Capstone Project

Predicting the effectiveness of bank marketing campaigns

Member Name

Soumabha Sarkar

Points to discuss

- Agenda
- Data summary
- EDA and feature engineering
- Preparing dataset for modelling
- Applying Model
- Model validation and selection
- Conclusion

Agenda

The data is related with direct marketing campaigns (phone calls) 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

Bank Client data:

Age: (numeric)

Job: type of job (categorical: ‘admin’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’).

Marital: marital status (categorical: ‘divorced’, ‘married’, ‘single’, ‘unknown’; note: ‘divorced’ means divorced or widowed).

Data Summary(continued)

education: (categorical: 'tertiary', 'secondary', 'unknown', 'primary‘)

default: has credit in default? (categorical: 'no', 'yes', 'unknown')

housing: has housing loan? (categorical: 'no‘, 'yes', 'unknown')

loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Data Summary(continued)

Related with the last contact of the current campaign:

contact: contact communication type (categorical: ‘cellular’, ‘telephone’)

month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, ..., ‘nov’, ‘dec’)

day: last contact day of the week (categorical: ‘mon’, ‘tue’, ‘wed’, ‘thu’, ‘fri’)

duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Data Summary(continued)

Other attributes:

campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

previous: number of contacts performed before this campaign and for this client (numeric)

poutcome: outcome of the previous marketing campaign (categorical: ‘failure’, ‘nonexistent’, ‘success’)

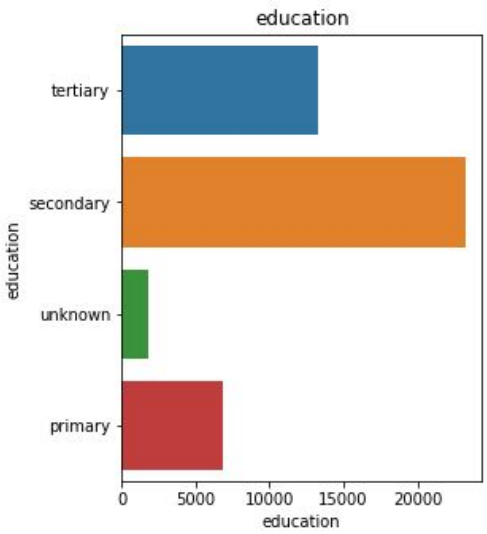
Output variable (desired target):

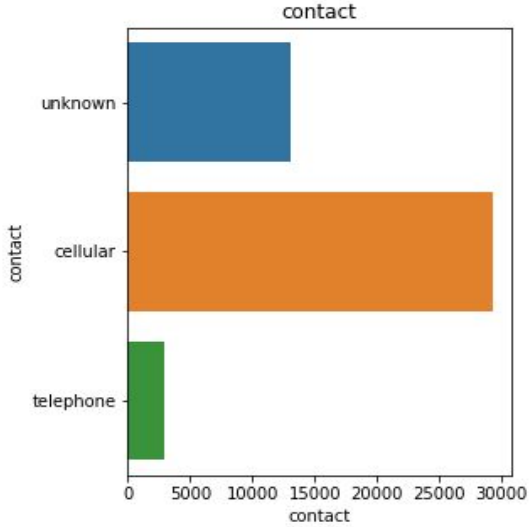
y - has the client subscribed a term deposit? (binary: ‘yes’, ‘no’)

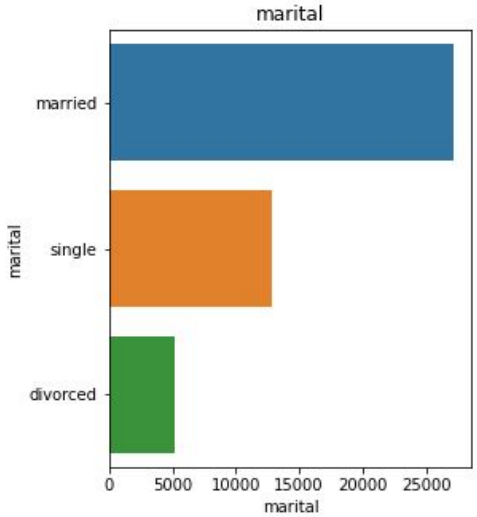
EDA and feature engineering

- Counting Categorical features
- Relationship of categorical features with the output variable.
- Distribution of numerical variables.
- Outliers detection.
- Correlation between numerical features.

Counting categorical features

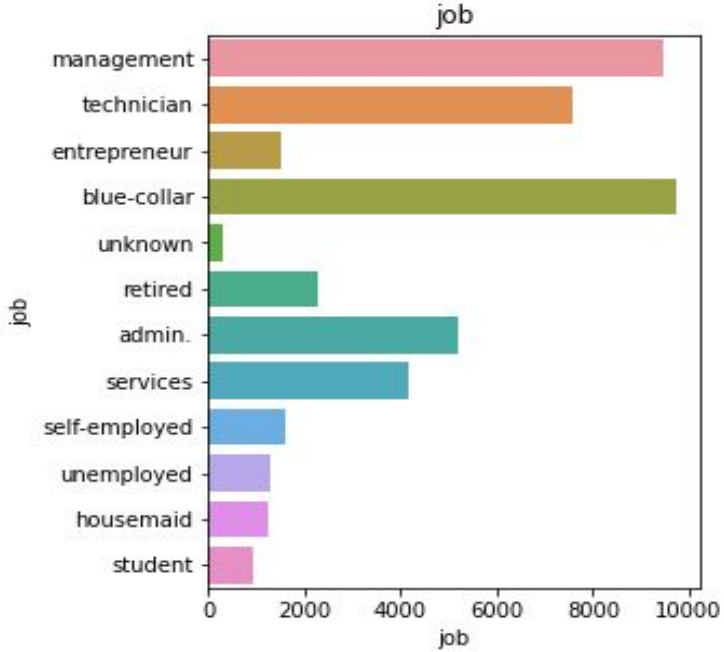


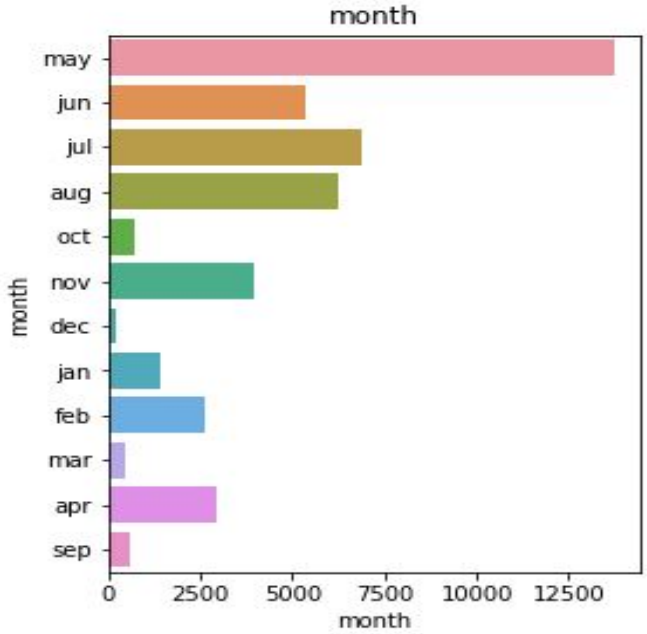




I. Most of the surveyed people have education level of secondary level.
II. Most of the contacts have cellular communication system.
III. Client who married are high in records in given dataset and divorced are less.

Counting categorical features

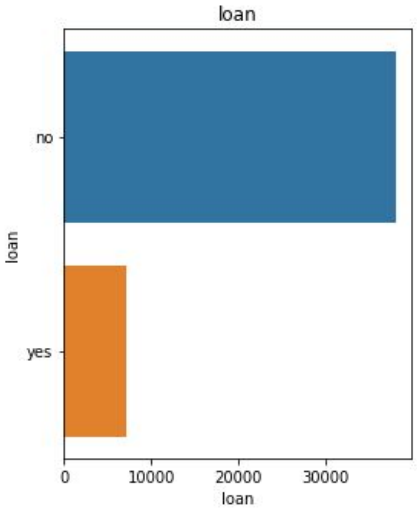


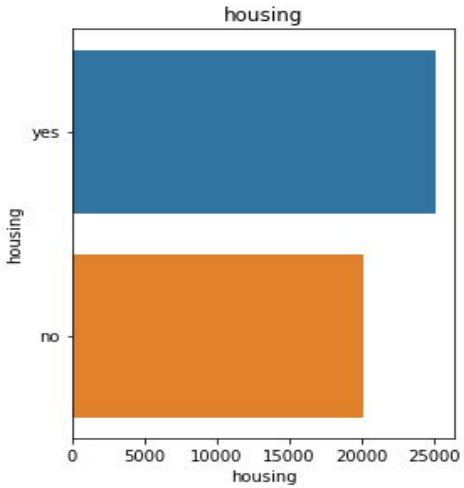


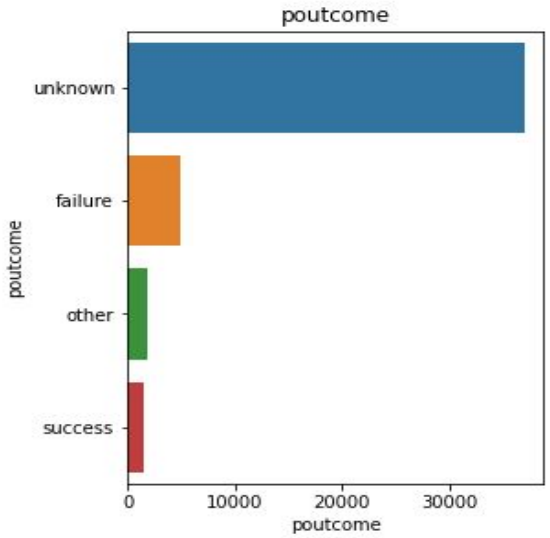
I. Most of the clients are doing blue-collar jobs i.e. hard manual labor works typically agriculture or manufacturing or construction or mining etc.

II. Data in month of May is high and less in December.

Counting categorical features

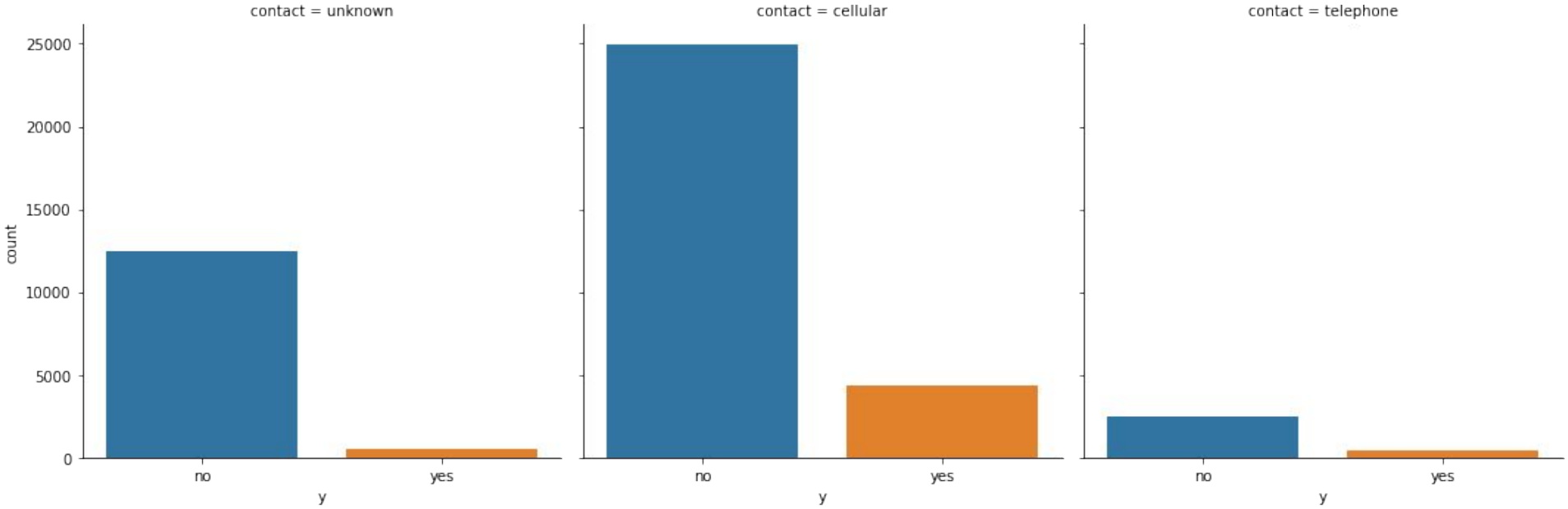






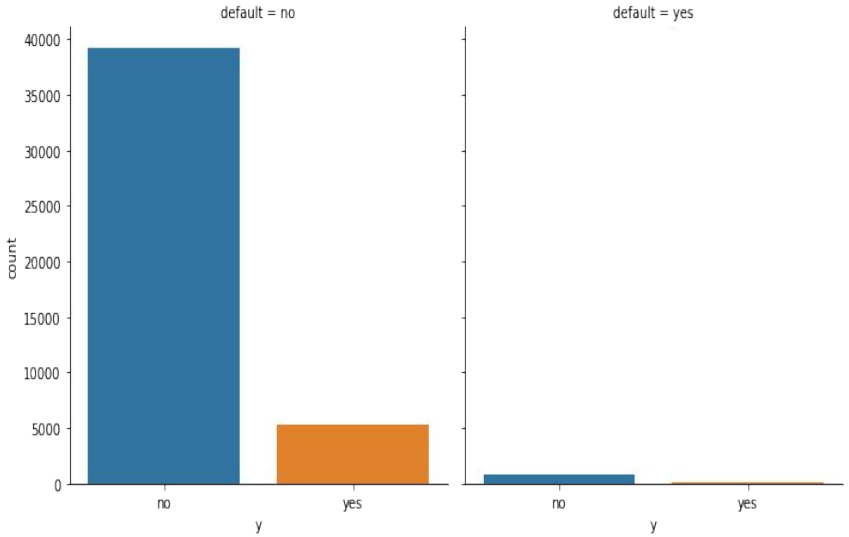
I. Most of the clients does not have personal loan.
II. The difference between the number of clients who have housing loan and who do not have housing loan is marginal.
III. In most of the cases the outcome of previous campaign is unknown.

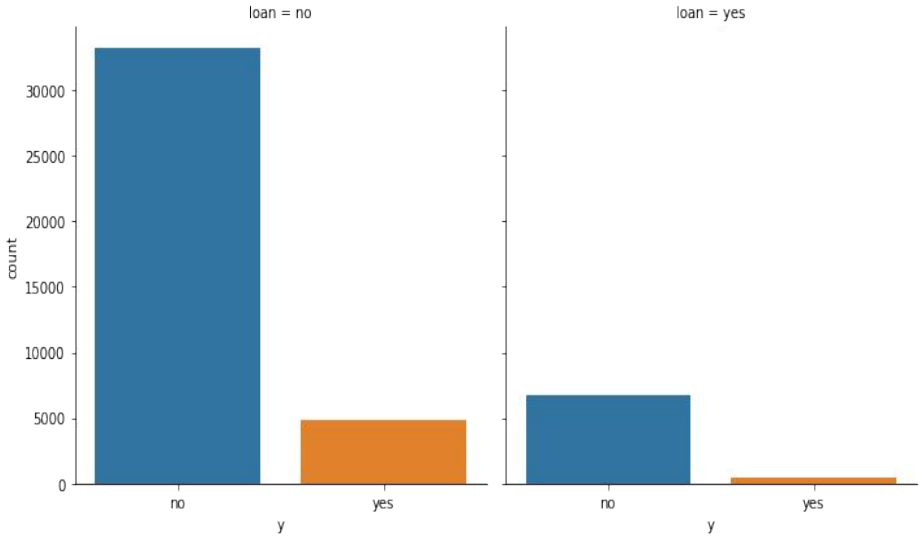
Relationship of categorical features with the output variable.



The contacts of type cellular have subscribed the term deposit mostly.

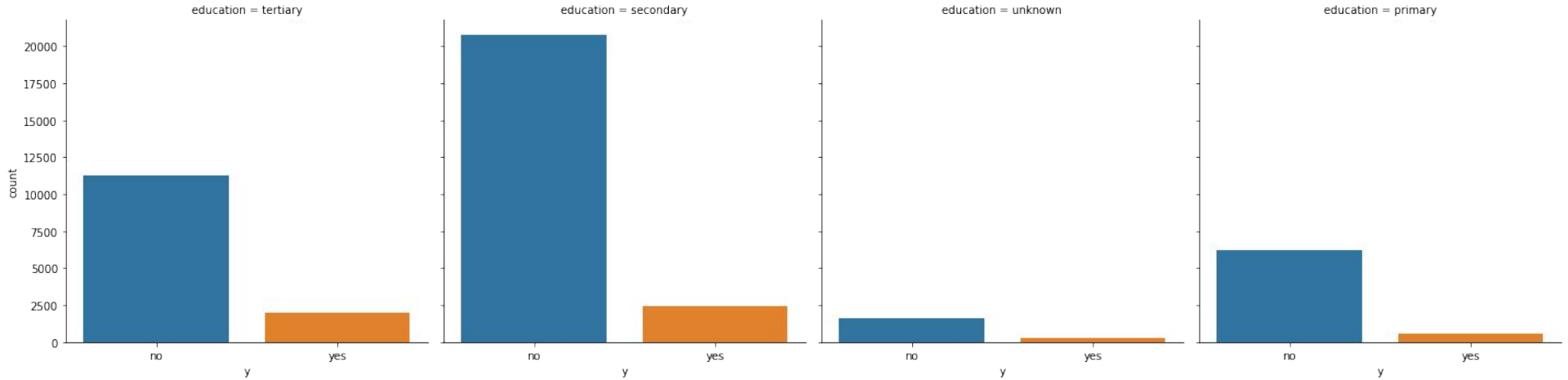
Relationship of categorical features with the output variable.





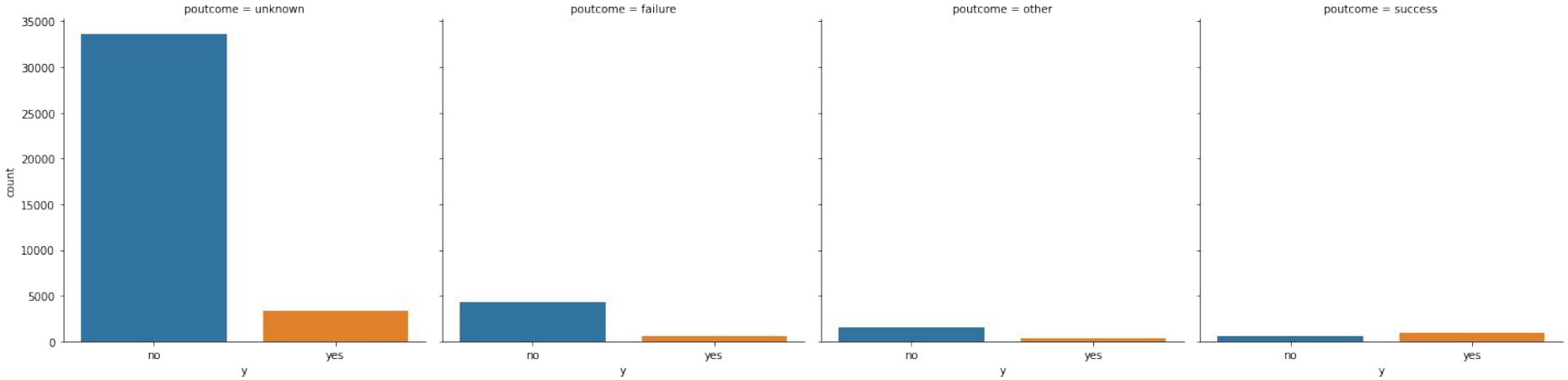
I. The clients who have no defaults subscribed the term deposit mostly.
II. The clients who does not have any outstanding loan amount have subscribed mostly.

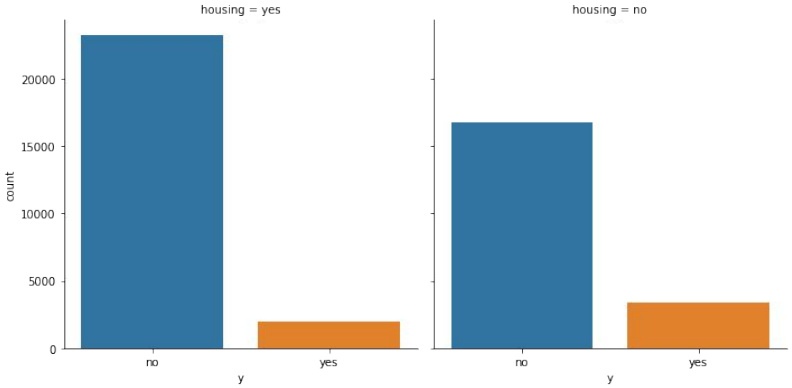
Relationship of categorical features with the output variable.



Clients who have educational qualification of secondary standard have subscribed the term deposit followed by the clients who have educational qualification of tertiary education

Relationship of categorical features with the output variable.

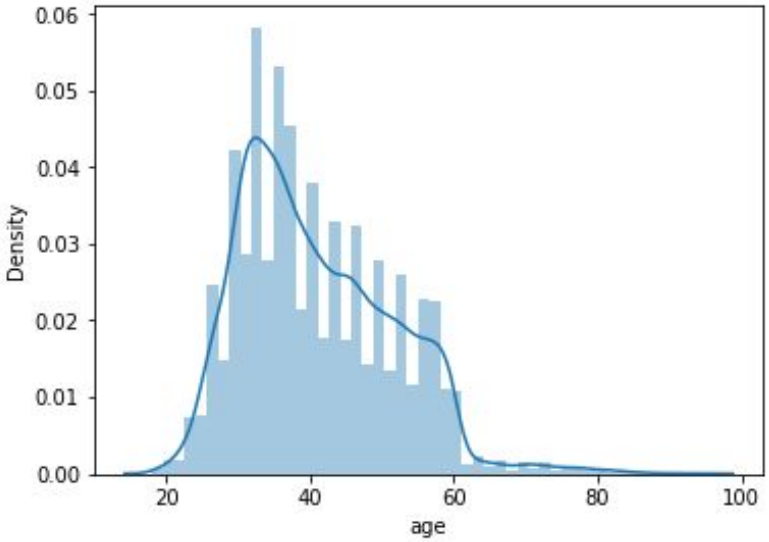


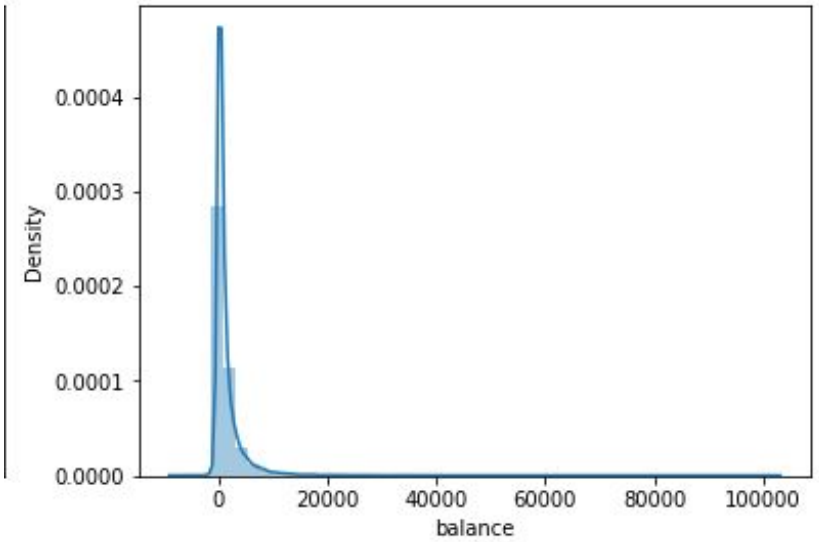


I. Clients with unknown previous campaign outcome said yes to subscribe the term deposit.

II. Clients with no housing loan subscribed the term deposit mostly.

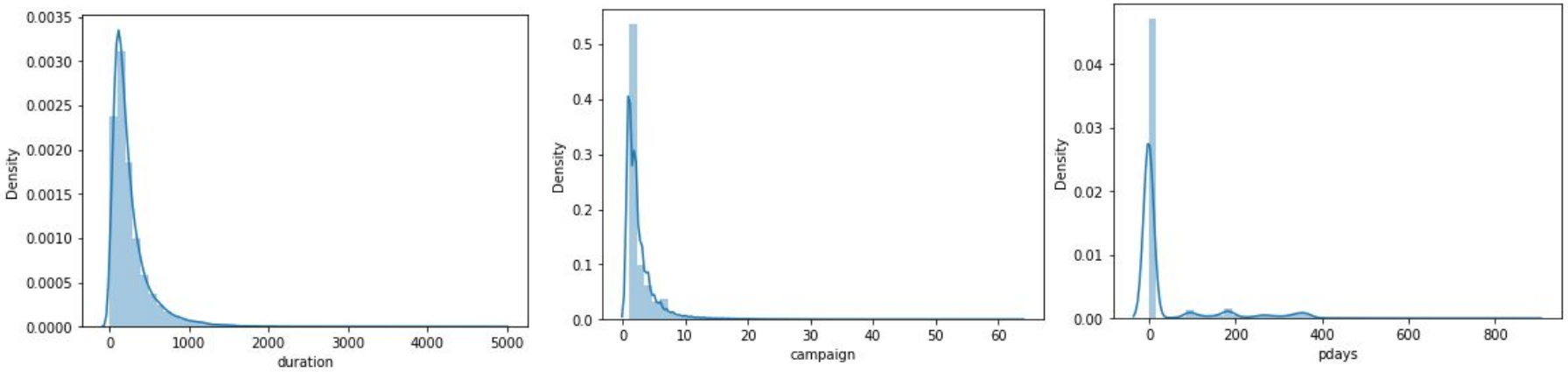
Distribution of numerical variables.





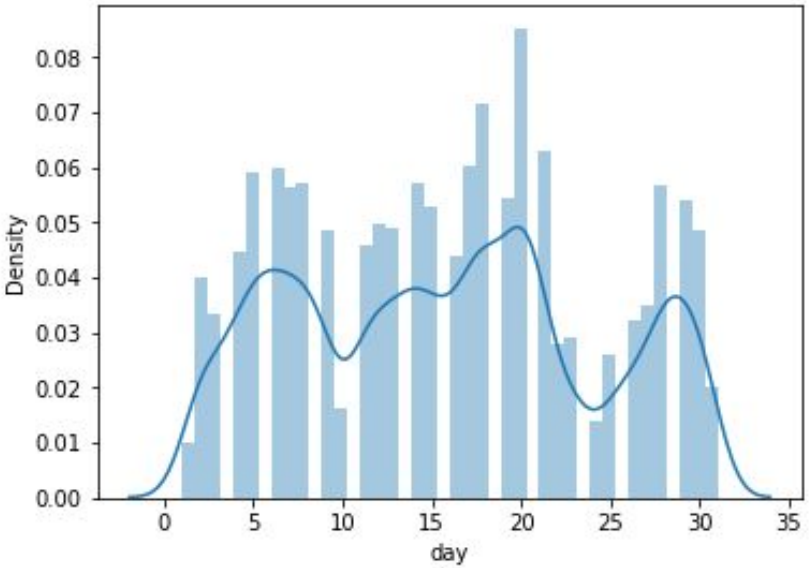
Distribution of age is seemed to be normally distributed but have outliers in it.
Distribution of balance is left skewed and seemingly have outliers.

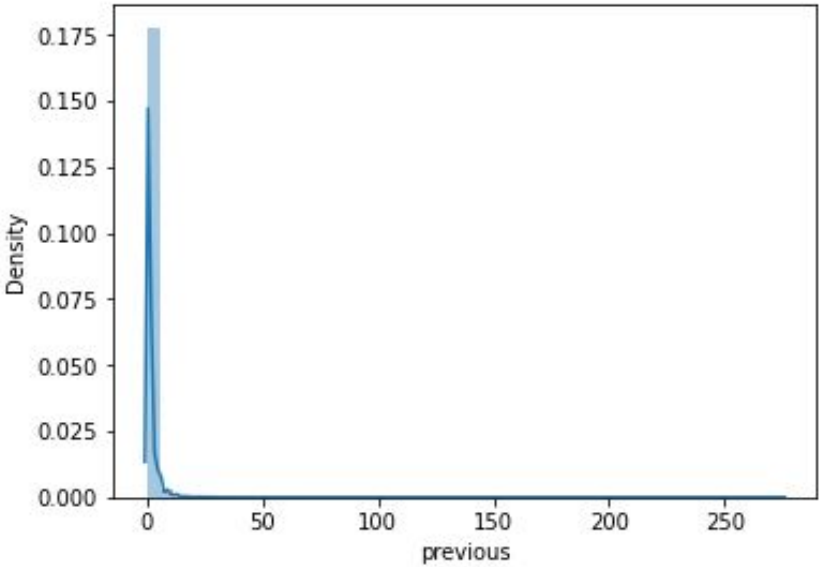
Distribution of numerical variables.



Duration, campaign and pdays variables are left skewed and probably have outliers.

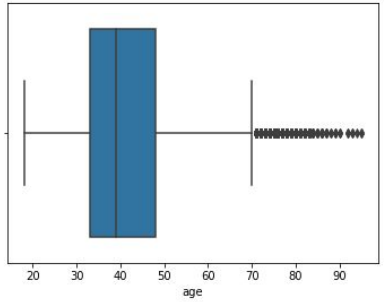
Distribution of numerical variables.

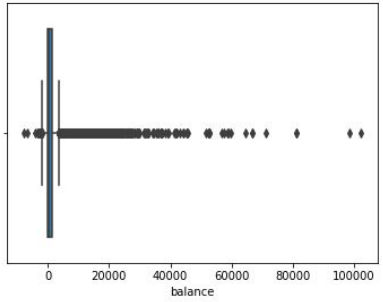


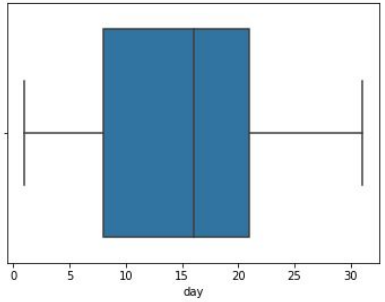


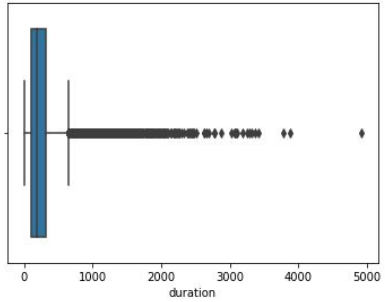
Day column is almost normally distributed.
Previous column is left skewed and probably have outliers.

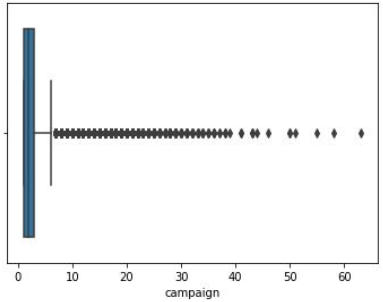
Outliers detection.

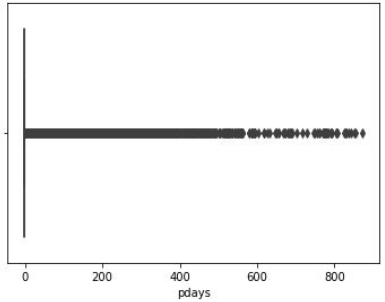


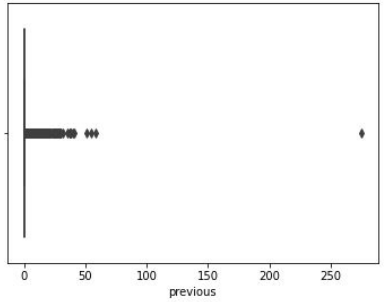












Apart from day column all the numerical columns have outliers. So we will try to remove the outliers using Z score.

Outliers handling

```
from scipy import stats
import numpy as np
z = np.abs(stats.zscore(df[['age','balance','duration','campaign','pdays','previous']]))
print(z)
df=df[(z<3).all(axis=1)]
df.shape
```

```
[[1.60696496 0.25641925 0.0110161  0.56935064 0.41145311 0.25194037]
 [0.28852927 0.43789469 0.41612696 0.56935064 0.41145311 0.25194037]
 [0.74738448 0.44676247 0.70736086 0.56935064 0.41145311 0.25194037]
 ...
 [2.92540065 1.42959305 3.37379688 0.72181052 1.43618859 1.05047333]
 [1.51279098 0.22802402 0.97014641 0.39902023 0.41145311 0.25194037]
 [0.37068857 0.52836436 0.39932797 0.24656035 1.4761376  4.52357654]]
(40209, 15)
```

We have tried to remove outliers using Z score.

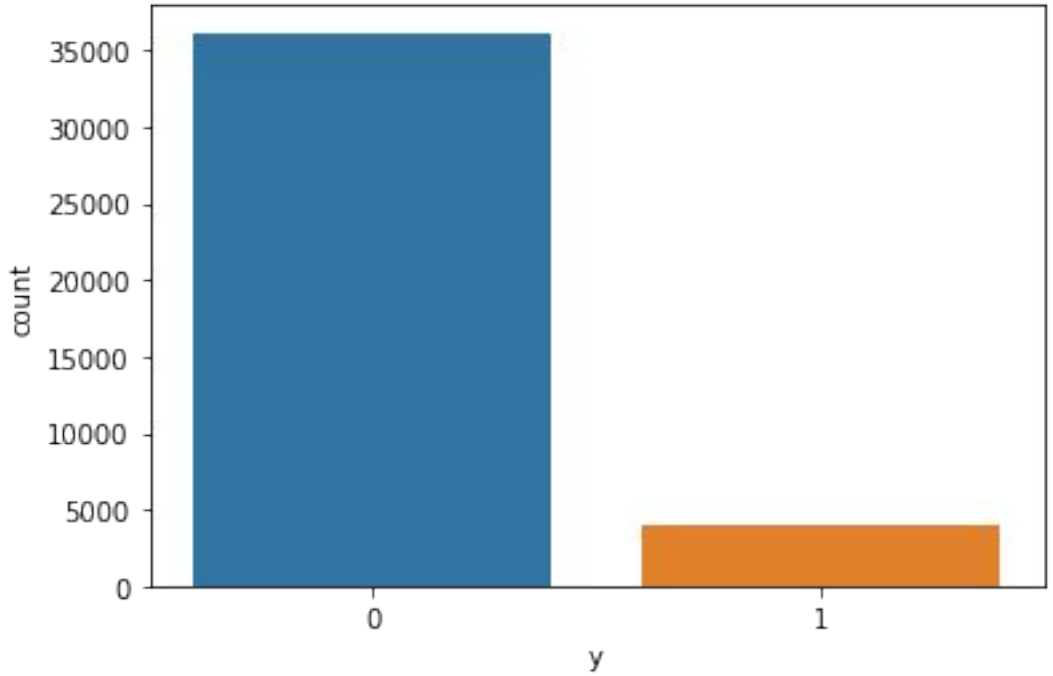
Feature engineering

- Drop unwanted Features
- Handle Missing Values
- Handle Categorical Features
- Handle Feature Scaling
- Remove Outliers

As per Exploratory Data Analysis EDA,

- no missing value found
- no feature found with one value
- 9 categorical features
- default features does not play imp role
- it seems some outliers found (age, balance, duration, campaign, pdays and previous has some outliers)

Preparing dataset for modelling



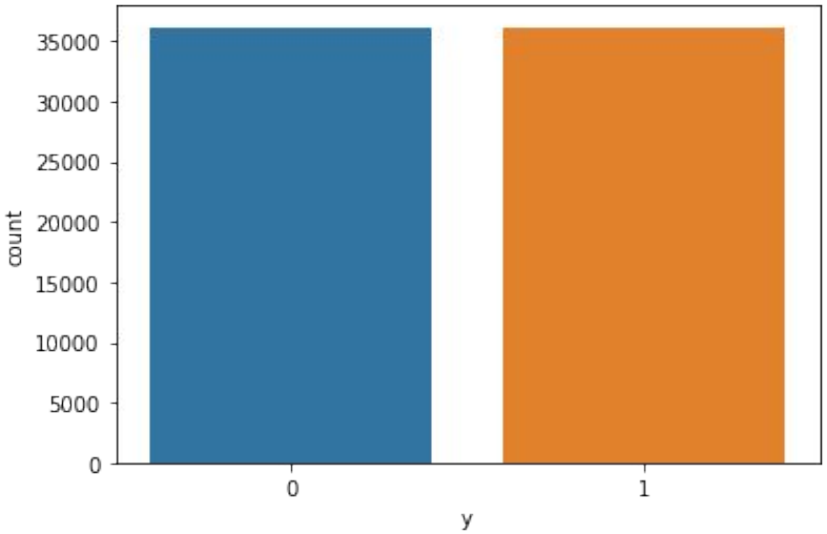
The dataset is not balanced properly so we will try RandomOverSampler to balance the dataset.

Preparing dataset for modelling

```
from imblearn.over_sampling import RandomOverSampler

os =  RandomOverSampler()
x_new,y_new=os.fit_resample(x,y)
```

```
from collections import Counter
print('Original dataset shape {}'.format(Counter(y)))
print('Resampled dataset shape {}'.format(Counter(y_new)))
sns.countplot(y_new)
plt.show()
```

```
Original dataset shape Counter({0: 36155, 1: 4054})
Resampled dataset shape Counter({0: 36155, 1: 36155})
```



After applying RandomOverSampler the output data is balanced now.

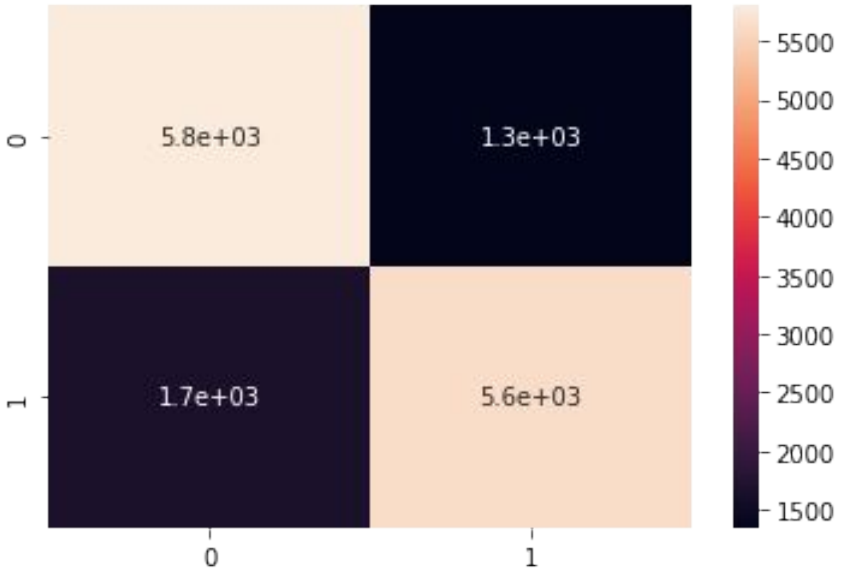
Preparing dataset for modelling

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix,roc_auc_score

#dividing the dataset into training and testing
X_train,X_test,y_train,y_test=train_test_split(x_new,y_new,test_size=.20,random_state=0)
print(X_train.shape,X_test.shape,y_train.shape,y_test.shape)

#feature scaling
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)
```

```
(57848, 14) (14462, 14) (57848,) (14462,)
```

Splitting the dataset into Train and Test data

Applying Logistic Regression

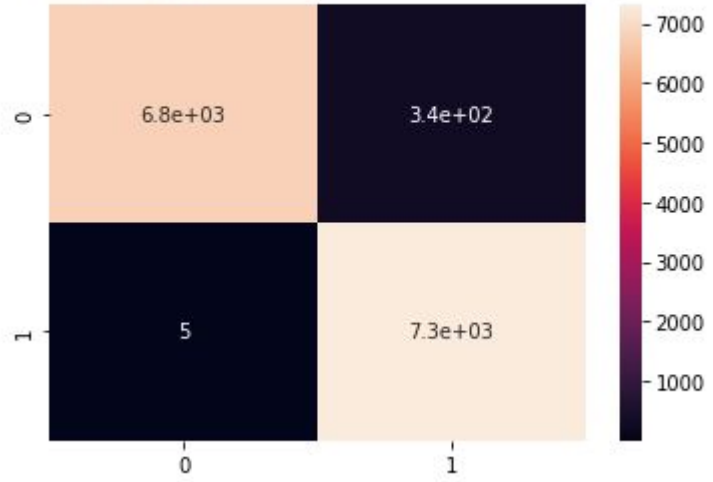


	precision	recall	f1-score	support
0	0.81	0.78	0.80	7455
1	0.78	0.81	0.79	7007
accuracy			0.79	14462
macro avg	0.80	0.80	0.79	14462
weighted avg	0.80	0.79	0.79	14462

Logistic Regression has an accuracy of 79%

Applying Random Forest classifier

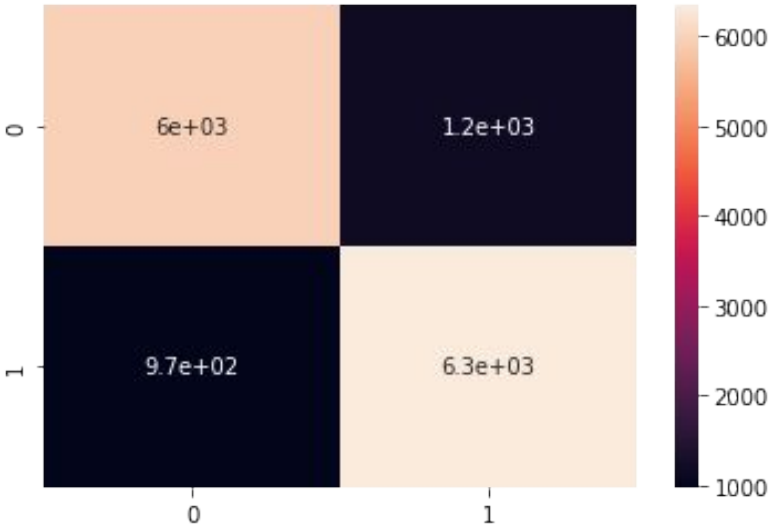
```
ROC_AUC Score: 0.9770984396993259
[[6802  345]
 [   5 7310]]
```



	precision	recall	f1-score	support
0	0.95	1.00	0.98	6802
1	1.00	0.95	0.98	7660
accuracy			0.98	14462
macro avg	0.98	0.98	0.98	14462
weighted avg	0.98	0.98	0.98	14462

Random Forest has accuracy of 98%

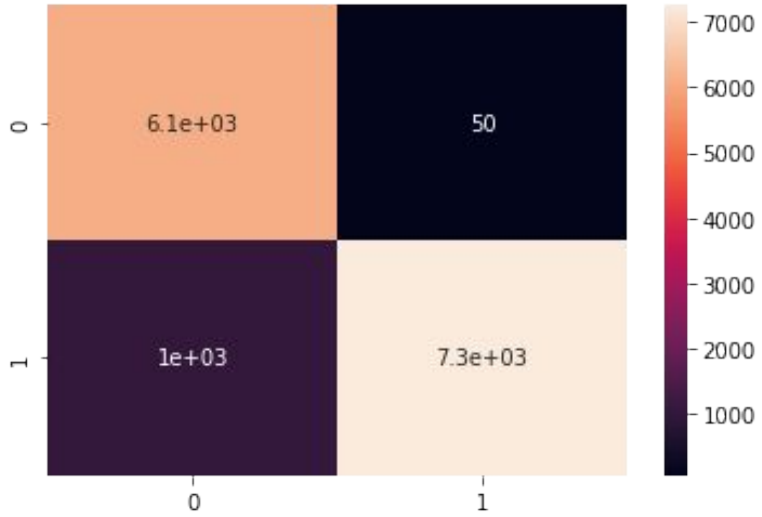
Applying XGB classifier



	precision	recall	f1-score	support
0	0.84	0.87	0.86	6856
1	0.88	0.85	0.86	7606
accuracy			0.86	14462
macro avg	0.86	0.86	0.86	14462
weighted avg	0.86	0.86	0.86	14462

XGB Classifier has accuracy of 86%

Applying KNN classifier



	precision	recall	f1-score	support
0	0.86	0.99	0.92	6182
1	0.99	0.88	0.93	8280
accuracy			0.93	14462
macro avg	0.93	0.93	0.93	14462
weighted avg	0.94	0.93	0.93	14462

KNN classifier has accuracy of 93%

Cross validation

```
[{'best_params': {'criterion': 'entropy',
    'max_depth': 3,
    'max_features': 'auto',
    'n_estimators': 50},
   'best_score': 0.797918781699342,
   'model': 'RandomForestClassifier'},
  {'best_params': {'learning_rate': 0.5, 'max_depth': 20, 'n_estimators': 200},
   'best_score': 0.9670688641067053,
   'model': 'XGBClassifier'}]
```

Cross validation on Random forest And XGB classifier

Model selection

	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic regression	0.796985	0.780314	0.811141	0.795429	0.797355
Randomforest	0.975799	0.999316	0.954931	0.976620	0.977098
KNNeighbors	0.925598	0.992344	0.876797	0.930999	0.933870
XGB Classifier	0.860393	0.881887	0.848146	0.864687	0.861063

Randomforest classifier is giving higher accuracy compared to other model so we will use this model in our data.

Conclusion

This sums up for the classification task of bank marketing dataset. We find that RandomForest gives us the best value for accuracy which is 0.98 while KNNClassifier gives us the second best accuracy value. The best AUCscore of 0.97 comes from RandomForest followed by KNN classifier.

The results of RandomForest and KNNClassifier are better while rest of the algorithms are giving more or less same result with minor differences.

As per algorithms importance of whether client uses cellular phone or not and the month in which client is being called play a vital role and the strategies of marketing campaign should be decided accordingly.