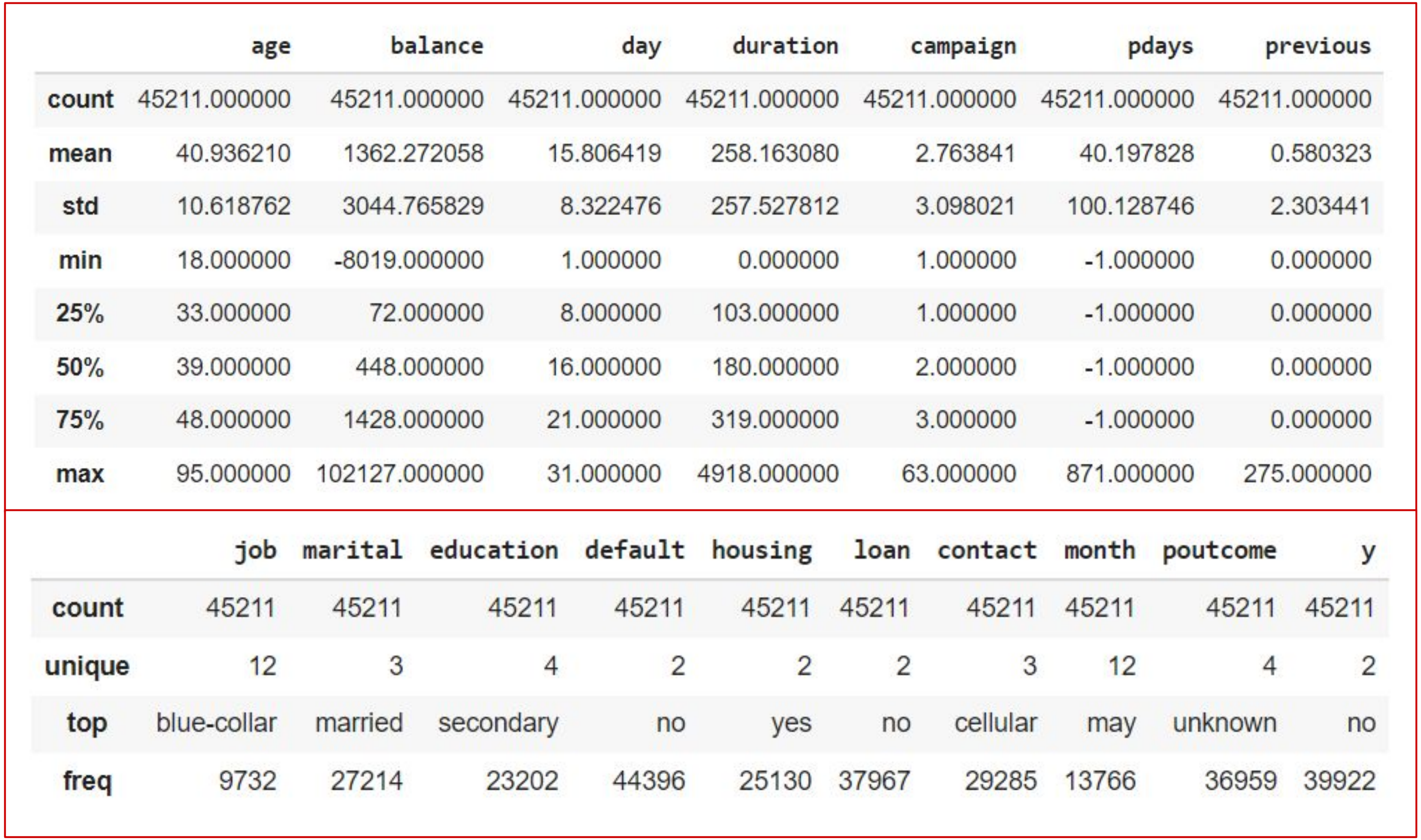
Attribute Information : Unique Values



```
# Unique values in all columns
for column in data.columns:
    print(column,data[column].nunique())
```

```
age 77
job 12
marital 3
education 4
default 2
balance 7168
housing 2
loan 2
contact 3
day 31
month 12
duration 1573
campaign 48
pdays 559
previous 41
poutcome 4
y 2
```

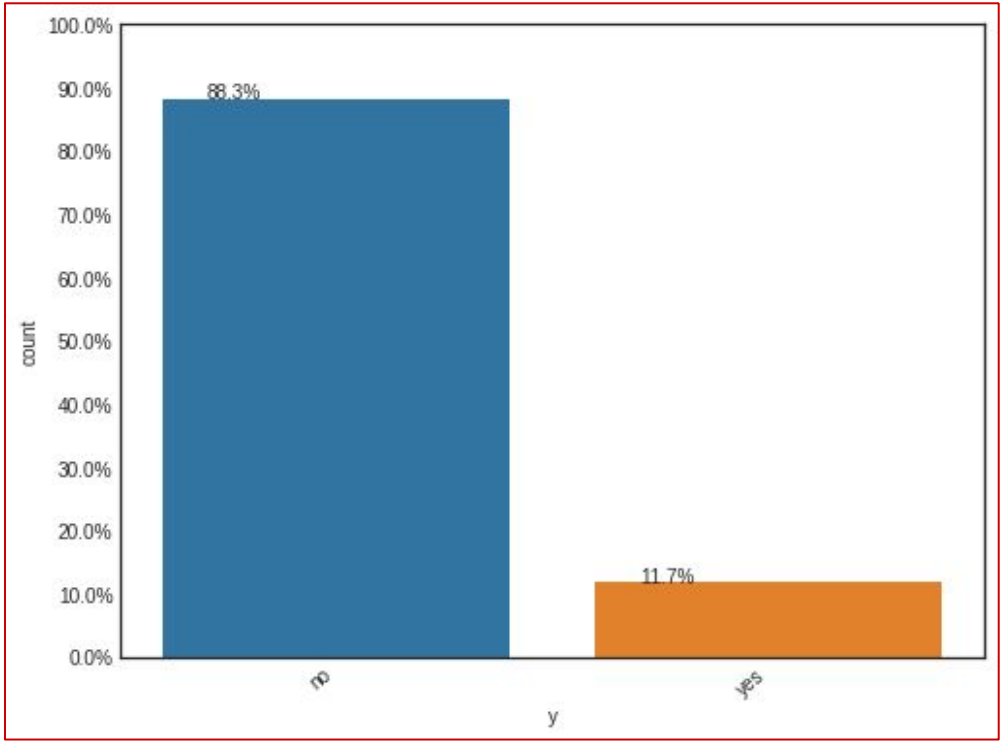
Descriptive Stats :



	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

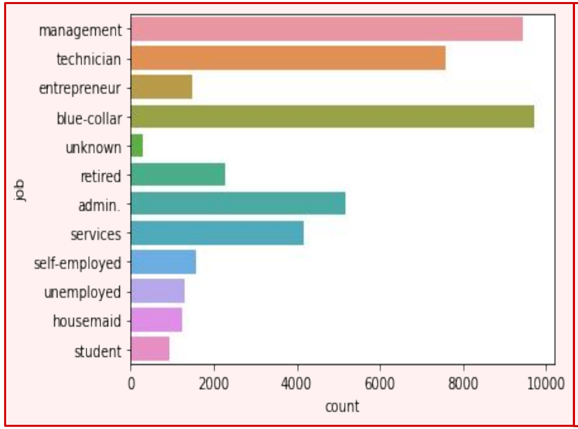
	job	marital	education	default	housing	loan	contact	month	poutcome	y
count	45211	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	12	3	4	2	2	2	3	12	4	2
top	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	13766	36959	39922

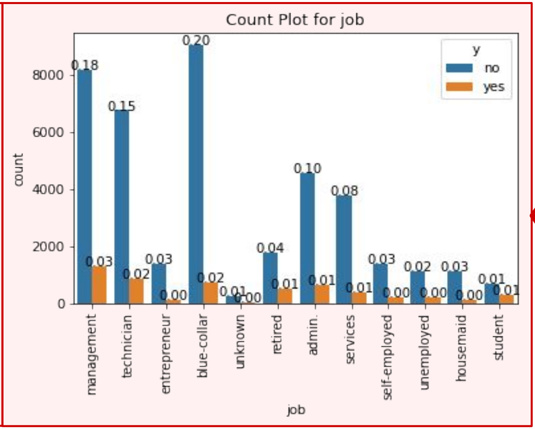
EDA : Defining the Target



- The target variable tells us the outcome of the campaign whether they went ahead for the term deposit or not.
- From this data we can see that 88% customers did not subscribed for Term deposit
- We can say that the percentage of people subscribing to the term deposit is quite low, thus creating an imbalance in the data.

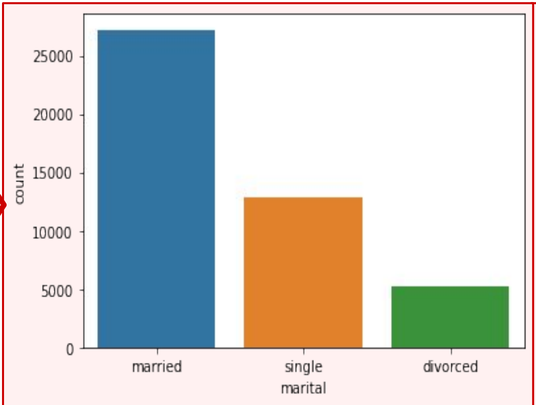
EDA(continued) :

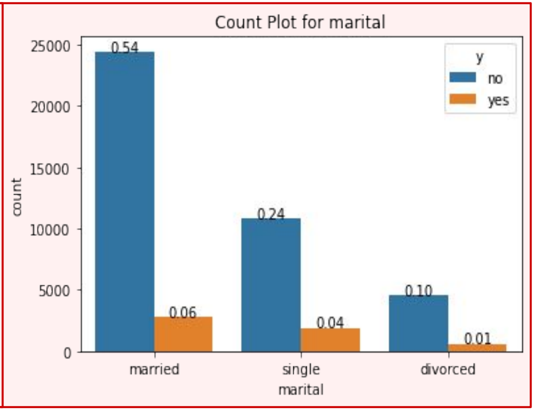




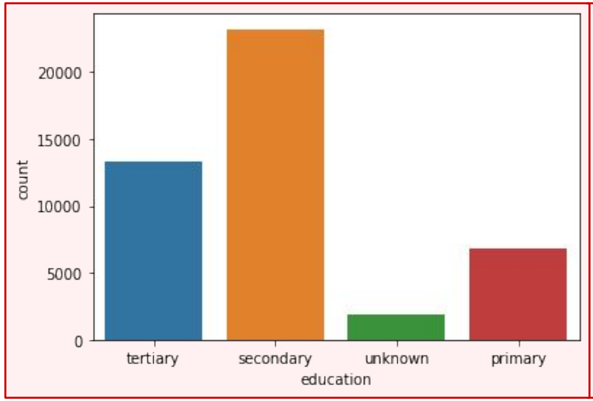
- Most of the customers have jobs as "management", "blue-collar" and "technician“.
- People with management jobs have subscribed more for the deposits.

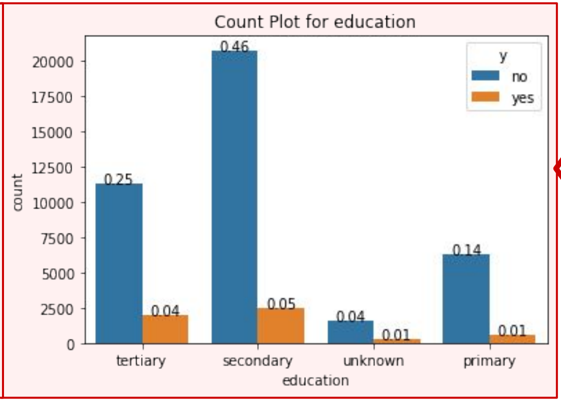
- Client who married are high in records.
- People who are married have subscribed for deposits more than people with any other marital status.





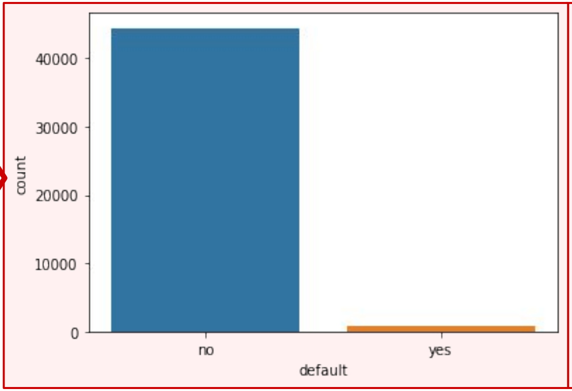
EDA(continued) :

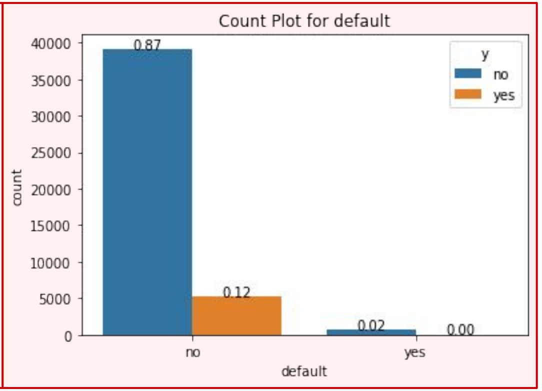




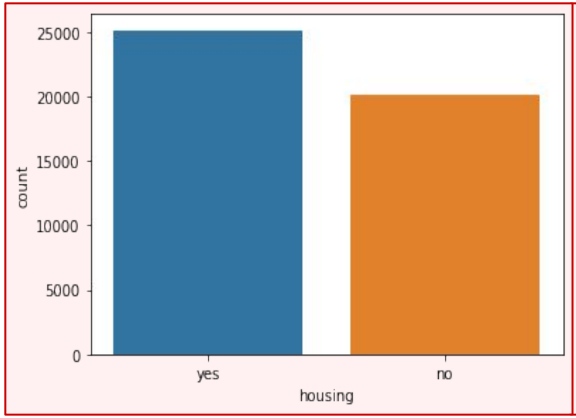
- Client whose education background is secondary are in high numbers.
- People with Secondary education qualification are the most who have subscribed for the deposits.

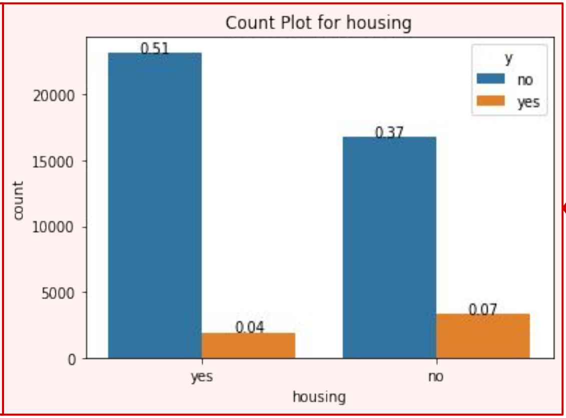
- default feature seems to be does not play important role.
- People with default status as no are the most ones who have and have not subscribed for bank deposits.





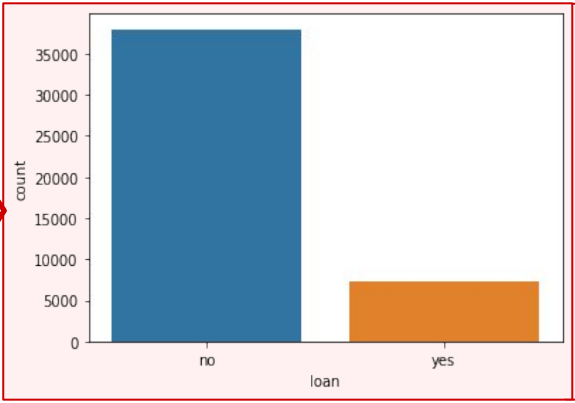
EDA(continued) :

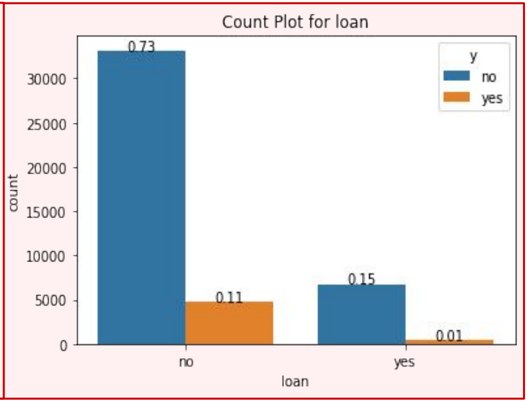




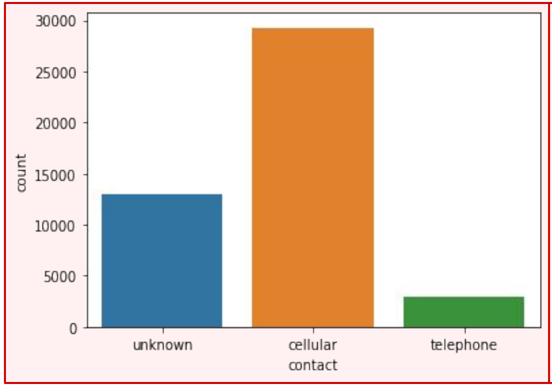
- People with housing loan are the most ones who have been contacted by the bank.
- People with no housing loan are the most ones who have subscribed for deposits.

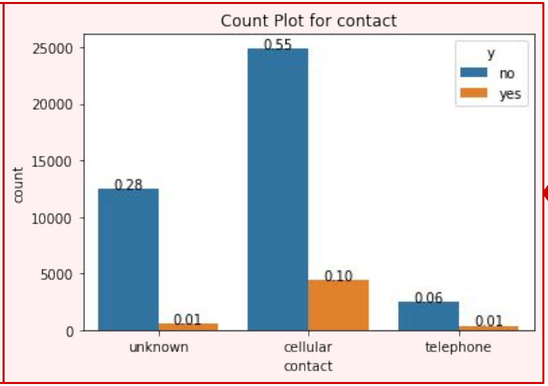
- People with no personal loan are the most ones who have been contacted by the bank for the deposits.
- People with no personal loan are the most ones who have not subscribed and are also the most ones who have subscribed for the deposits.





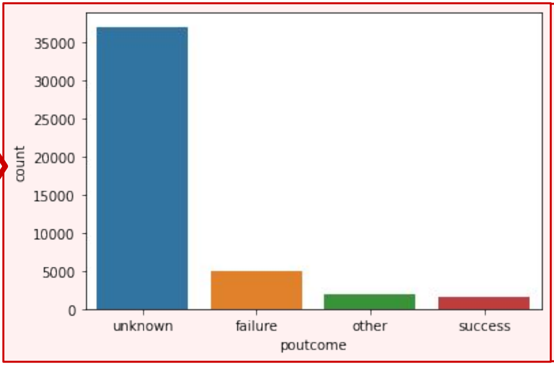
EDA(continued) :

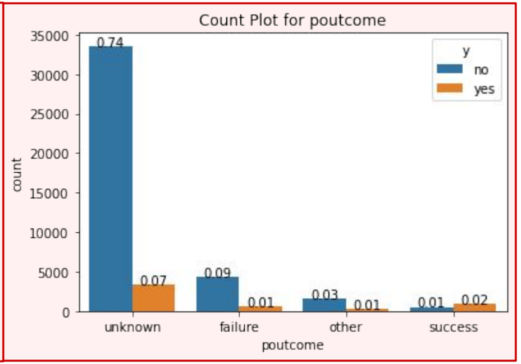




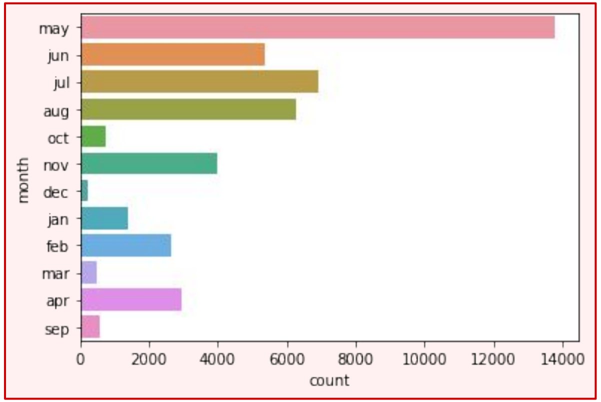
- Most people are contacted more in cellular than telephone.
- More people contacted on cellular by bank have subscribed the deposits.

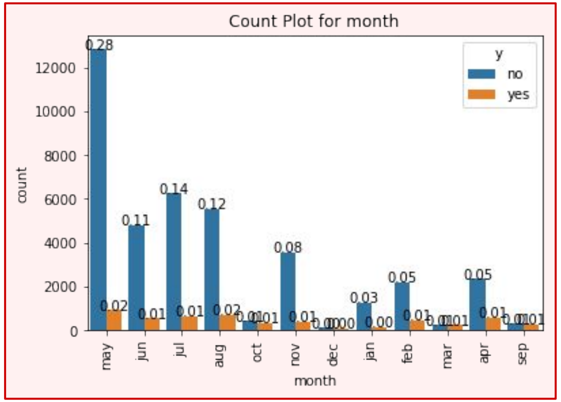
- Majority of the outcome of the previous campaign is Non-Existent.
- People whose previous outcome is non-existent have actually subscribed more.





EDA(continued) :

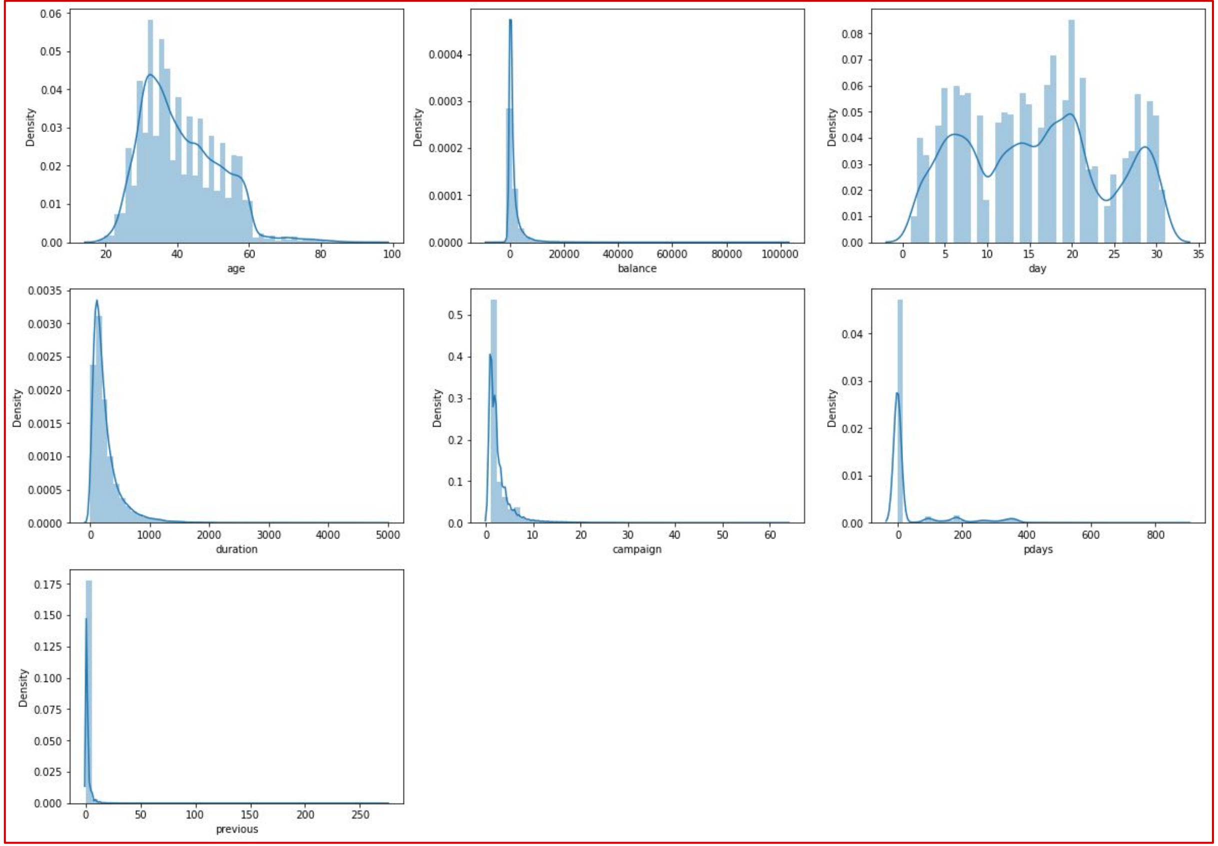




- Data in month of may is high and less in Dec

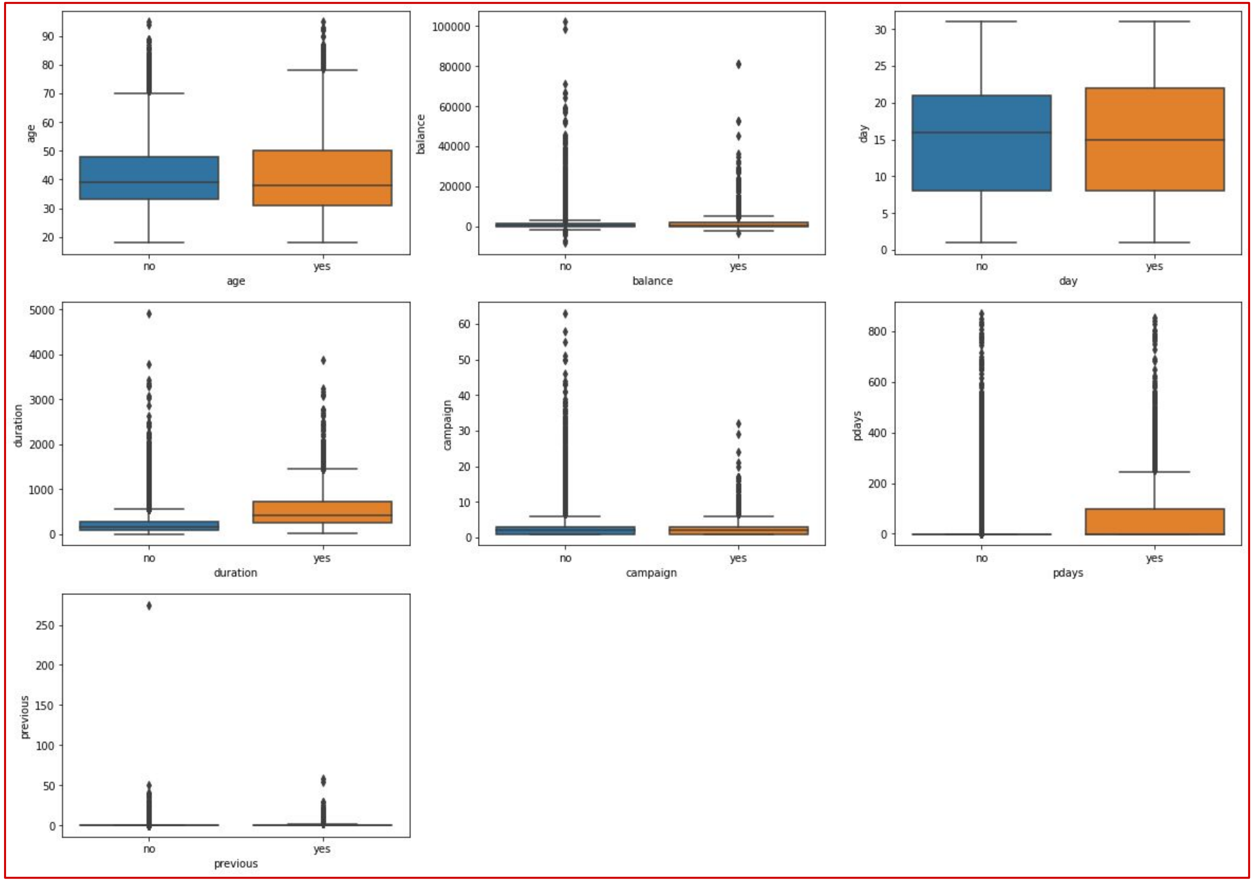
- The month of the highest level of marketing activity was the month of May.

EDA(continued) :



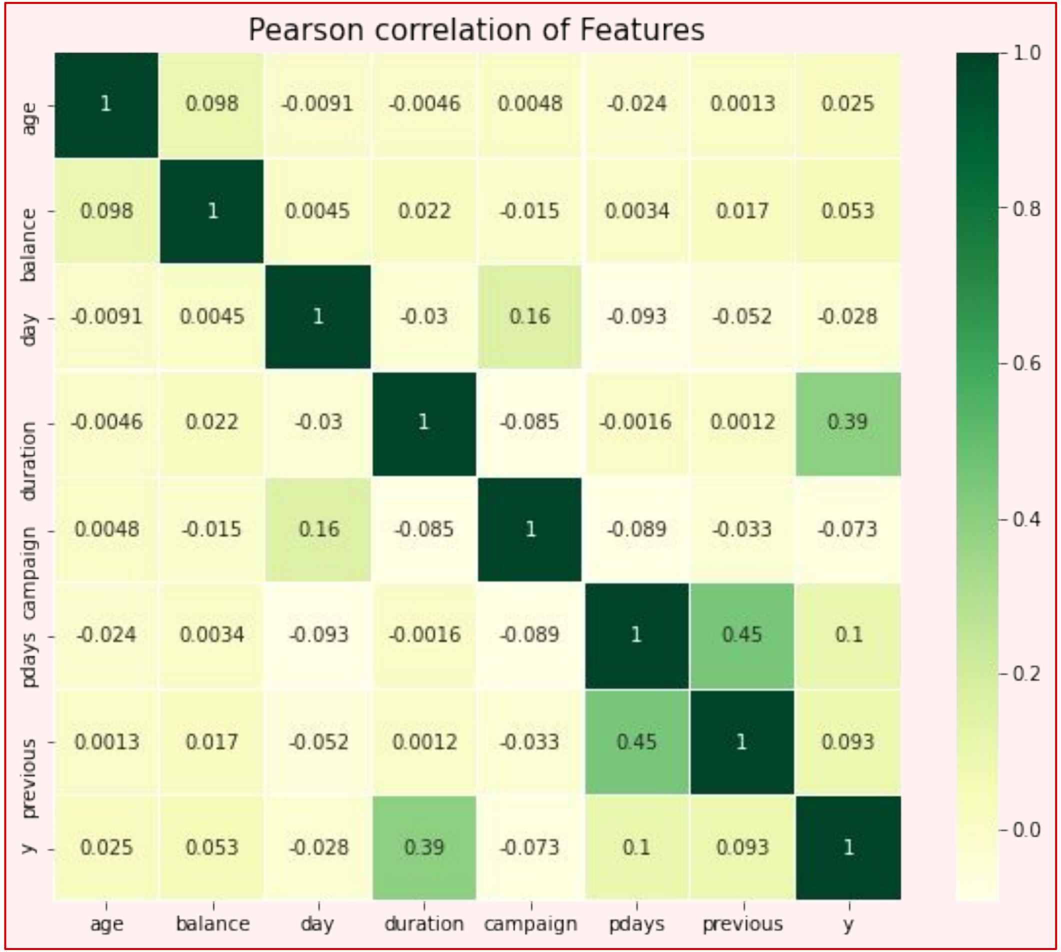
- It seems age, days distributed normally
- balance, duration, campaign, pdays and previous heavily skewed towards left and seems to be have some outliers.
- Most of the customers are in the age range of 30-40.
- for any class labels, previous with greater than 4 are outliers

EDA(continued) :



- Client shows interest on deposit who had discussion for longer duration
- Both the customers that subscribed or didn't subscribe a term deposit, has a median age of around 38-40
- Very people has been contacted by the bank and number of days passed for previous campaign is between 0–100
- average subscription rate is below 50% if the number of contacts during the campaign exceeds 4

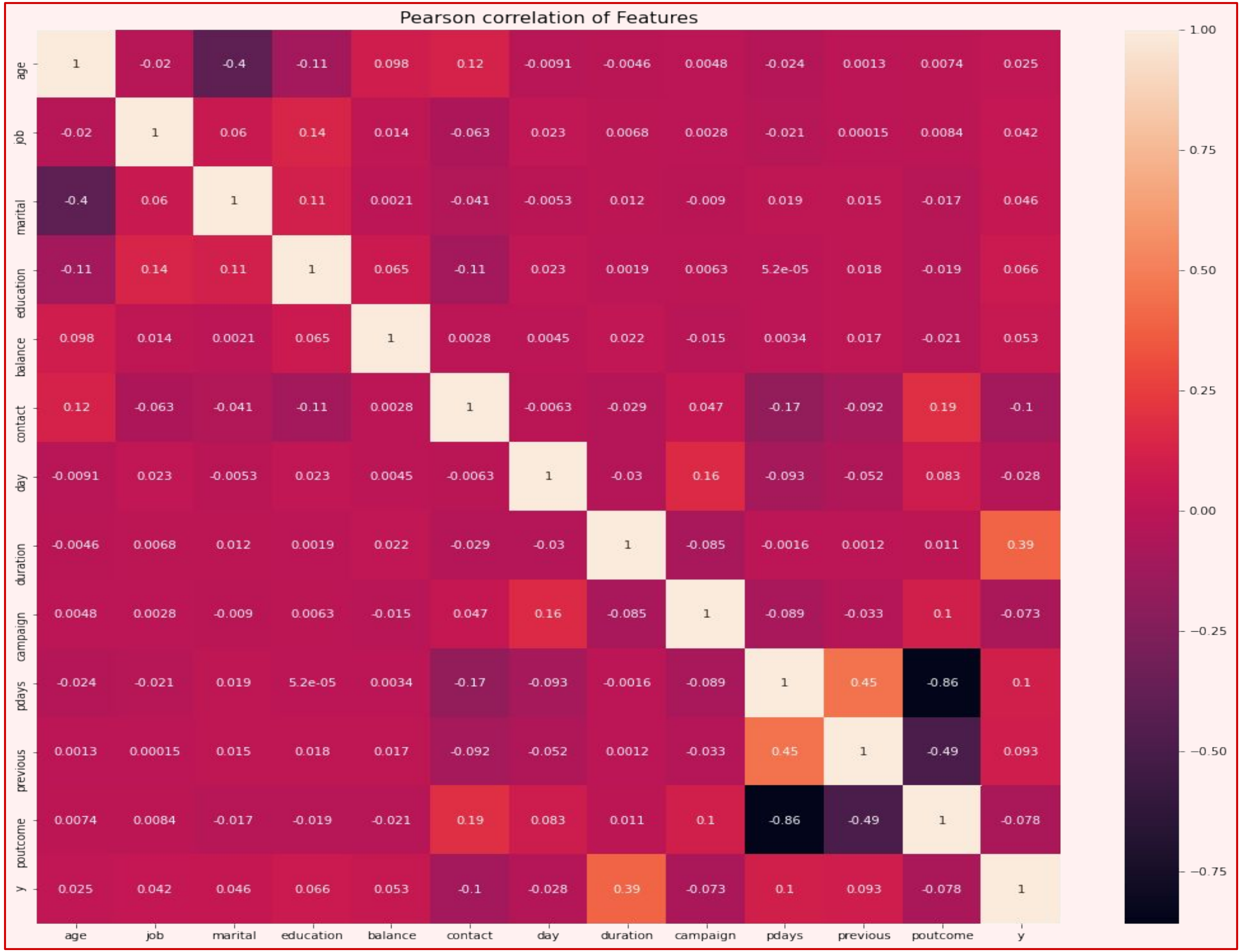
Correlation Heatmap :



Pearson correlation of Features

	age	balance	day	duration	campaign	pdays	previous	y
age	1	0.098	-0.0091	-0.0046	0.0048	-0.024	0.0013	0.025
balance	0.098	1	0.0045	0.022	-0.015	0.0034	0.017	0.053
day	-0.0091	0.0045	1	-0.03	0.16	-0.093	-0.052	-0.028
duration	-0.0046	0.022	-0.03	1	-0.085	-0.0016	0.0012	0.39
campaign	0.0048	-0.015	0.16	-0.085	1	-0.089	-0.033	-0.073
pdays	-0.024	0.0034	-0.093	-0.0016	-0.089	1	0.45	0.1
previous	0.0013	0.017	-0.052	0.0012	-0.033	0.45	1	0.093
y	0.025	0.053	-0.028	0.39	-0.073	0.1	0.093	1

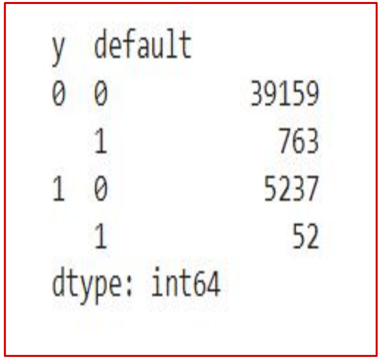
Correlation Heatmap (Contd.) :



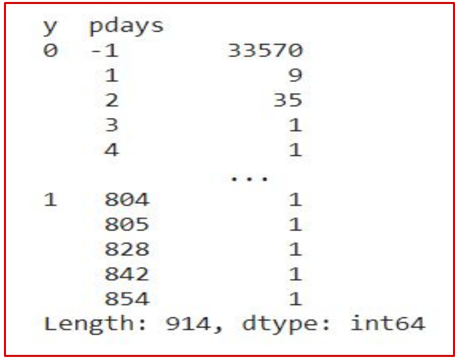
Pearson correlation of Features

	age	job	marital	education	balance	contact	day	duration	campaign	pdays	previous	poutcome	y
age	1	-0.02	-0.4	-0.11	0.098	0.12	-0.0091	-0.0046	0.0048	-0.024	0.0013	0.0074	0.025
job	-0.02	1	0.06	0.14	0.014	-0.063	0.023	0.0068	0.0028	-0.021	0.00015	0.0084	0.042
marital	-0.4	0.06	1	0.11	0.0021	-0.041	-0.0053	0.012	-0.009	0.019	0.015	-0.017	0.046
education	-0.11	0.14	0.11	1	0.065	-0.11	0.023	0.0019	0.0063	5.2e-05	0.018	-0.019	0.066
balance	0.098	0.014	0.0021	0.065	1	0.0028	0.0045	0.022	-0.015	0.0034	0.017	-0.021	0.053
contact	0.12	-0.063	-0.041	-0.11	0.0028	1	-0.0063	-0.029	0.047	-0.17	-0.092	0.19	-0.1
day	-0.0091	0.023	-0.0053	0.023	0.0045	-0.0063	1	-0.03	0.16	-0.093	-0.052	0.083	-0.028
duration	-0.0046	0.0068	0.012	0.0019	0.022	-0.029	-0.03	1	-0.085	-0.0016	0.0012	0.011	0.39
campaign	0.0048	0.0028	-0.009	0.0063	-0.015	0.047	0.16	-0.085	1	-0.089	-0.033	0.1	-0.073
pdays	-0.024	-0.021	0.019	5.2e-05	0.0034	-0.17	-0.093	-0.0016	-0.089	1	0.45	-0.86	0.1
previous	0.0013	0.00015	0.015	0.018	0.017	-0.092	-0.052	0.0012	-0.033	0.45	1	-0.49	0.093
poutcome	0.0074	0.0084	-0.017	-0.019	-0.021	0.19	0.083	0.011	0.1	-0.86	-0.49	1	-0.078
y	0.025	0.042	0.046	0.066	0.053	-0.1	-0.028	0.39	-0.073	0.1	0.093	-0.078	1

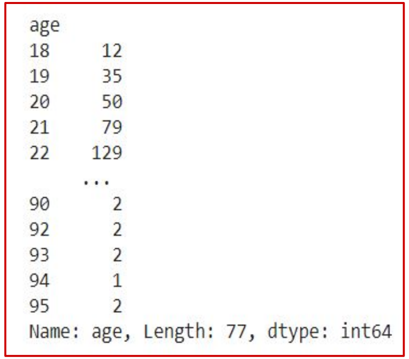
Data Cleaning :



```
y  default
0  0          39159
   1            763
1  0           5237
   1             52
dtype: int64
```

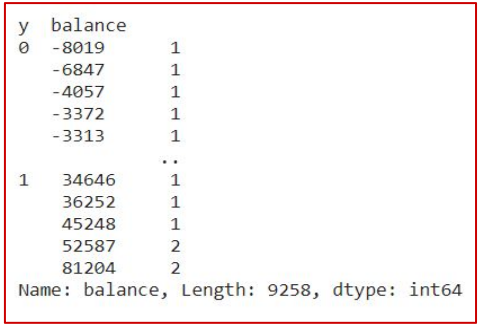


```
y   pdays
0   -1         33570
     1             9
     2            35
     3             1
     4             1
                ...
1    804           1
     805           1
     828           1
     842           1
     854           1
Length: 914, dtype: int64
```

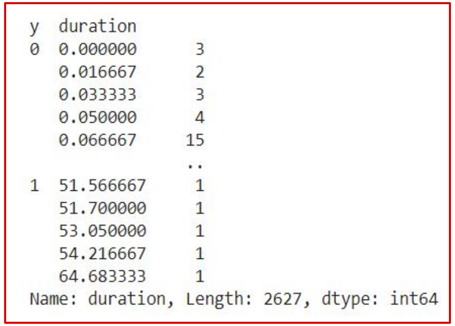


```
age
18      12
19      35
20      50
21      79
22     129
      ...
90       2
92       2
93       2
94       1
95       2
Name: age, Length: 77, dtype: int64
```

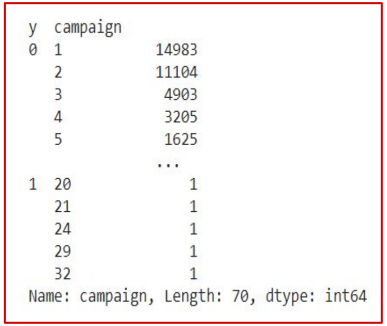
- default features does not play imp role
- drop pdays as it has -1 value for around 75%+ .
- Age can be ignored and values lies in between 18 to 95
- these outlier should not be remove as balance goes high, client show interest on deposit
- these outlier should not be remove as duration goes high, client show interest on deposit
- remove outliers in feature campaign
- remove outliers in feature previous.



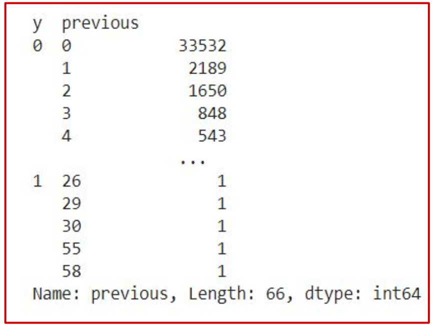
```
y  balance
0  -8019      1
   -6847      1
   -4057      1
   -3372      1
   -3313      1
             ..
1   34646     1
    36252     1
    45248     1
    52587     2
    81204     2
Name: balance, Length: 9258, dtype: int64
```



```
y  duration
0  0.000000     3
   0.016667     2
   0.033333     3
   0.050000     4
   0.066667    15
               ..
1  51.566667    1
   51.700000    1
   53.050000    1
   54.216667    1
   64.683333    1
Name: duration, Length: 2627, dtype: int64
```



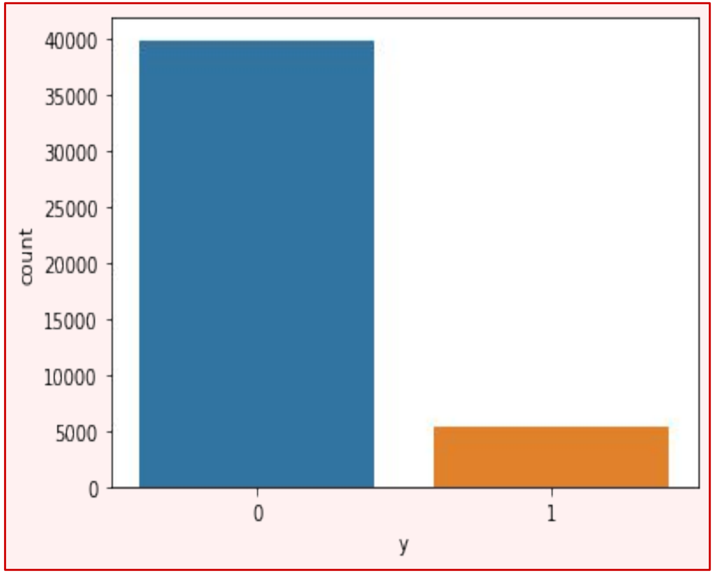
```
y  campaign
0  1          14983
   2          11104
   3           4903
   4           3205
   5           1625
               ...
1  20             1
   21             1
   24             1
   29             1
   32             1
Name: campaign, Length: 70, dtype: int64
```



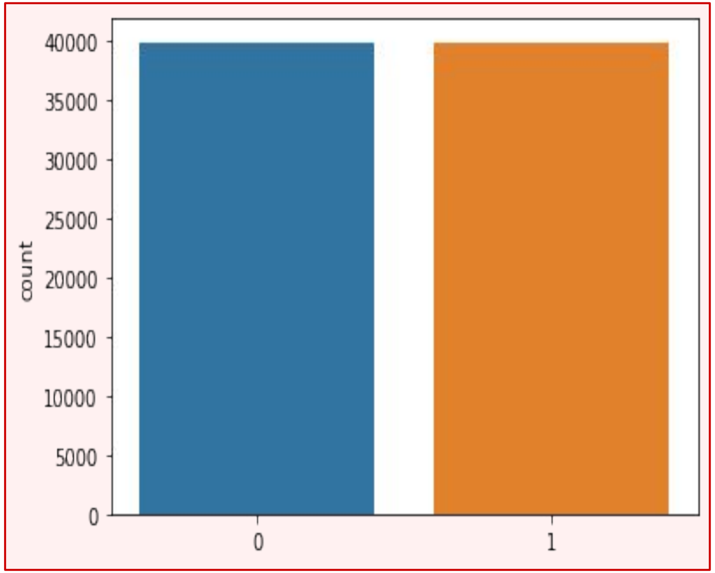
```
y  previous
0  0          33532
   1           2189
   2           1650
   3            848
   4            543
               ...
1  26             1
   29             1
   30             1
   55             1
   58             1
Name: previous, Length: 66, dtype: int64
```

Target Feature : Random Oversampling

Before

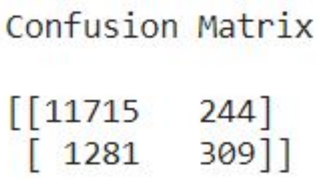


After

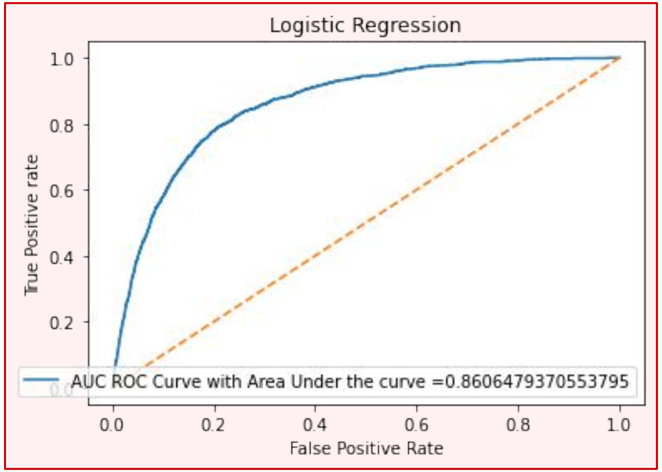


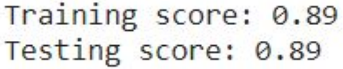
ML Model Evaluation : Logistic Regression

Before



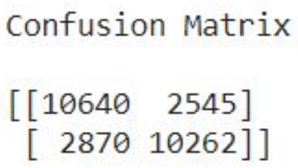
```
Confusion Matrix

[[11715   244]
 [ 1281   309]]
```



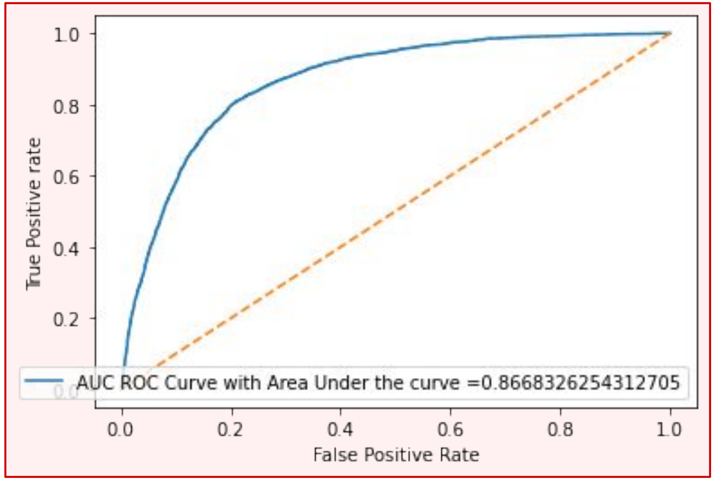


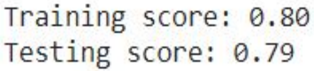
```
Training score: 0.89
Testing score: 0.89
```

After



```
Confusion Matrix

[[10640  2545]
 [ 2870 10262]]
```



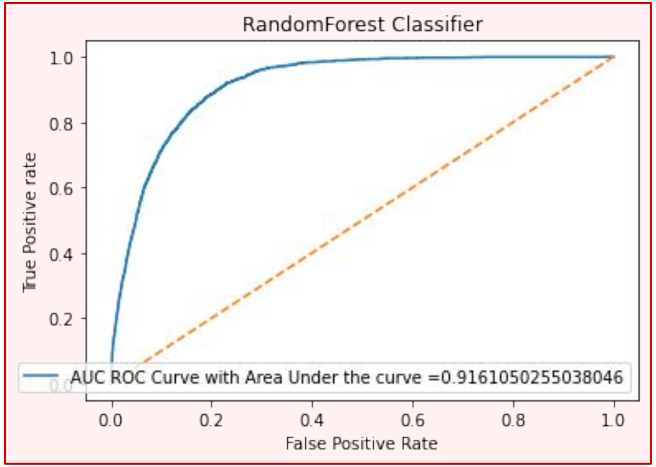


```
Training score: 0.80
Testing score: 0.79
```

ML Model Evaluation : Random Forests

Before

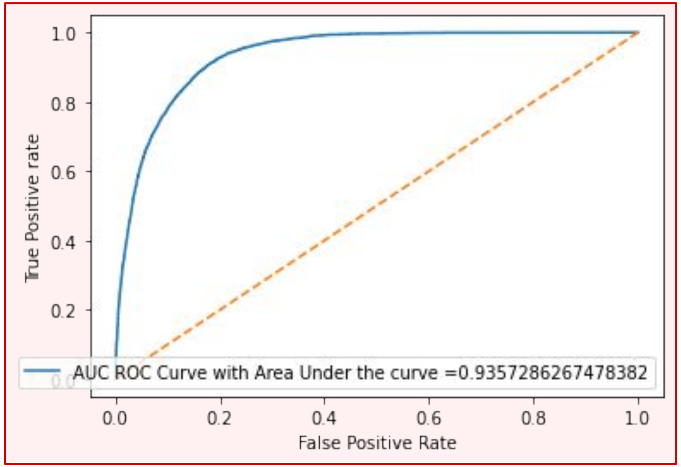
```
Confusion Matrix

[[11874    85]
 [ 1342   248]]
```



```
Training score: 0.90
Testing score: 0.89
```

After

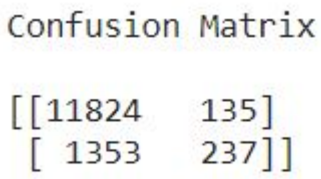
```
Confusion Matrix

[[10713  2472]
 [ 1261 11871]]
```



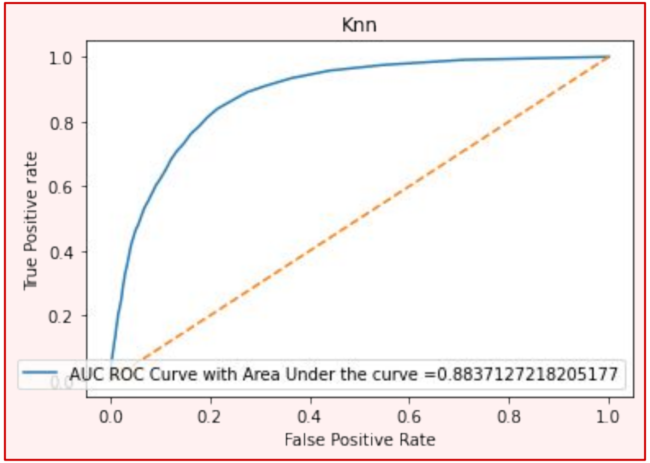
```
Training score: 0.86
Testing score: 0.86
```

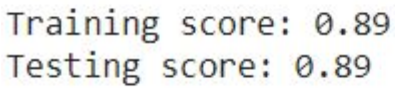
ML Model Evaluation : k-Nearest Neighbors

Before



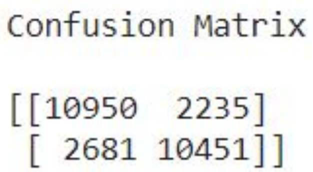
```
Confusion Matrix

[[11824   135]
 [ 1353   237]]
```



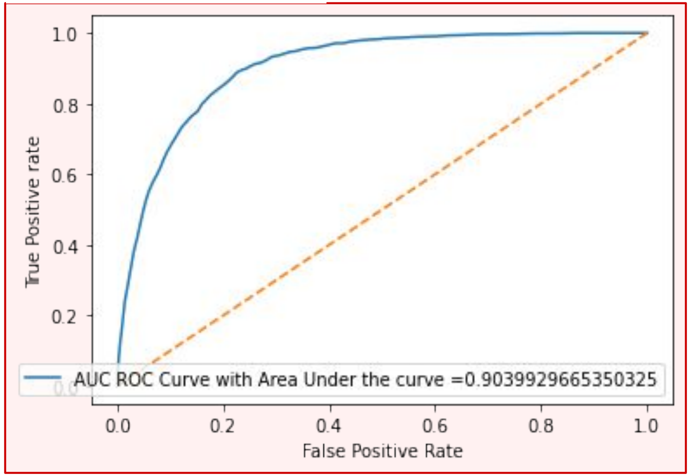


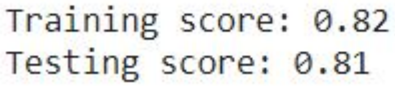
```
Training score: 0.89
Testing score: 0.89
```

After



```
Confusion Matrix

[[10950  2235]
 [ 2681 10451]]
```





```
Training score: 0.82
Testing score: 0.81
```

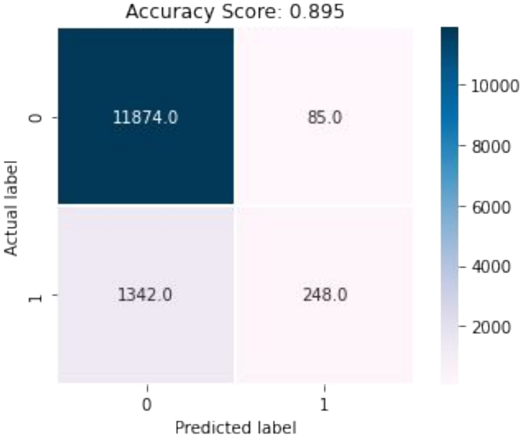
ML Model Evaluation Result :

Among this 3 model we can see that Accuracy and AUC_ROC score are almost high for every model so we had compare models with F1 Score and found that Random Forest is clear winner.

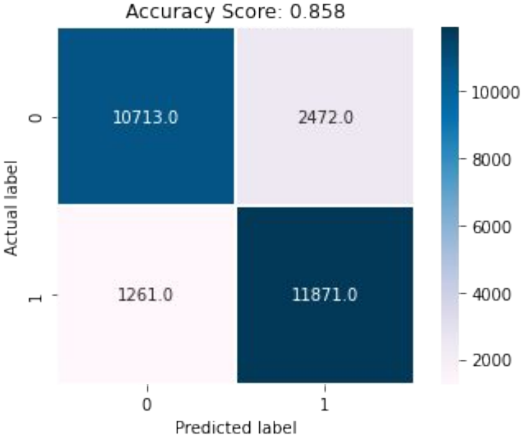
Algorithm	Accuracy		Precision		Recall		F1 Score		AUC ROC Score	
	IMBALANCED	BALANCED	IMBALANCED	BALANCED	IMBALANCED	BALANCED	IMBALANCED	BALANCED	IMBALANCED	BALANCED
Logistic Regression	88	79	55	80	19	78	28	79	86	86
Random Forest	89	85	74	82	15	90	25	86	91	93
k-Nearest N eighbors	89	81	63	82	14	79	24	80	88	90

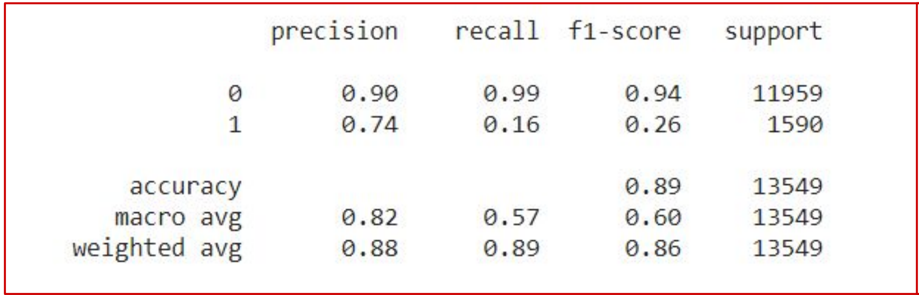
Winner Classification Report : Random Forest

Before

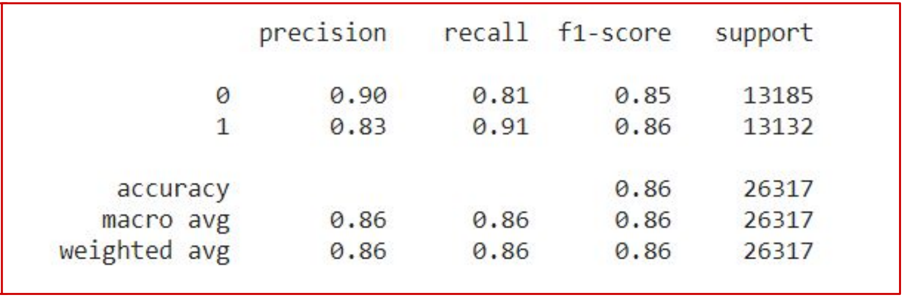


After





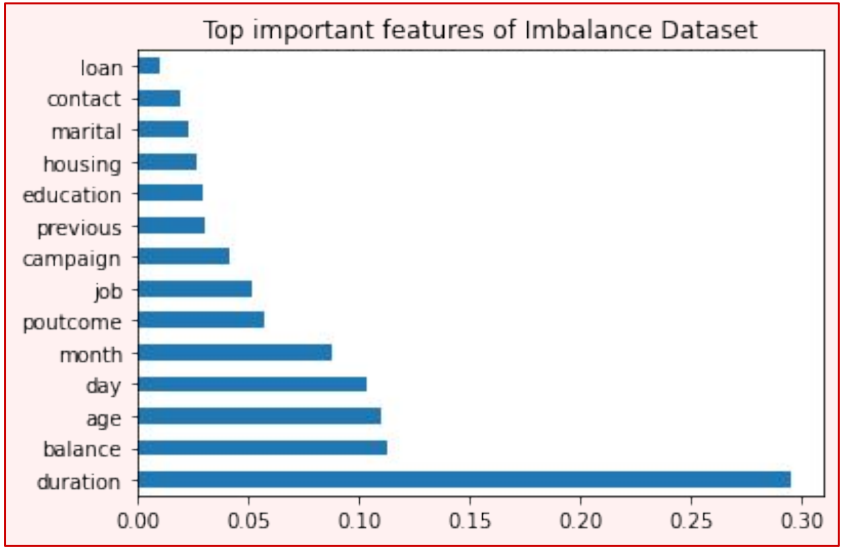
```
              precision    recall  f1-score   support

           0       0.90      0.99      0.94     11959
           1       0.74      0.16      0.26      1590

    accuracy                           0.89     13549
   macro avg       0.82      0.57      0.60     13549
weighted avg       0.88      0.89      0.86     13549
```



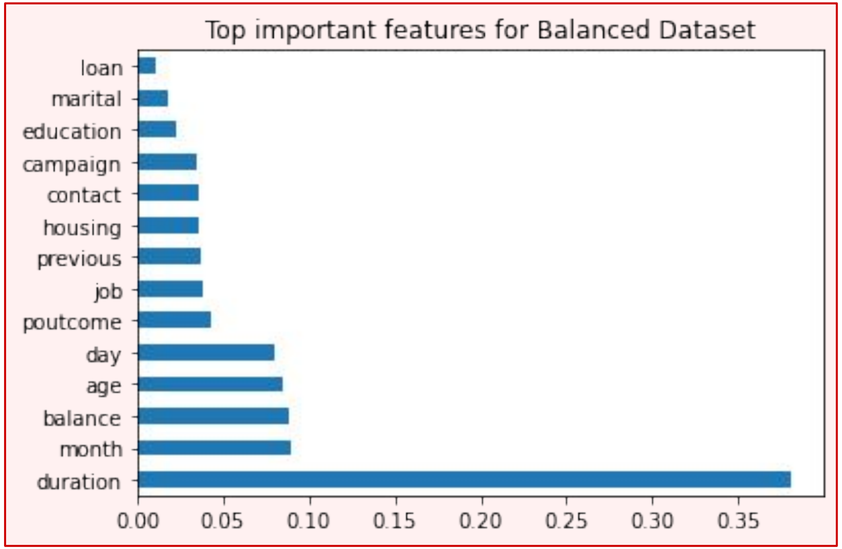
```
              precision    recall  f1-score   support

           0       0.90      0.81      0.85     13185
           1       0.83      0.91      0.86     13132

    accuracy                           0.86     26317
   macro avg       0.86      0.86      0.86     26317
weighted avg       0.86      0.86      0.86     26317
```

Important Feature : Using Random Forest

Before



After



Proposed solutions for the Next Marketing Campaign :

- Months of Marketing Activity: For the next marketing campaign, it will be wise for the bank to focus the marketing campaign during the months of March, September, October, and December. (December should be under consideration because it was the month with the lowest marketing activity, there might be a reason why December is the lowest.)

- Campaign Calls: A policy should be implemented that states that no more than 3 calls should be applied to the same potential client. Remember, the more we call the same potential client, the higher the likelihood he or she will decline to open a term deposit. This might be as a result of a potential client getting tired/pissed at being disturbed. It also saves us time and effort in getting new potential clients.

- Age Category: The customer's age affects campaign outcome as well. The next marketing campaign of the bank should target potential clients in their 20s or younger and 60s or older. This will increase the likelihood of more term deposits subscriptions.

- Occupation: Potential clients that were students or retired were the most likely to subscribe to a term deposit. Retired individuals, tend to have more term deposits in order to gain some cash through interest payments. Retired individuals tend to not spend so much of their money as responsibilities are usually reduced, so they are more likely to lend it to the financial institution. Students were the other group that used to subscribe to term deposits.

- Balances: We see those potential clients on average and high balances are more likely to open a term deposit. Lastly, the next marketing campaign should focus on individuals of average and high balances in order to increase the likelihood of subscribing to a term deposit as they have more money to spare.

Conclusion :

- From the study conducted, the results are impressive and convincing in terms of using a machine-learning algorithm to decide on the marketing campaign of the bank. Among the three classification approach used to model the data, we found that Accuracy and AUC ROC score is high for almost all models whether it is for balanced or Imbalance dataset.
    - Random Forest with balance set is highest with 93% of AUC ROC score.
    - Random Forest and Knn with Imbalance set both contain 89% of Accuracy.
- But for fair model performance among all balance and Imbalance datasets, we decided the best model using F1 Score here also Random Forest is the clear winner on balance dataset with a score of 86%.
- And also when we talk about Balance data set comparison Random Forest is the best model in all Evaluation metrics.
- Further, we can also improve using hyper tuning on our Algorithm to get the most out of it but Hyper Tuning consumes a lot of time and resources of the system depending upon how big the Data we have and what algorithm we're using. It will go through a number of Iterations and try to come up with the best possible value for us.
- The bank marketing manager can identify the potential client by using the model if the client's information like education, housing loan, Personal loan, duration of the call, number of contacts performed during this campaign, previous outcomes, etc is available. This will help the bank to predict the success of subscribing to a long-term deposit even before the telemarketing call is executed.
- Thank you for your time.

Q & A