<u>Project Report on:</u>

“Predicting the Probability of bank Deposit Subscription”

Submitted by :

Ahana Koshe : 336

Neha Mankani :340

Shafiya Sayyed :354

Shoieb Shaikh :357

May, 2021

Under The Guidance of :

Mr. Jubair Nadaf

Department of Computer Engineering .

DEPARTMENT OF COMPUTER ENGINEERING .

KEYSTONE SCHOOL OF ENGINEERING ,

Handewadi Chowk , Uruli Devachi , Shewalewadi

Pune, Maharashtra 412308

CERTIFICATE

This is to certify that the project report entitled
"Predicting the Probability of bank Deposit Subscription"
submitted by Ahana Koshe, Neha Mankani ,Shafiya Sayyed ,Shoieb Shaikh
of the Keystone School of Engineering, Pune
in partial fulfilment for the award of Degree of Bachelor of Engineering
in Computer Engineering is a bonafide record of the project work
carried out by the students under my supervision during the
Academic year 2020 - 2021.

Table of Contents:

TITLE

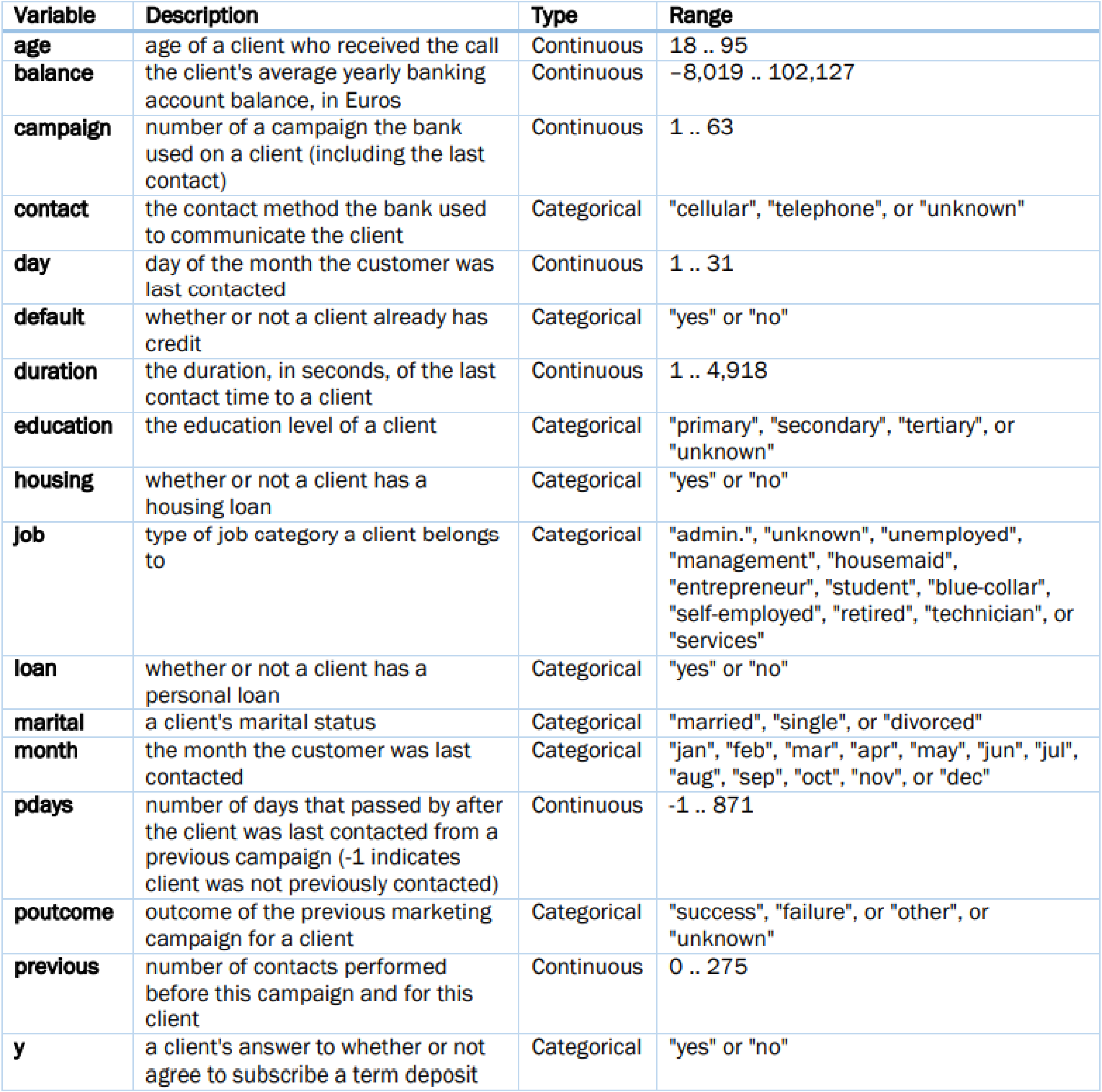
Predicting the Probability of Bank Deposit Subscription

ABSTRACT

As the number of marketing campaigns to which consumers are subjected continues to explode, forprofit businesses, non-profit charitable and community organizations, and political candidates are becoming increasingly dependent upon targeted direct-to-consumer (DTC) campaigns. In order for these to be successful, these organizations must invest heavily in strategies that select the best possible prospects. For this study, we examined 45,211 records related to direct marketing campaigns of a Portuguese Banking institution and attempted to define a reliable model for predicting consumer intent to subscribe to a term deposit.

INTRODUCTION AND DATA COLLECTION

The original study uses data collected from a Portuguese bank for 17 direct marketing campaigns spanning from May 2008 to November 2010. The purpose was to determine the effectiveness of direct marketing campaigns (DM) by observing whether a client decided to make a long-term bank deposit after receiving a DM phone call. [1] The original data set contained 17 explanatory variables and 45,211 rows. The response variable is 'y' and is categorical. After removing corrupted or partial data rows, the cleaned data set contained 45,191 observations. Due to the size of the data, we create several data sets by utilizing the random function in Excel to split the original data into training, validation, and test datasets.



Variable	Description	Type	Range
age	age of a client who received the call	Continuous	18 .. 95
balance	the client's average yearly banking account balance, in Euros	Continuous	−8,019 .. 102,127
campaign	number of a campaign the bank used on a client (including the last contact)	Continuous	1 .. 63
contact	the contact method the bank used to communicate the client	Categorical	"cellular", "telephone", or "unknown"
day	day of the month the customer was last contacted	Continuous	1 .. 31
default	whether or not a client already has credit	Categorical	"yes" or "no"
duration	the duration, in seconds, of the last contact time to a client	Continuous	1 .. 4,918
education	the education level of a client	Categorical	"primary", "secondary", "tertiary", or "unknown"
housing	whether or not a client has a housing loan	Categorical	"yes" or "no"
job	type of job category a client belongs to	Categorical	"admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", or "services"
loan	whether or not a client has a personal loan	Categorical	"yes" or "no"
marital	a client's marital status	Categorical	"married", "single", or "divorced"
month	the month the customer was last contacted	Categorical	"jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", or "dec"
pdays	number of days that passed by after the client was last contacted from a previous campaign (-1 indicates client was not previously contacted)	Continuous	-1 .. 871
poutcome	outcome of the previous marketing campaign for a client	Categorical	"success", "failure", or "other", or "unknown"
previous	number of contacts performed before this campaign and for this client	Continuous	0 .. 275
y	a client's answer to whether or not agree to subscribe a term deposit	Categorical	"yes" or "no"

Table 1. List of variables in the original dataset

STUDY OBJECTIVE

We will explore the research question does poutcome alone adequately predict whether a DM phone call will result in success (y = 'yes') or does a combination of multiple variables (and which variables) better predict success?

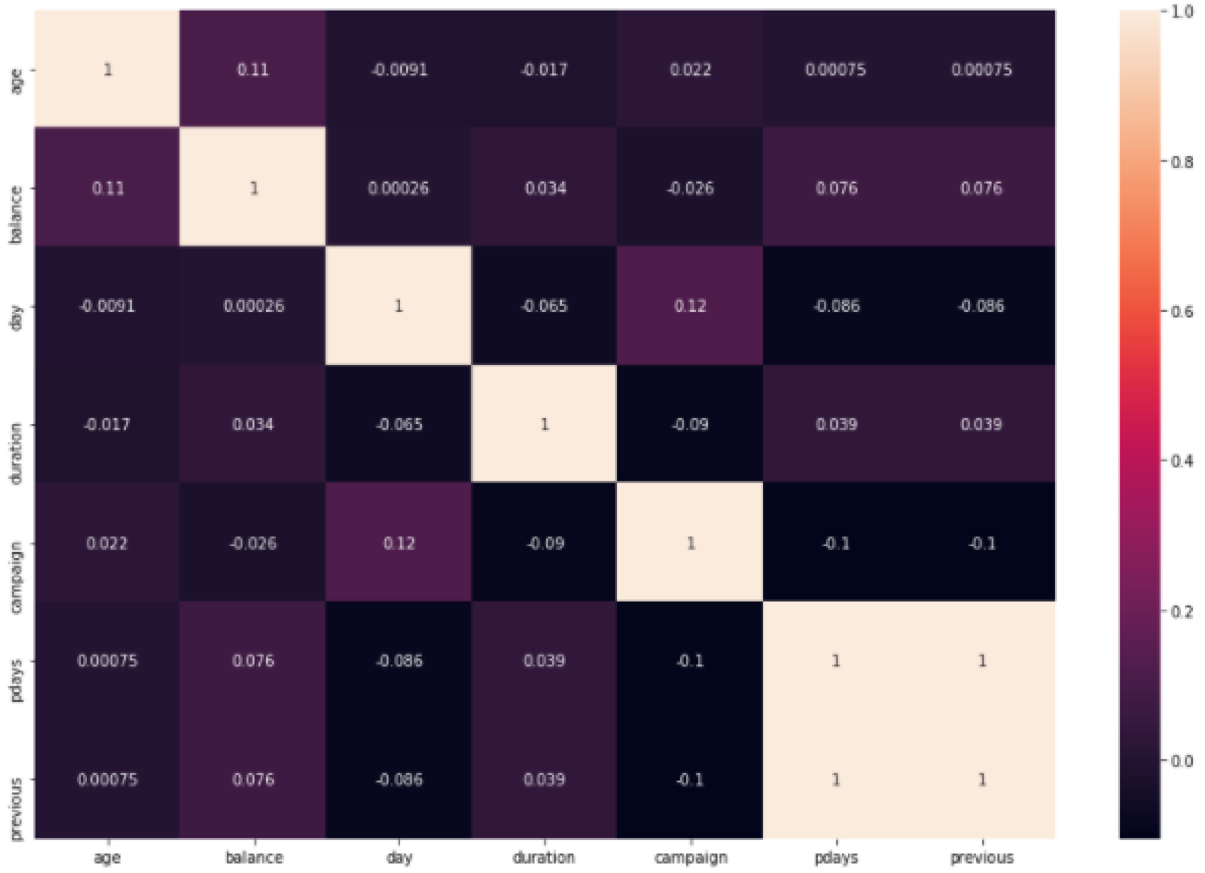
THEORY & ANALYSIS

1. PRELIMINARY ANALYSIS

The data set contains a mixture of continuous and categorical potential predictive variables. It is interesting to note that the longer time frames between contacts (pdays), higher account balance

(balance), length of previous contact (duration), and higher numbers of previous contact appear to hold the most notable differences and lead to a response value of ‘yes’.

Additionally, from Table 3, it shows the correlations for the continuous variables. The correlations are not particularly high in most cases, but the correlation between pdays and previous is moderately strong and statistically significant from 0 (correlation=0.57633). (This collinearity is expected, especially since there is an exact one-to-one correlation between pdays=–1 -- client was not previously contacted; and previous=0 -- no contact was made o this client before the current campaign.)



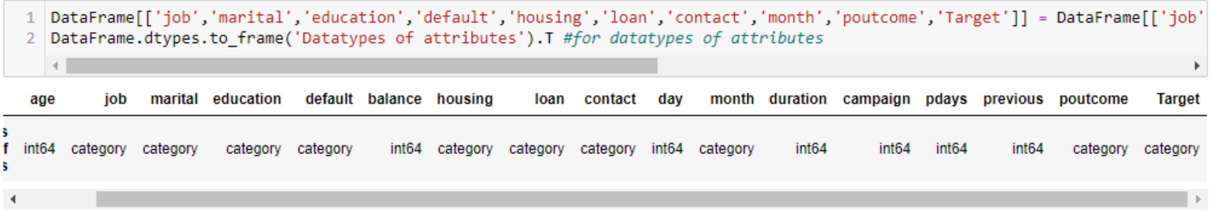
2. Cleaning Data

a. 'job','education' and 'contact' attributes contain multiple unknown values.
b. As they are specific string values and not numerical so cannot be treated with median, mean replacement
c. As per data description, duration attribute is removed to have a realistic prediction.
d. 'day' and 'month' of contact will not have any realtionship with customer opting for subscription

e. 'pdays' and 'previous' are highly correlated so removing any one from both of them.

f. Even though 'poutcome' has a huge number of 'Unknown' values, it is kept because 'poutcome' is outcome of the previous marketing campaign. Which can help in better training and testing of models.

3. Checking the data type of each attribute :



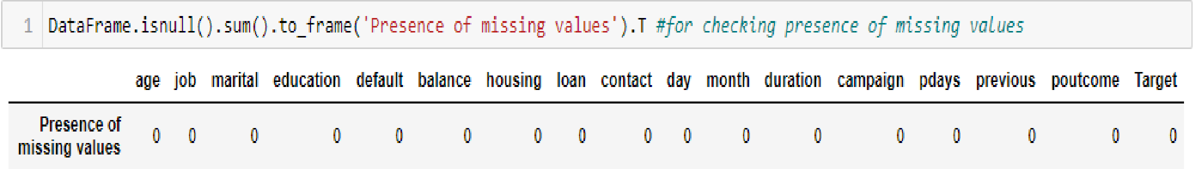
```
DataFrame[['job','marital','education','default','housing','loan','contact','month','poutcome','Target']] = DataFrame[['job'
DataFrame.dtypes.to_frame('Datatypes of attributes').T #for datatypes of attributes
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Target
s f s	int64	category	category	category	category	int64	category	category	category	int64	category	int64	int64	int64	int64	category	category

a. 10 features have object datatype and 7 have int datatype

b. As per given data we know that 10 features are of category type,so lets convert the datatype of those features

4. Checking the presence of missing values :

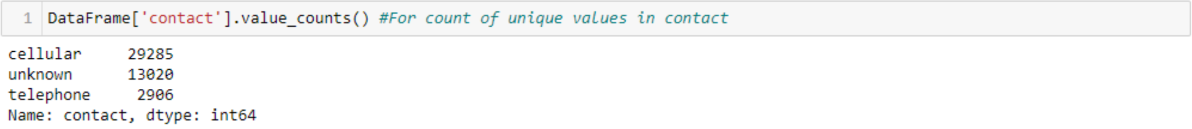


```
DataFrame.isnull().sum().to_frame('Presence of missing values').T #for checking presence of missing values
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Target
Presence of missing values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

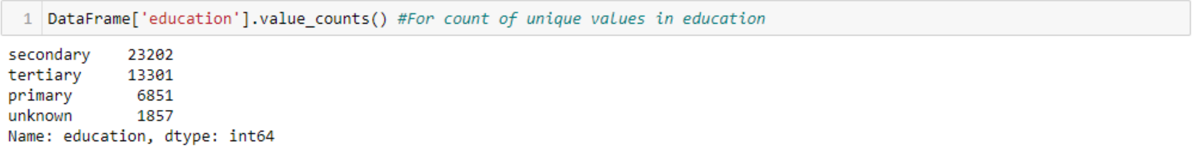
a. In above cell, missing values are not found, so checking columns where unknown is mentioned

b. Checking columns which contain string data



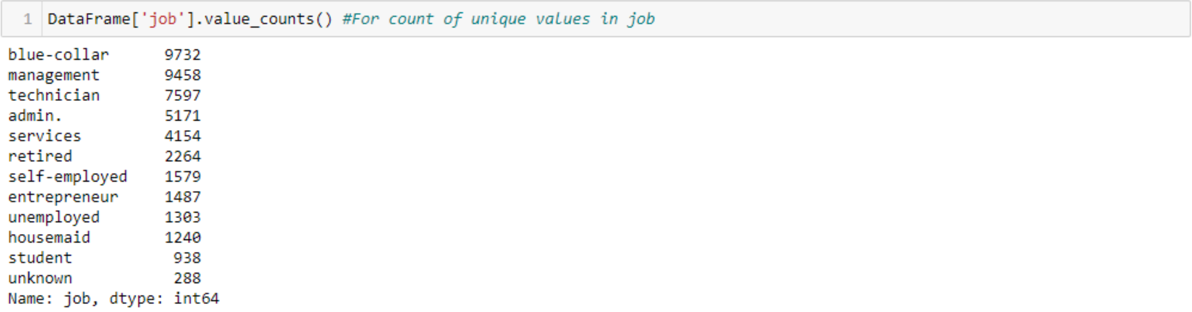
```
DataFrame['contact'].value_counts() #For count of unique values in contact
```

```
cellular     29285
unknown      13020
telephone     2906
Name: contact, dtype: int64
```



```
DataFrame['education'].value_counts() #For count of unique values in education
```

```
secondary    23202
tertiary     13301
primary       6851
unknown       1857
Name: education, dtype: int64
```

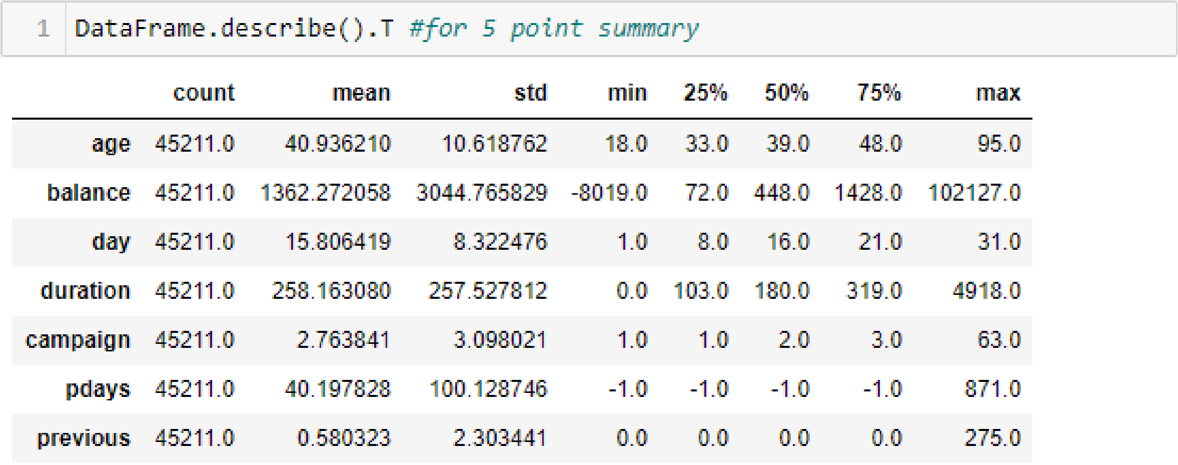


```
DataFrame['job'].value_counts() #For count of unique values in job
```

```
blue-collar      9732
management       9458
technician       7597
admin.           5171
services         4154
retired          2264
self-employed    1579
entrepreneur     1487
unemployed       1303
housemaid        1240
student           938
unknown           288
Name: job, dtype: int64
```

- The dataset has unknown values in 'job','education','contact' and 'poutcome' columns

5. Looking at the 5 Point summary :

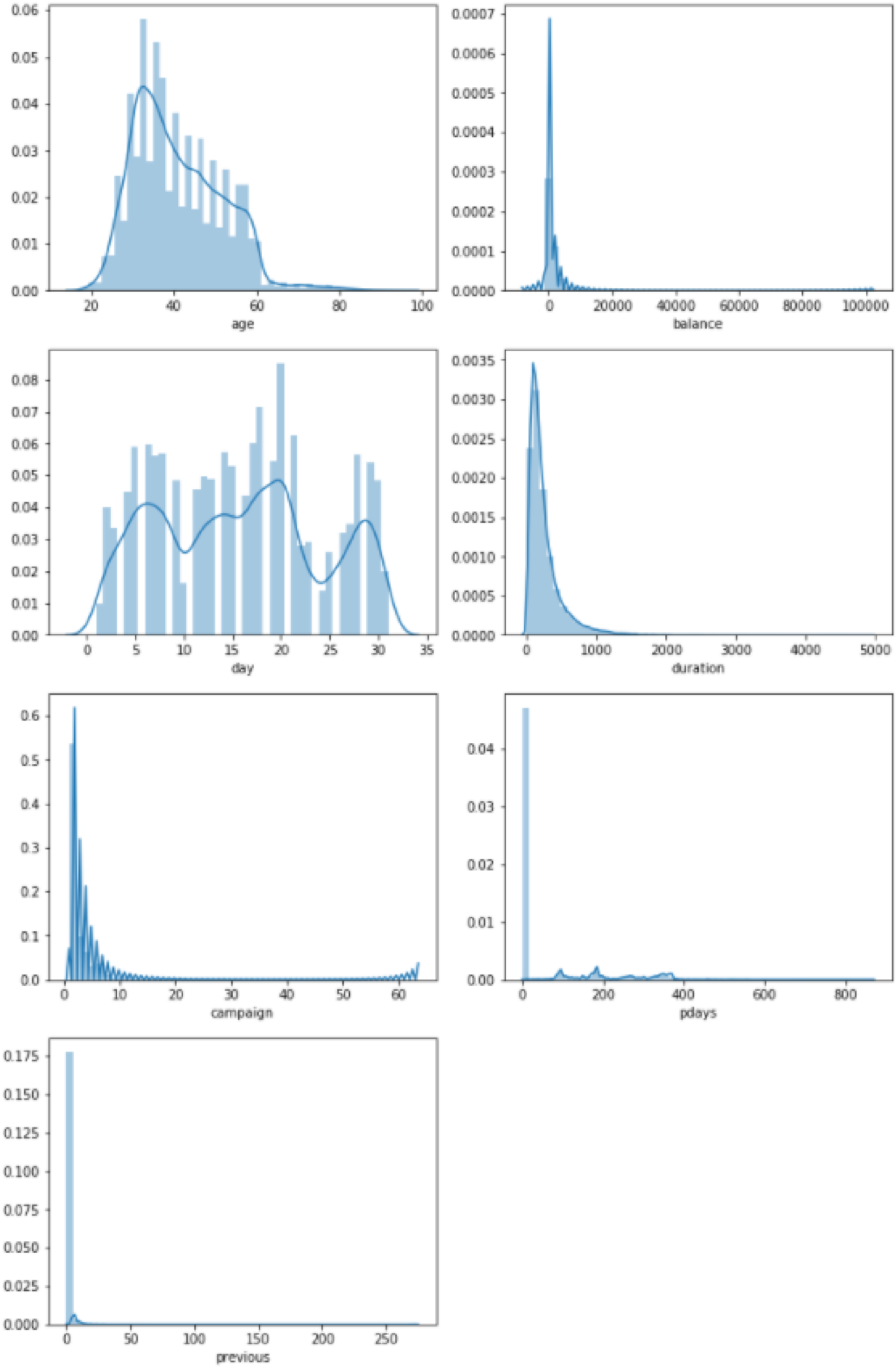


```
DataFrame.describe().T #for 5 point summary
```

	count	mean	std	min	25%	50%	75%	max
age	45211.0	40.936210	10.618762	18.0	33.0	39.0	48.0	95.0
balance	45211.0	1362.272058	3044.765829	-8019.0	72.0	448.0	1428.0	102127.0
day	45211.0	15.806419	8.322476	1.0	8.0	16.0	21.0	31.0
duration	45211.0	258.163080	257.527812	0.0	103.0	180.0	319.0	4918.0
campaign	45211.0	2.763841	3.098021	1.0	1.0	2.0	3.0	63.0
pdays	45211.0	40.197828	100.128746	-1.0	-1.0	-1.0	-1.0	871.0
previous	45211.0	0.580323	2.303441	0.0	0.0	0.0	0.0	275.0

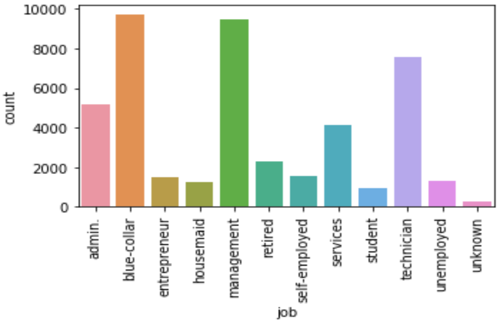
a. Outliers are present in 'age', 'balance', 'duration', 'campaign', 'pdays' and 'previous' columns.
b. 'balance', 'duration', 'campaign', 'pdays' and 'previous' are right skewed.
c. More than 75% people have been contacted in a day after previous campaign as pdays is -1 till 75th precentile
d. Minimum balance is -8019 and maximum balance is 102127
e. Minimum age is 18 years and maximum is 95 years

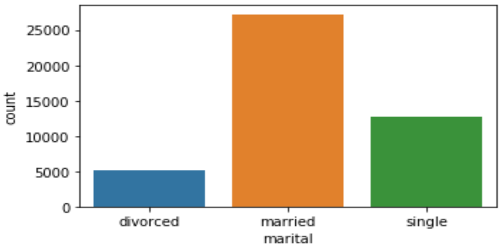
6. Checking the distribution of numerical columns.

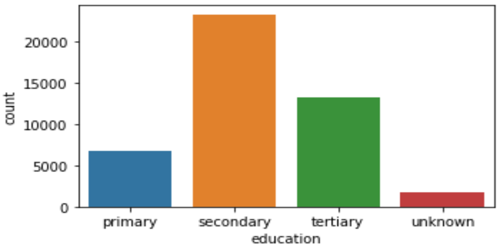


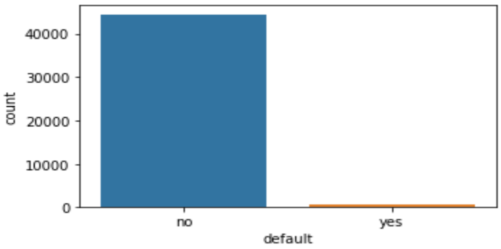
a. 'balance', 'duration', 'campaign', 'pdays' and 'previous' are right skewed.
b. 'age' is somewhat normally distributed

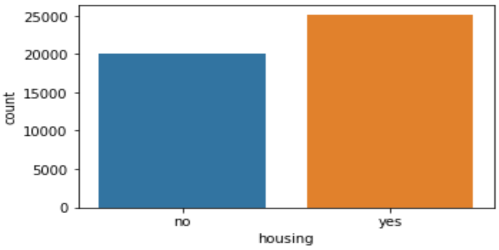
7. Understanding the distribution of categorical data:

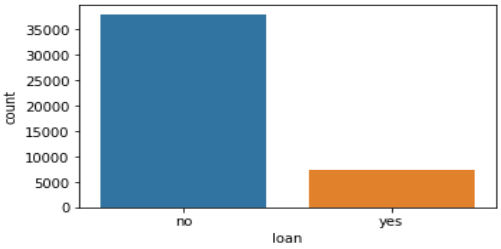


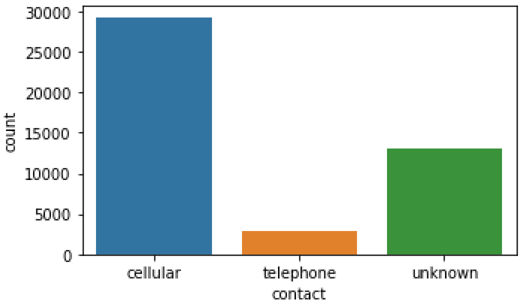


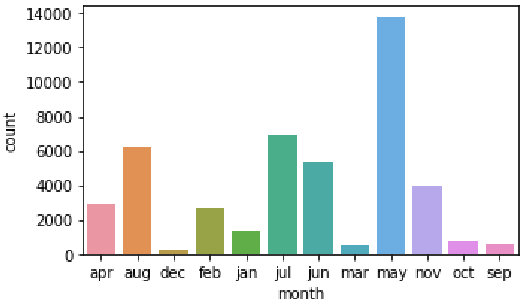


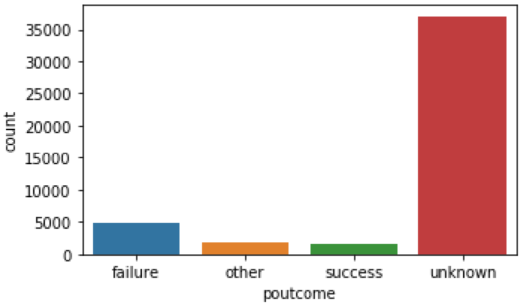


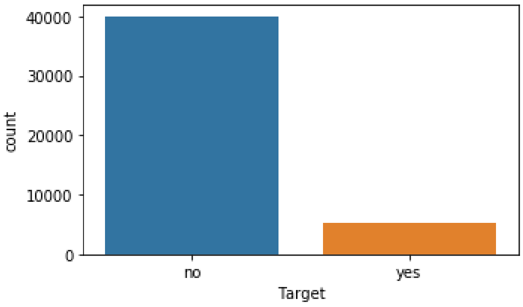






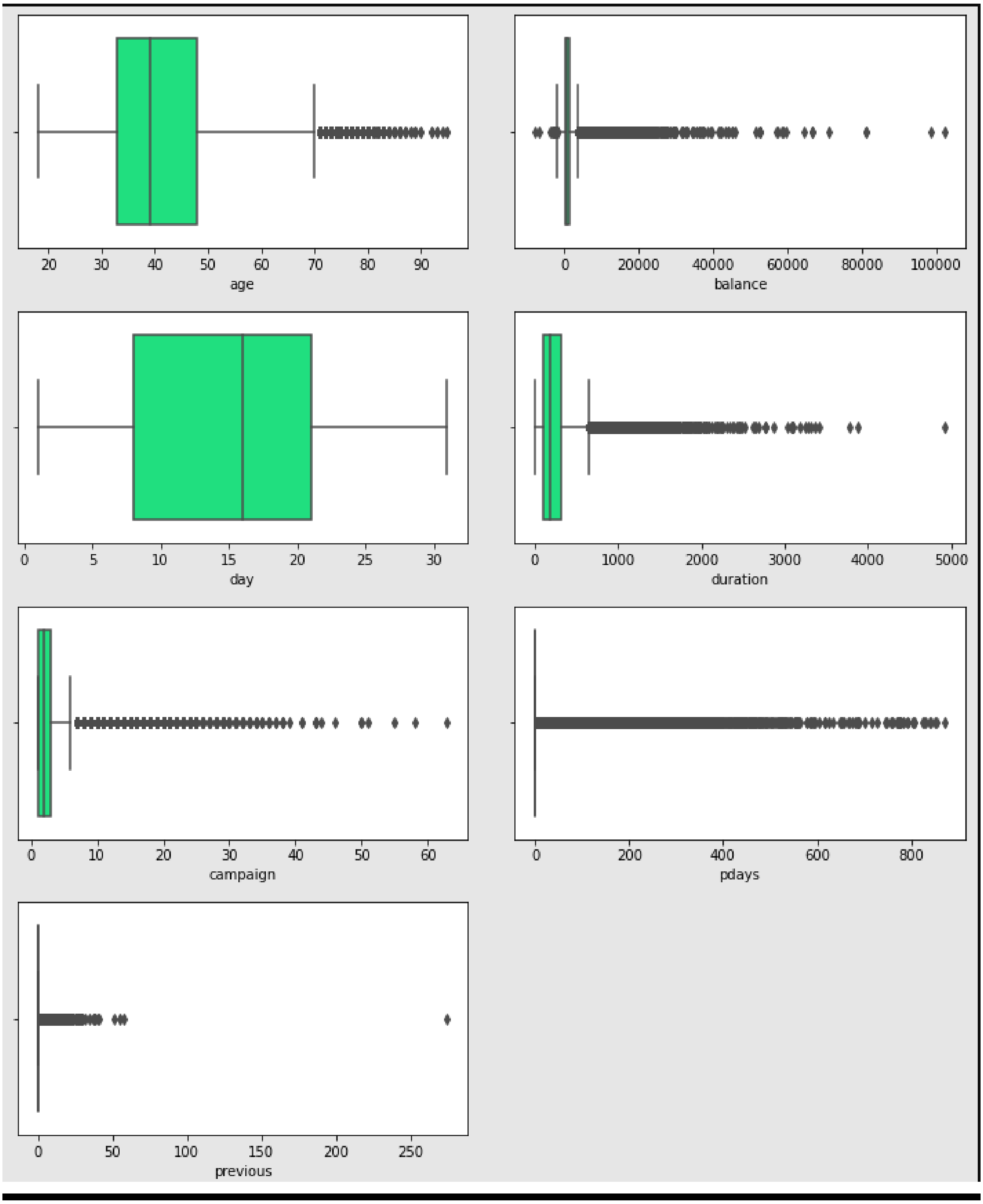






a. More than 90% customers have no default credit
b. Around 88% customers have not subscribed for term deposit
c. Most customers have been contacted in may
d. Most customers have been contacted by cellular network(mobile phone)
e. Number of customers who have housing loan is more than the number of customers who don't have housing loan
f. Around 6% customers have credit in default

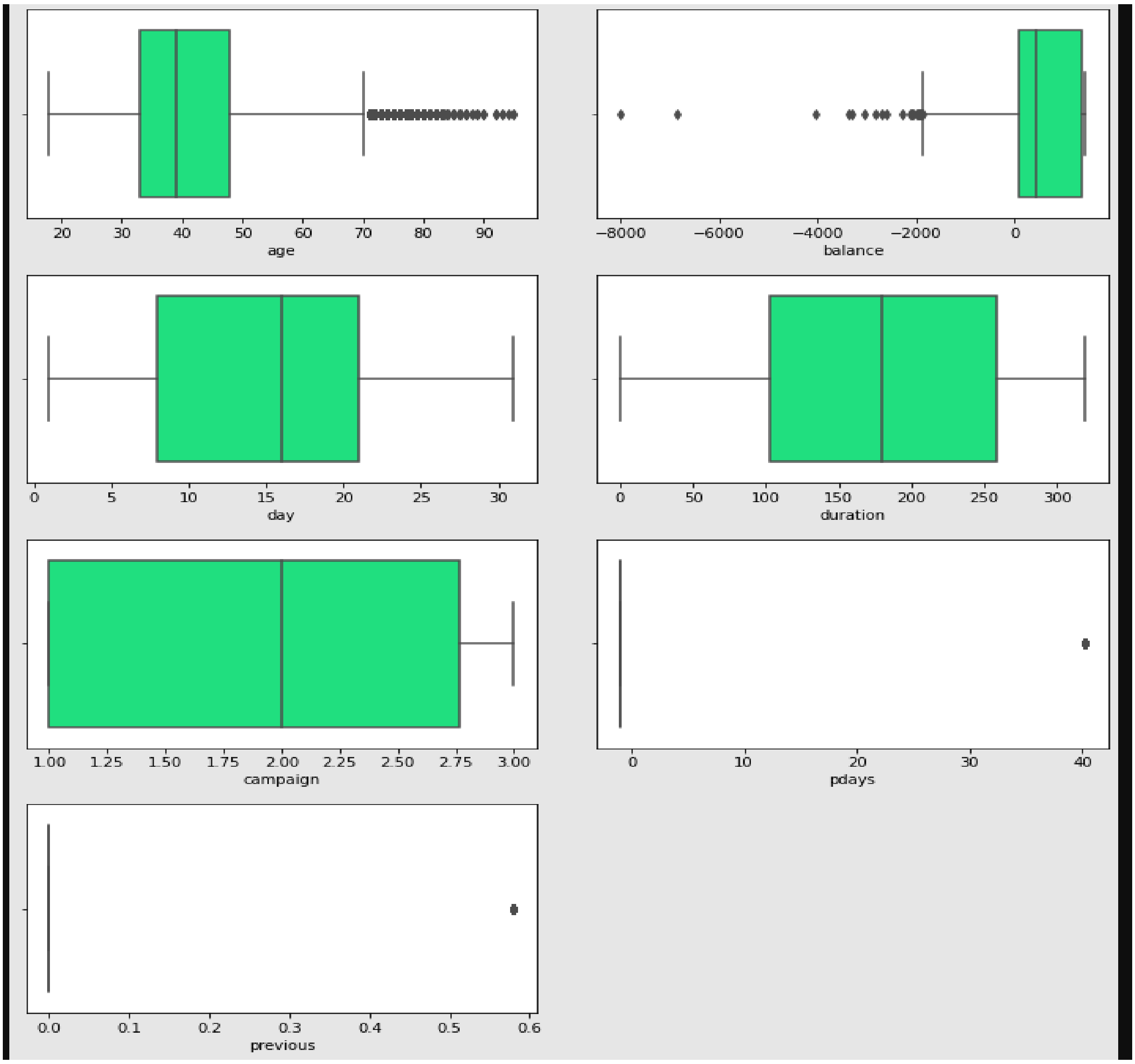
8. Checking the presence of outliers :



a. Prove the existance of outliers in 'age', 'balance', 'duration', 'campaign', 'pdays' and 'previous' columns.
b. Values less than 0 are present in 'Balance' column

9. Handling Outliers with mean value

a. Here we replaces outliers with the mean values of that feature to get better accuracy from the model and the standard deviation is maintained .
b. After replacing the outliers this was the final boplot .



10. Removing Columns:

a. 'job','education' and 'contact' attributes contain multiple unknown values.
b. As they are specific string values and not numerical so cannot be treated with median, mean replacement
c. As per data description, duration attribute is removed to have a realistic prediction.

d. 'day' and 'month' of contact will not have any realtionship with customer opting for subscription
e. 'pdays' and 'previous' are highly correlated so removing any one from both of them.
f. Even though 'poutcome' has a huge number of 'Unknown' values, it is kept because 'poutcome' is outcome of the previous marketing campaign. Which can help in better training and testing of models.

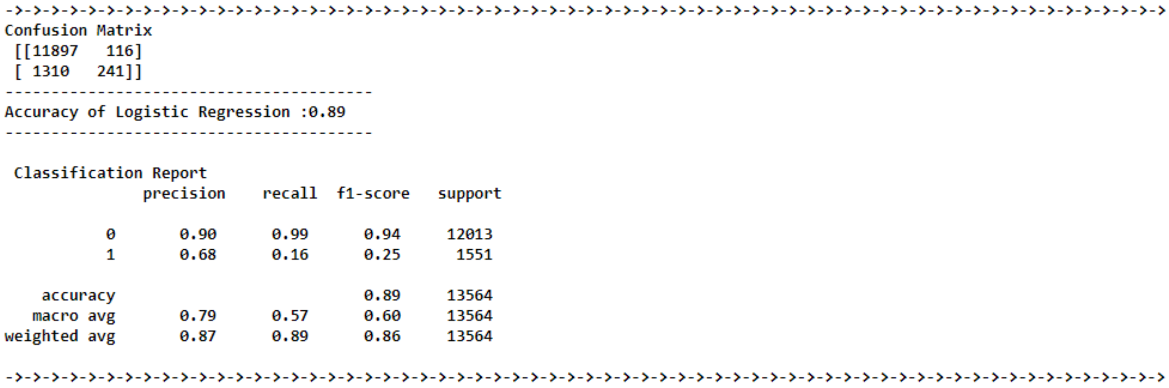
11. MODEL SELECTION

At the outset of this analysis, we undertook the task of identifying a rich model based upon the Portuguese banking dataset that might provide greater insights into the customers likely to purchase a new deposit subscription. To devise such a model, we worked independently through the training phase to identify three distinct models, which we describe briefly in this section.

Model 1&2 (Neha)

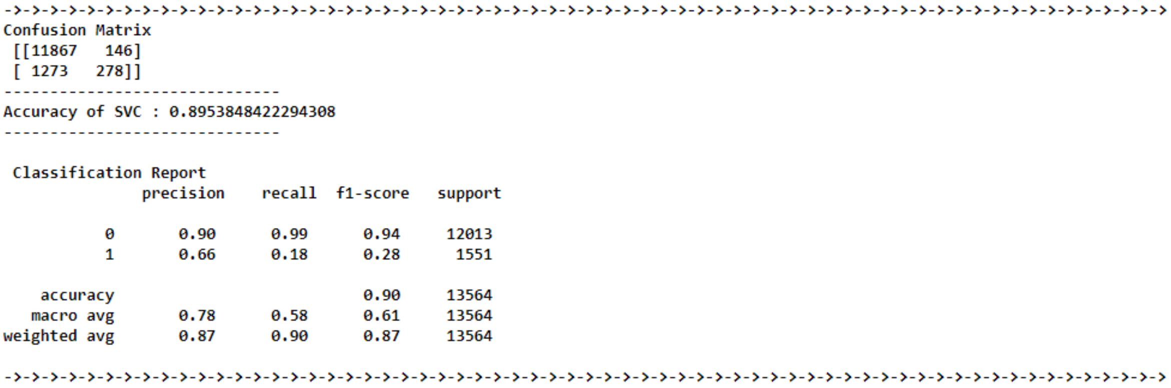
Neha worked on 2 machine learning models namely logistic regression and support vector classifier and devised the accuracy score as mentioned below.

1.Logistic regression :



```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[11897   116]
 [ 1310   241]]
----------------------------------------
Accuracy of Logistic Regression :0.89
----------------------------------------

 Classification Report
               precision    recall  f1-score   support

           0       0.90      0.99      0.94     12013
           1       0.68      0.16      0.25      1551

    accuracy                           0.89     13564
   macro avg       0.79      0.57      0.60     13564
weighted avg       0.87      0.89      0.86     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

2.Support Vector classifier :



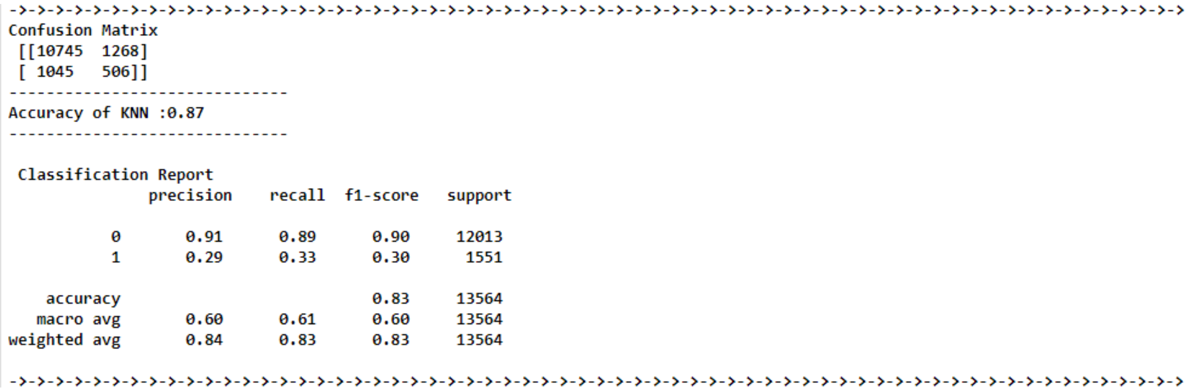
```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[11867   146]
 [ 1273   278]]
------------------------------
Accuracy of SVC : 0.8953848422294308
------------------------------

 Classification Report
               precision    recall  f1-score   support

           0       0.90      0.99      0.94     12013
           1       0.66      0.18      0.28      1551

    accuracy                           0.90     13564
   macro avg       0.78      0.58      0.61     13564
weighted avg       0.87      0.90      0.87     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

Neha noticed a bit more accuracy in support vector classifier and we considered this as an option for prediction as well.

Model 3&4 (Shafiya)

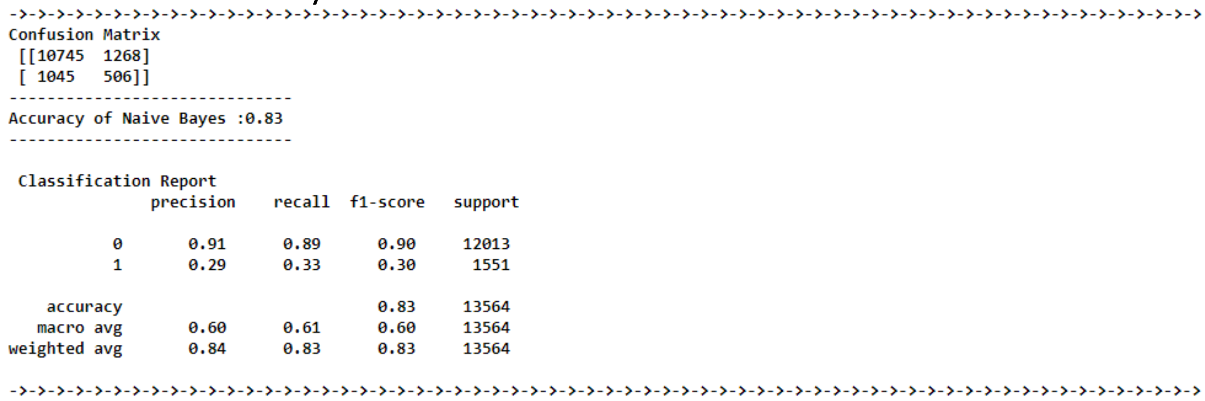
Shafiya had a different method of approach she choose the K nearest neighbour first to check what accuracy it gives and next she used the naïve bayes classifier specifically gaussian naïve bayes classifier for prediction and below are the scores .

1. KNN:



```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[10745  1268]
 [ 1045   506]]
------------------------------
Accuracy of KNN :0.87
------------------------------

 Classification Report
              precision    recall  f1-score   support

           0       0.91      0.89      0.90     12013
           1       0.29      0.33      0.30      1551

    accuracy                           0.83     13564
   macro avg       0.60      0.61      0.60     13564
weighted avg       0.84      0.83      0.83     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

2.Gaussion Naïve Bayes Classifier :



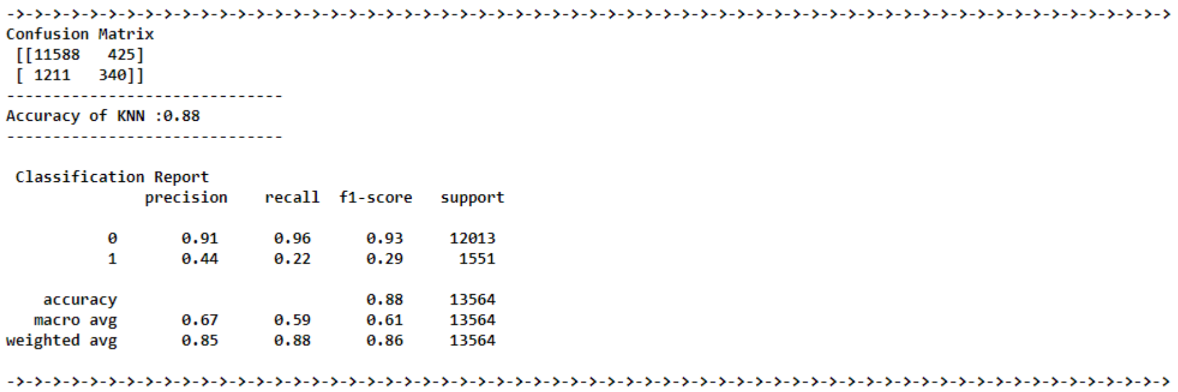
```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[10745  1268]
 [ 1045   506]]
------------------------------
Accuracy of Naive Bayes :0.83
------------------------------

 Classification Report
              precision    recall  f1-score   support

           0       0.91      0.89      0.90     12013
           1       0.29      0.33      0.30      1551

    accuracy                           0.83     13564
   macro avg       0.60      0.61      0.60     13564
weighted avg       0.84      0.83      0.83     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

Her there were no greater finding and accuracy result in the model made by Shafiya .

Model 5 (Shoieb)

For the next models we wanted to enhance the score and go beyond the models so we thought of using ensemble techniques to check if that would affect the accuracy of the model . Shoieb used Random forest technique over the KNN model made by shafiya to find out there was not much increase in the accuracy .

1.Below are the result when we used Random forest ensemble technique for feature selection :

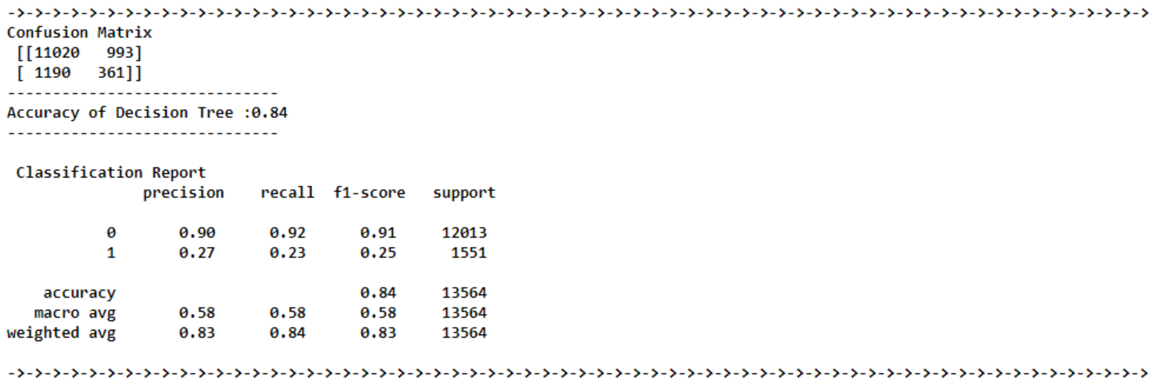


```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[11588   425]
 [ 1211   340]]
------------------------------
Accuracy of KNN :0.88
------------------------------

 Classification Report
              precision    recall  f1-score   support

           0       0.91      0.96      0.93     12013
           1       0.44      0.22      0.29      1551

    accuracy                           0.88     13564
   macro avg       0.67      0.59      0.61     13564
weighted avg       0.85      0.88      0.86     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

Model 6 (Ahana)

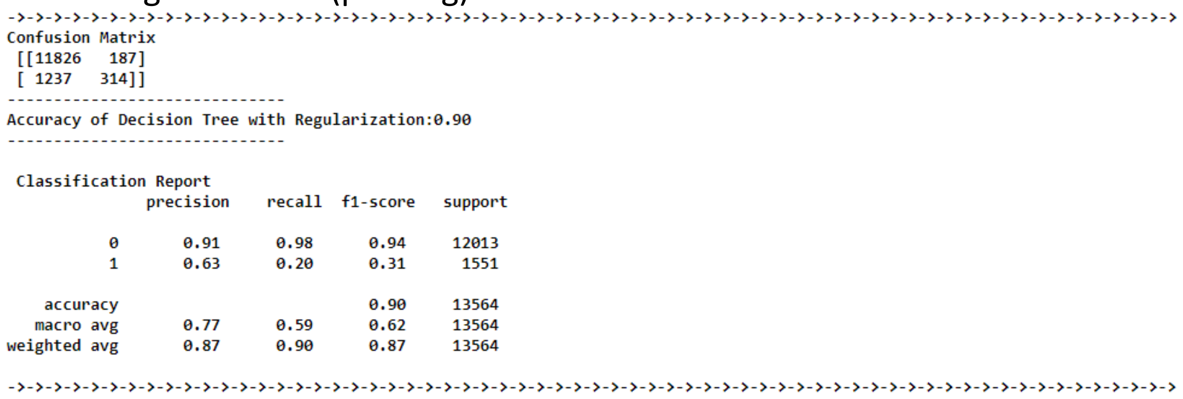
Ahana went for Decision Tree classifier and worked on the pruning of the decision tree as well so we could achieve more and she went on to do regularization on decision tree by pruning it and we found good results :

1.Normal Decision tree:



```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[11020   993]
 [ 1190   361]]
------------------------------
Accuracy of Decision Tree :0.84
------------------------------

 Classification Report
              precision    recall  f1-score   support

           0       0.90      0.92      0.91     12013
           1       0.27      0.23      0.25      1551

    accuracy                           0.84     13564
   macro avg       0.58      0.58      0.58     13564
weighted avg       0.83      0.84      0.83     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

2.After Regularization (pruning):



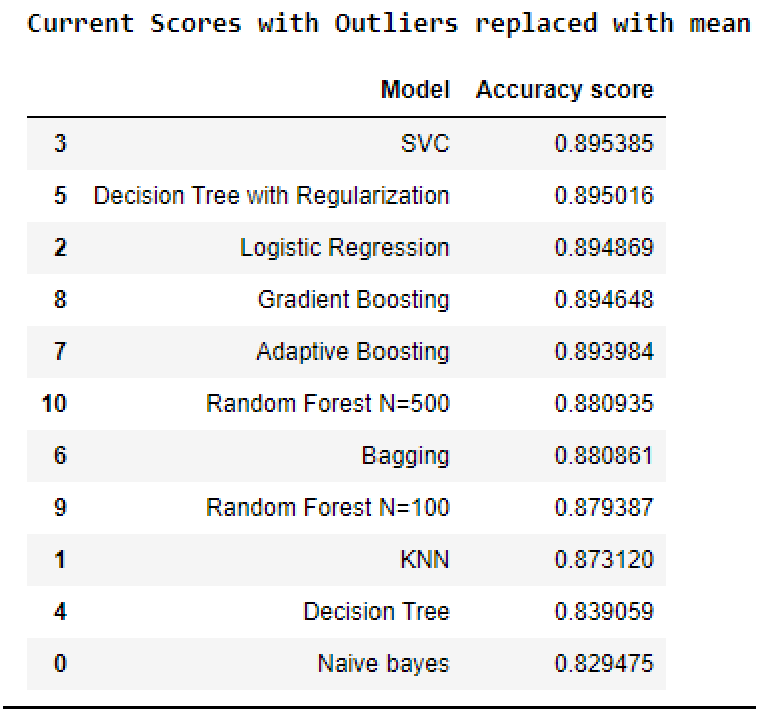
```
->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
Confusion Matrix
 [[11826   187]
 [ 1237   314]]
------------------------------
Accuracy of Decision Tree with Regularization:0.90
------------------------------

 Classification Report
              precision    recall  f1-score   support

           0       0.91      0.98      0.94     12013
           1       0.63      0.20      0.31      1551

    accuracy                           0.90     13564
   macro avg       0.77      0.59      0.62     13564
weighted avg       0.87      0.90      0.87     13564

->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->->
```

Here we realised the accuracy had increased more than other

.

12. Selecting a Final Model

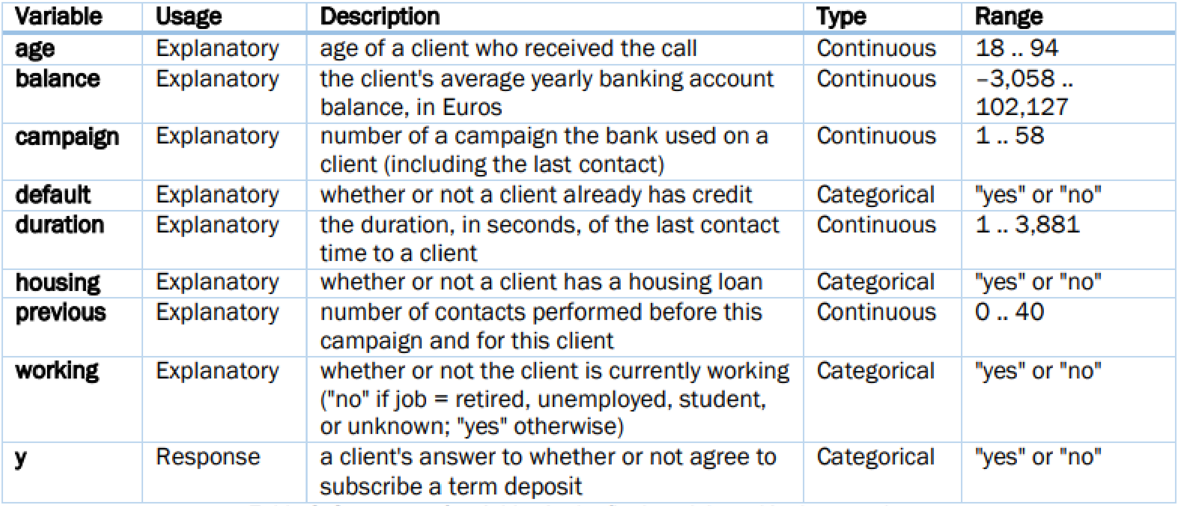
We went on to test all the model and get an score out of them and below are the results of the testing and we decided to take the decision tree model with regularization as that gave the highest accuracy .



```
Current Scores with Outliers replaced with mean
```

	Model	Accuracy score
3	SVC	0.895385
5	Decision Tree with Regularization	0.895016
2	Logistic Regression	0.894869
8	Gradient Boosting	0.894648
7	Adaptive Boosting	0.893984
10	Random Forest N=500	0.880935
6	Bagging	0.880861
9	Random Forest N=100	0.879387
1	KNN	0.873120
4	Decision Tree	0.839059
0	Naive bayes	0.829475

This table shows the comparison of all the models and techniques we tried and here we can see that when we replaced the outliers with the mean value and used decision tree with regularization we had an amazing accuracy and hence we thought of going on with that model.



Variable	Usage	Description	Type	Range
age	Explanatory	age of a client who received the call	Continuous	18 .. 94
balance	Explanatory	the client's average yearly banking account balance, in Euros	Continuous	−3,058 .. 102,127
campaign	Explanatory	number of a campaign the bank used on a client (including the last contact)	Continuous	1 .. 58
default	Explanatory	whether or not a client already has credit	Categorical	"yes" or "no"
duration	Explanatory	the duration, in seconds, of the last contact time to a client	Continuous	1 .. 3,881
housing	Explanatory	whether or not a client has a housing loan	Categorical	"yes" or "no"
previous	Explanatory	number of contacts performed before this campaign and for this client	Continuous	0 .. 40
working	Explanatory	whether or not the client is currently working ("no" if job = retired, unemployed, student, or unknown; "yes" otherwise)	Categorical	"yes" or "no"
y	Response	a client's answer to whether or not agree to subscribe a term deposit	Categorical	"yes" or "no"

LIMITATIONS

The study is observational, so causal relationships cannot be drawn from the results. In the original study, success (y='yes') was only 8% of the total data set. Since success was such a small margin in the data, it created difficulty in identifying the best model to achieve success. There were several variables (contact, education, job, poutcome) with values 'unknown' in the data set. These presented difficulties, as we were most interested in success or failure in poutcome to best investigate the study objective. We kept the 'unknown' in the data sets, as removing the data could have costly effects in determining the final model and conclusions that can be drawn from the model

CONCLUSION

Comments on dataset:

- The models perform well in predicting the class 0 i.e. customer not subscribing to term deposit which can be seen in the confusion matrix of all models.
- The models do not perform well in predicting the class 1 i.e. customer subscribing to term deposit which can be seen in the confusion matrix of all models.
- Above situation occured because the Dataset is imbalanced. i.e. The ratio difference between class 0 and class 1 is huge. Which trained models to effectively identify class 0 but did not train suffuiciently to classify class 1.
- This situation could have been avoided if the datset was balanced.
- Along with imbalance, the dataset contained large number of unknown string values in 'job','education','contact' and 'poutcome' columns.

Comments on Models:

- When benchmarking with 'duration' column, Support Vector Classifier achieved 90% model accuracy while naive bayes score was 85% accurate.
- Decision Tree performed better than others in normal condition on the data set .
- After removing the 'duration' column, The highest model score dropped by 0.5%.
- The Outliers did not affect much on accuracy scores of all models. As can be seen in above accuracy scores, getting rid of outliers by mean/median replacement did not affect the scores.
- In Decision Trees, Gradient boosting method always performed better for this dataset.
- While visualizing Decision Tree, The Pruned decision tree was easy to visualize as it had lesser leaf nodes than Tree which was not pruned.

Miscellaneous Comments:

- After trying get_dummies the score did not show significant difference as well as I have skipped the get_dummies step because the dataset was creating more dimension, which was making the project more computationally intensive.
- If I had kept get_dummies step, then in production stage if the new dataset turned out to be huge in number of rows then this project would have taken a lot of time to execute.
- Outlier handeling did not make any significant difference in the accuracy scores of models.
- I have tried to keep minimum time complexity of this project.

REFERENCES

1. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS. 2. Kramer AA, and JE Zimmerman. “Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited.” Crit Care Med. 2007 Sep;35(9):2052-6. Internet: http://www.ncbi.nlm.nih.gov/pubmed/17568333 3. Allison, Paul. “Alternatives to the Hosmer-Lemeshow Test.” Blog post on Statistical Horizons, Apr. 9, 2014. Internet: http://statisticalhorizons.com/alternatives-to-the-hosmer-lemeshow-test 4. GOFLOGIT is a SAS macro for computing several alternative goodness-of-fit tests in logistic regression. Internet: https://github.com/friendly/SAS-macros/blob/master/goflogit.sas