CAPSTONE PROJECT

BANK MARKETING EFFECTIVENESS PREDICTION

(SUPERVIED MACHINE LEARNING CLASSIFICATION)

TEAM MEMBERS

Aditya Tadas
Nikhil Machave
Aishwarya Methe

AGENDA

- PROBLEM STATEMENT
- BUSINESS UNDERSTANDING
- FEATURE ANALYSIS
- DATA SUMMARY
- HANDLING MISSING AND DUPLICATE VALUES
- UNIVARIENT ANALYSIS
- BIVARIENT AND MULTIVARIENT ANALYSIS
- FEATURE ENGINEERING
- FEATURE SELECTION
- FREQUENCY COUNT ENCODING AND ONE HOT ENCODING
- FINDING COVARIENCE OF VARIABLES
- HANDLING CLASS IMBALANCE USING SMOTE
- SPLITTING DATA INTO TRAIN AND TEST
- FITTING CLASSIFICATION MODEL
- HYPERTUNING OF BEST FIT MODEL
- EXPLAIN FEATURE IMPORTANCE USING SHAPASH MODEL EXPLANATORY
- CONCLUSION

PROBLEM STATEMENT

- 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)

BUSINESS UNDERSTANDING

- Bank marketing is the design structure, layout and delivery of customer-needed Services worked out by checking out the corporate objectives of the bank and environmental constraints.
- A term deposit is fixed-term investment that include the deposit of money into an account at financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.
- The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases ,the account holder may allow the investor early termination or withdrawal if they give several days notification .Also, there will be a penalty assessed for early termination.

FEATURE ANALYSIS

- Age:(numeric)(Age of the person)
- 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 windowed)
- Education: categorial:basic.4y,basic.6y,basic.9y,’high school, ’illiterate’, professional course', university degree', 'unknown’
- 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’) retained with the last contact of the current campaigns:
- Contact: contact communication type(categorical:’ cellular, telephone’)
- Month: last contact month of year (categorical: jan, feb, mar, nov , dec)
- Campaign: number of contacts performed during this campaign and for this client (numeric, includes)
- duration: last contact duration in seconds (numeric).Important note: this attribute highly affected the output Target(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 obvious known thus this input should only be included for benchmark purposes and should be discarded if the intension is to have a realistic predictive model
- 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)
- Balance: (numeric) account balance of this client
- Day: (numeric) day of the week
- Poutcome: outcome of the previous marketing campaign (categorical: failure non existent success social and economic context attributes
- Y: has the client subscribed a term deposits?(binary, yes ,no)

DATA SUMMARY

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

- Here we have summary of our dataset
- There are 6 numerical variable present in our dataset i.e age,balance,day,duration,campaign,pdays,previous.
- There are 9 categorical variable present in our dataset which are job,marital,education,default,housing,loan,contact,month,poutcome.
- Our target variable is Binary class variable i.e ‘y’.

FINDING MISSING AND DUPLICATES VALUES

```
bank_df.isna().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
```

```
bank_df.shape
```

```
(45211, 17)
```

```
bank_df.duplicated().sum()
```

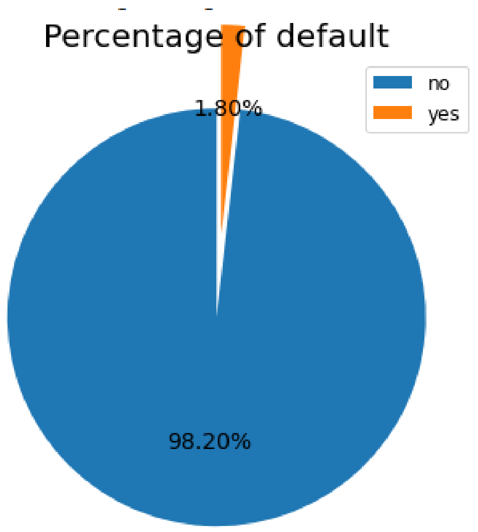
```
0
```

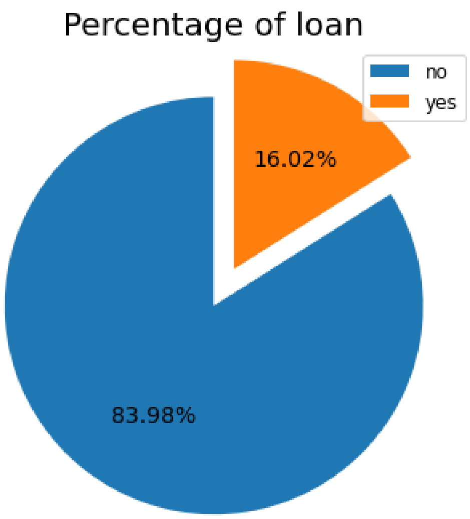
```
bank_df['y'].value_counts()
```

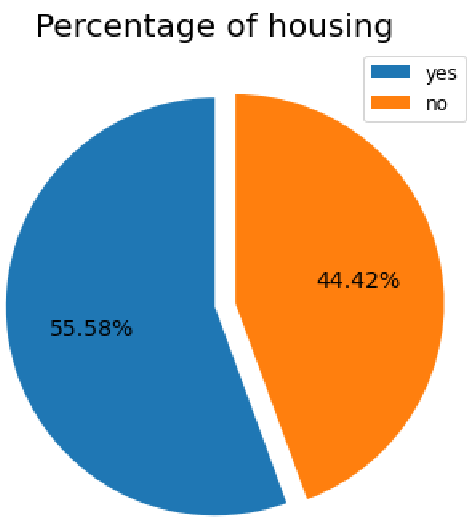
```
no     39922
yes     5289
Name: y, dtype: int64
```

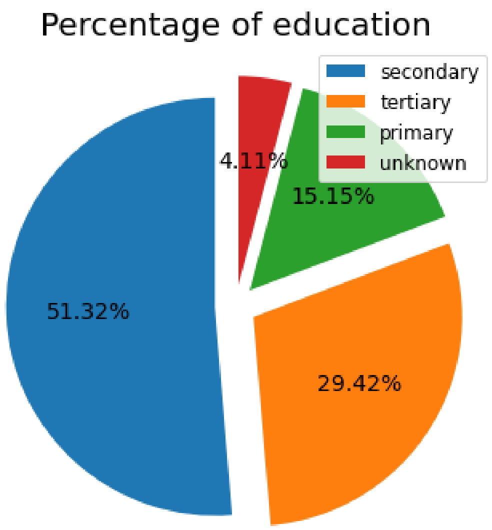
- Data contains 45211 records and 17 columns.
- There are no null values present in our dataset.
- There are no duplicated value present in our dataset.
- As we seen in the column 'y' which is out target variable there are very high class imbalance in that column so we have to done some oversampling or Undersampling method to overcome class imbalance problem.

UNIVARIENT ANALYSIS ON CATEGORICAL COLUMNS









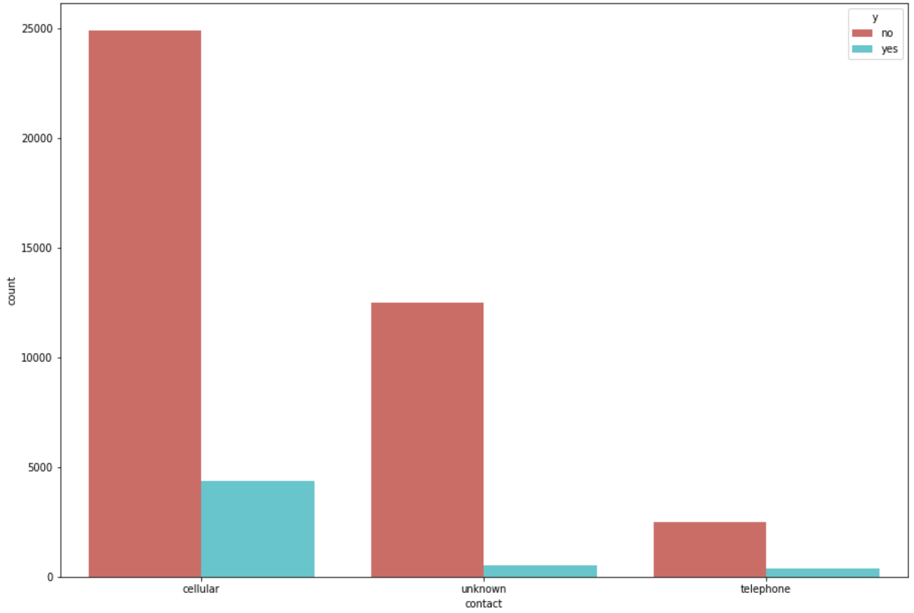
- As we can see that in the pie plot of percentage of default , Most number of clients in our dataset does not having default , 98.20% of clients have not done any default only 1.80 % clients are default.

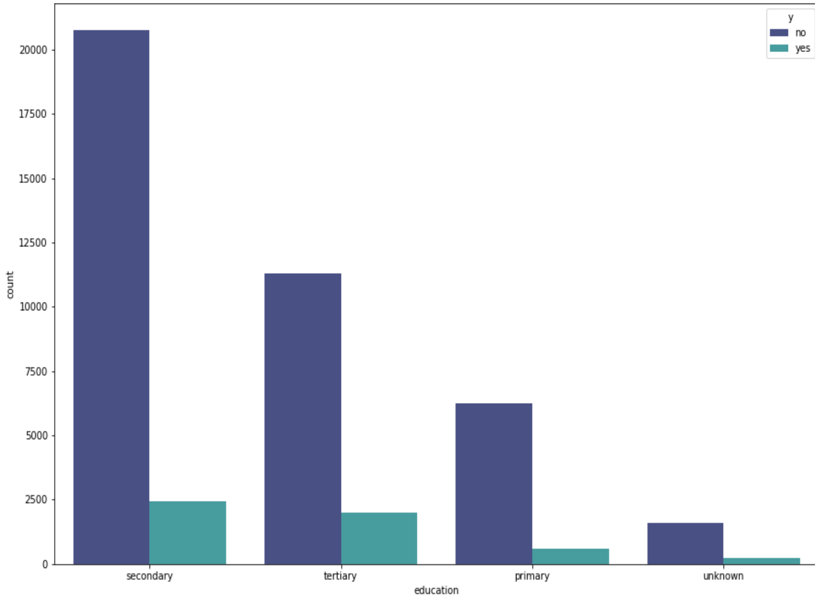
- In the loan pie plot , 83.98 % clients are not taking any personal long only 16.02 % of clients having personal loan.

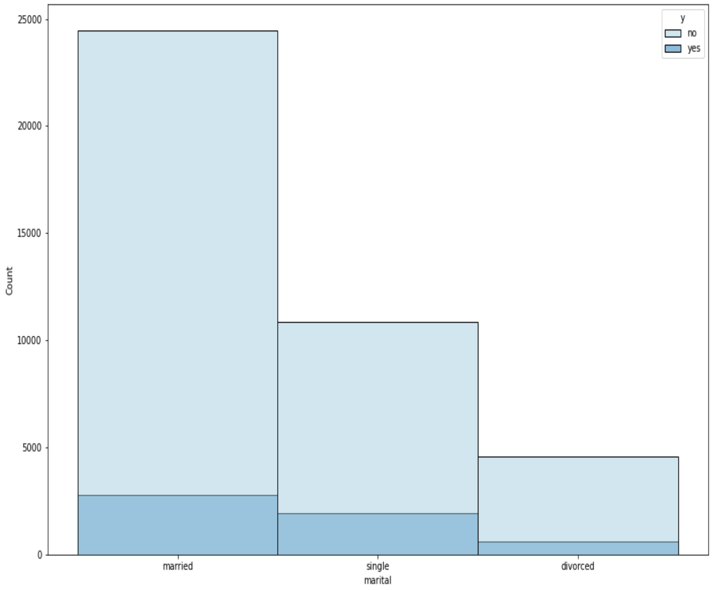
- In a pie plot of Housing, roughly above 50 % of clients have taken housing loan so we can say that half of the client taken housing loan.

- Among all the clients most of the clients have done secondary education , education of 4.11% of clients is unknown.

BIVARIENT ANALYSIS ON CATEGORICAL COLUMNS

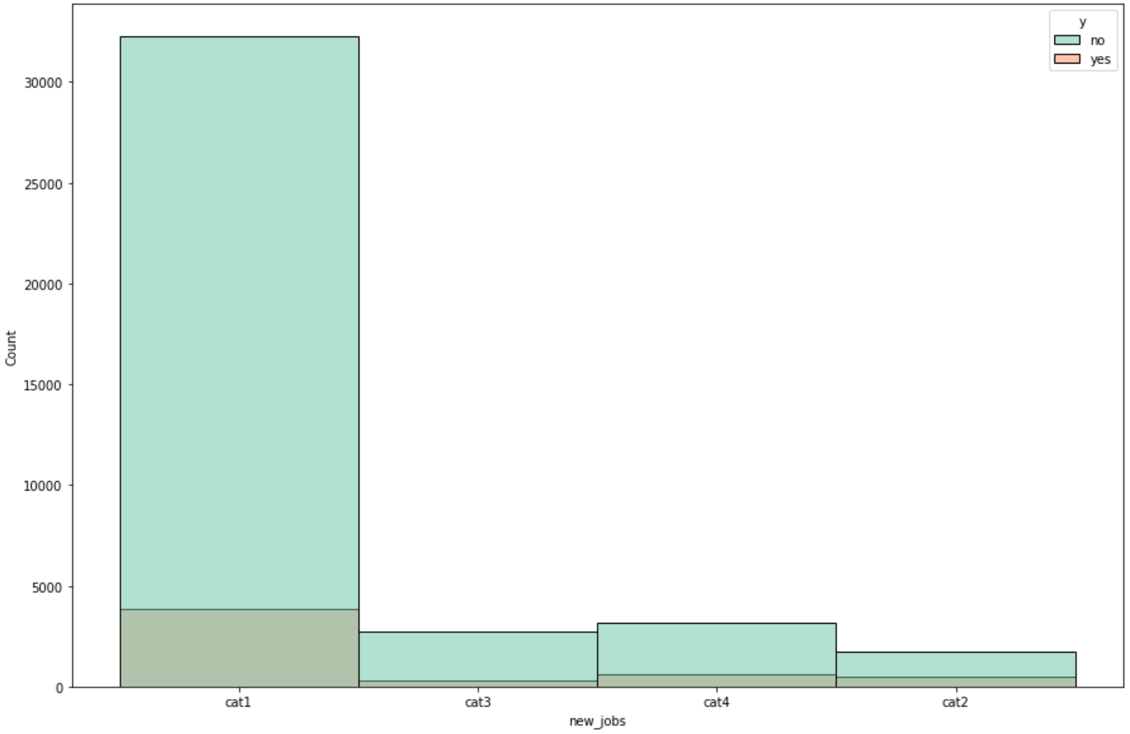






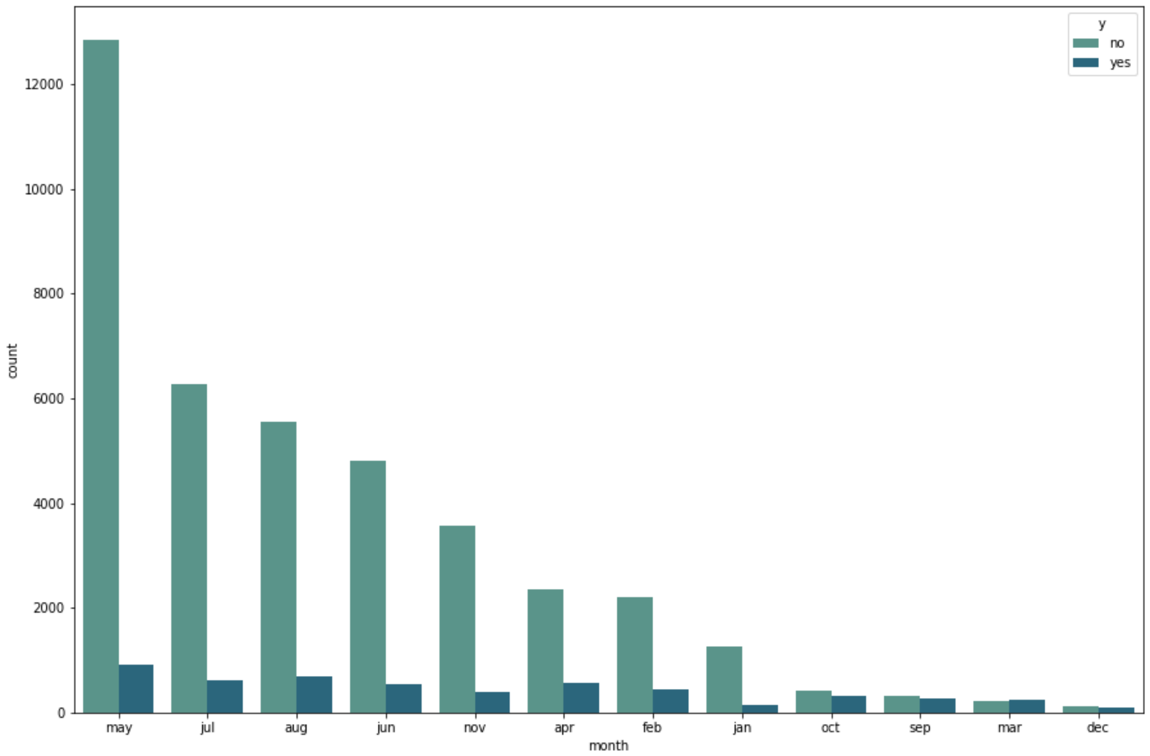
- Most of the clients contacted through cellular after that least clients are contacted through telephone there are some records having unknown contact also.
- Most of the clients who contacted through cellular agree to subscribe for the term Deposit, Very less clients which are contacted through telephone agree to subscribe for term deposit.
- Most of the clients having education secondary and terittary agree to subscribe for term deposite, very less clients having primary and unknown education subscribe for term deposit.
- Most of the clients who’s marital status is married agree to subscribe for term deposit after that single marital status clients agree to subscribe for term deposit but when the marital status of client is divorced then those clients have very less possibility to subscribe for term deposite

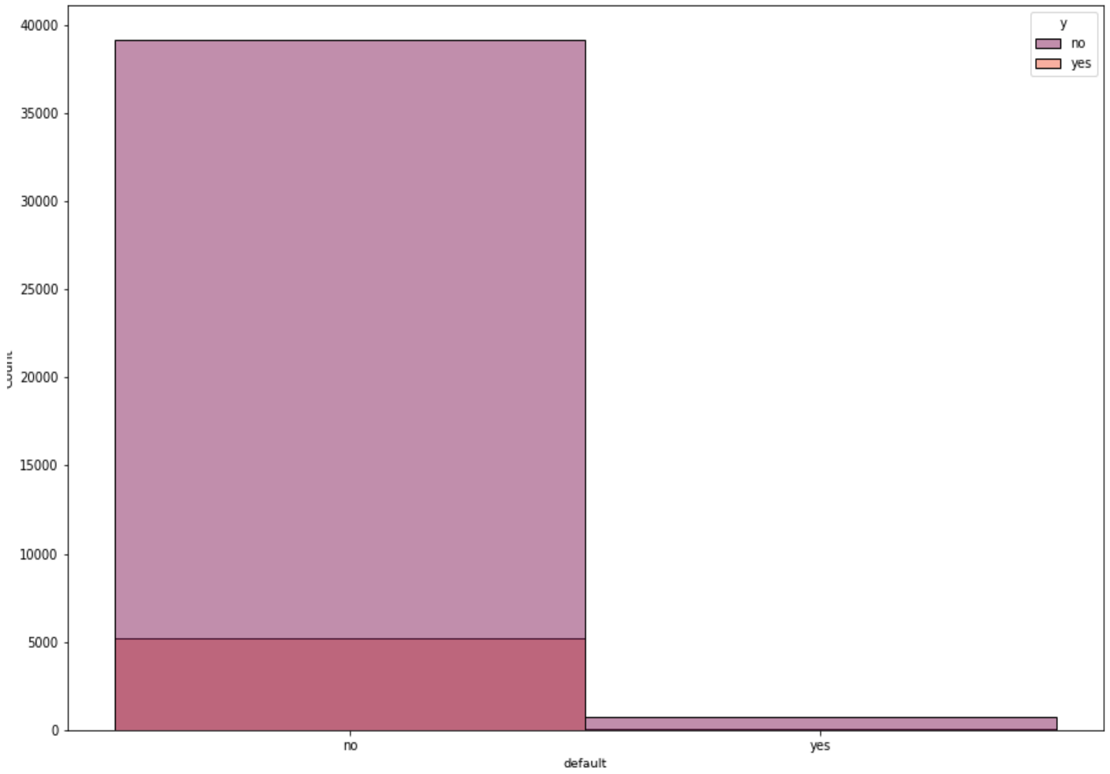
FEATURE ENGINEERING




```python
#converting job column into 4 categories
def cluster_job(job):
    cat_1=['blue-collar','management','technician','admin.','services']
    cat_2=['retired']
    cat_3=['self-employed','entrepreneur']
    cat_4=['unemployed','housemaid','student','unknown']

    if job in cat_1 :
        return 'cat1'
    if job in cat_2 :
        return 'cat2'
    if job in cat_3 :
        return 'cat3'
    if job in cat_4 :
        return 'cat4'
    return job
```

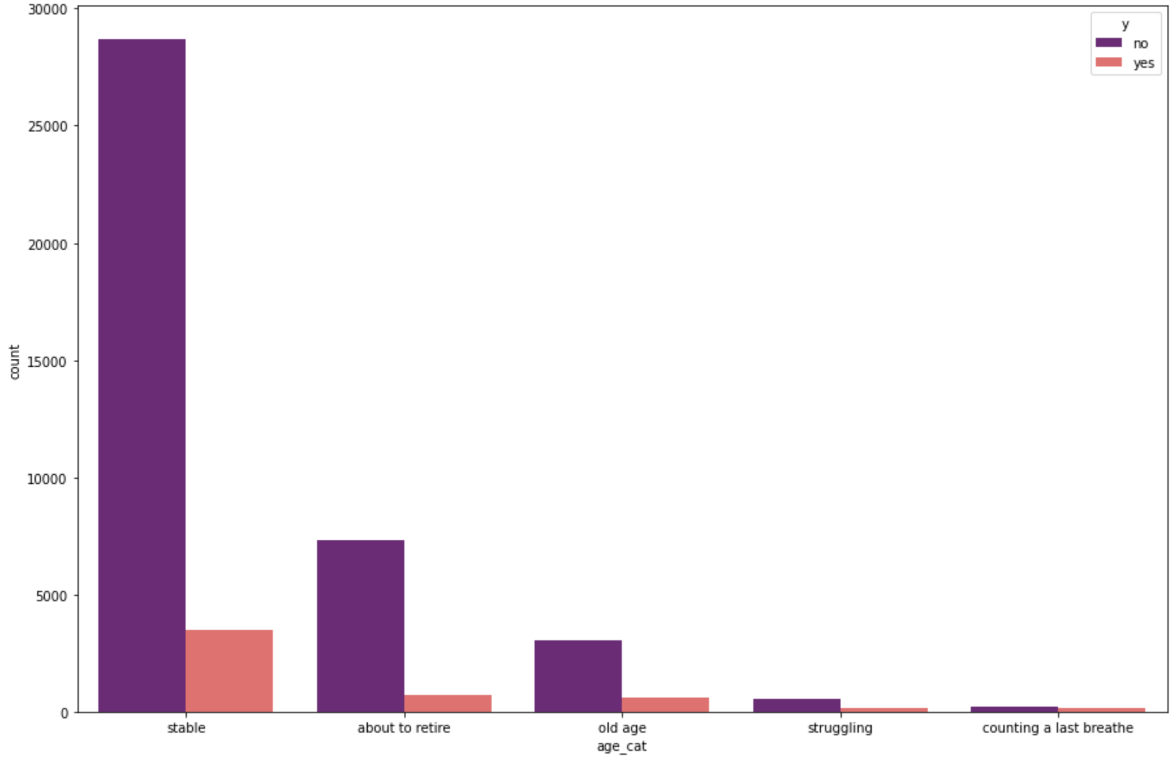
- We have done some feature engineering for job columns we have created 4 categories according to their attributes
- Job name blue collar , management , technician , admin and services are grouped as category 1.
- Retired client are named as category 2.
- Self employed and entrepreneur grouped as category 3.
- Unemployed , housemaid , student and unknown are grouped as category 4.
- Soo from the above categories we have seen that when the client subscribed for term deposit mostly belonging to category 1.





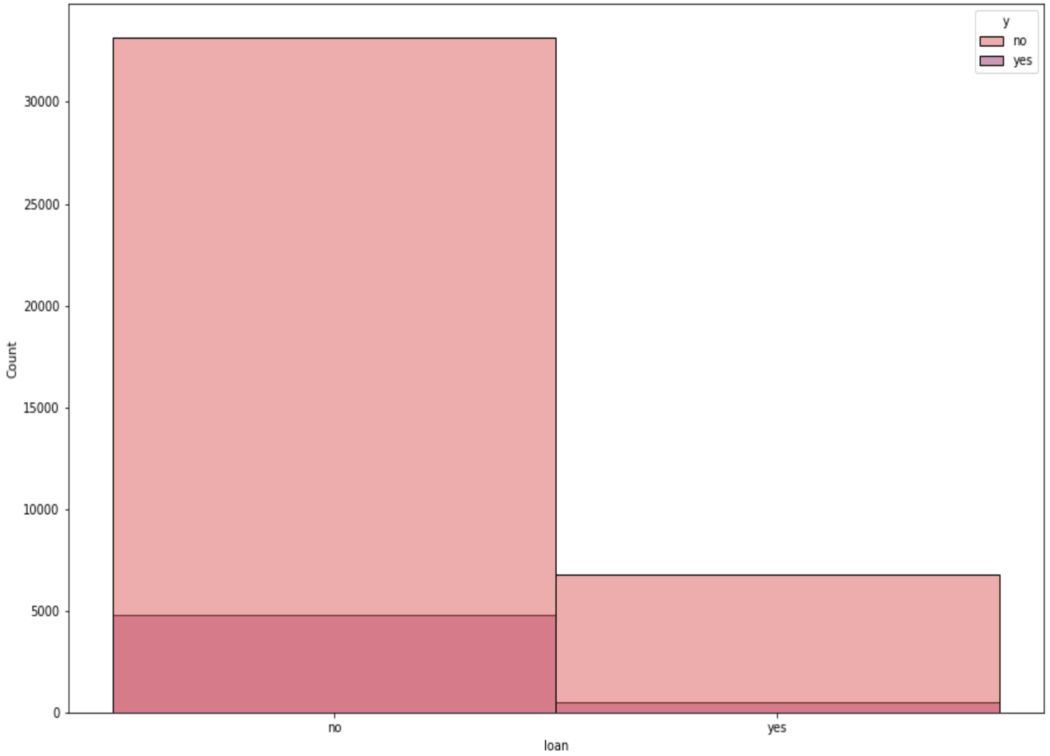
- In the month may, jul, aug, jan most of the clients subscribed for term deposit but in the month dec , mar , sep very less clients subscribed for term deposit.

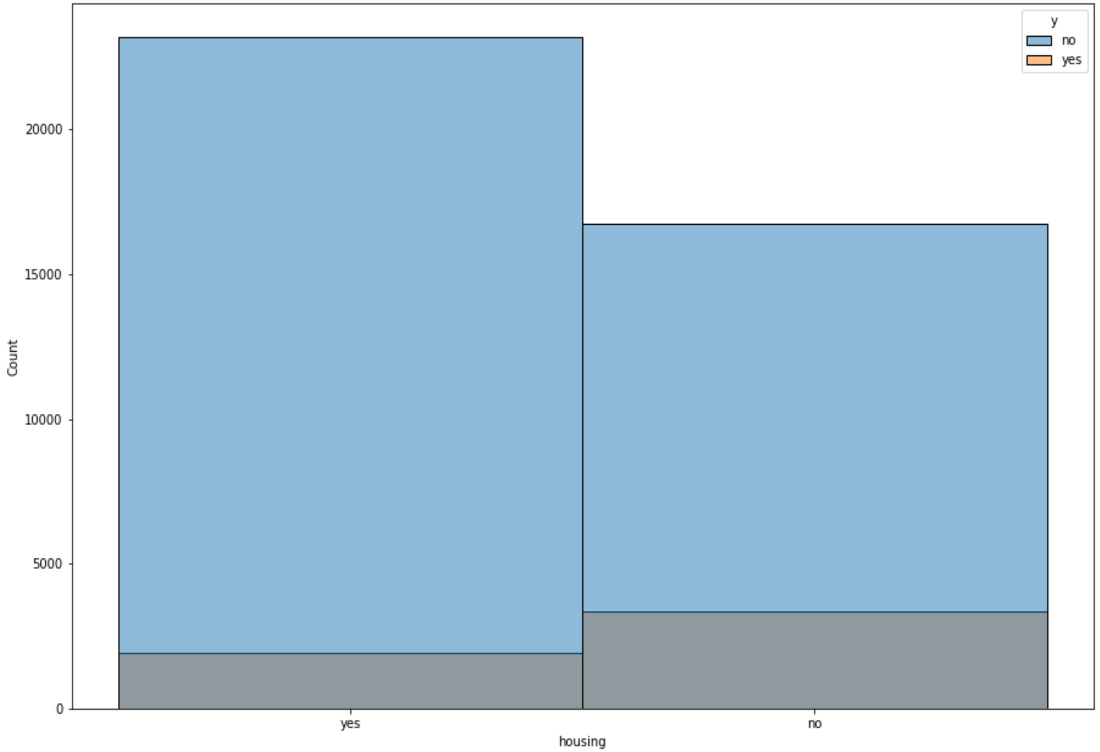
- When client not done any default those clients are more likely to subscribe for term deposit.



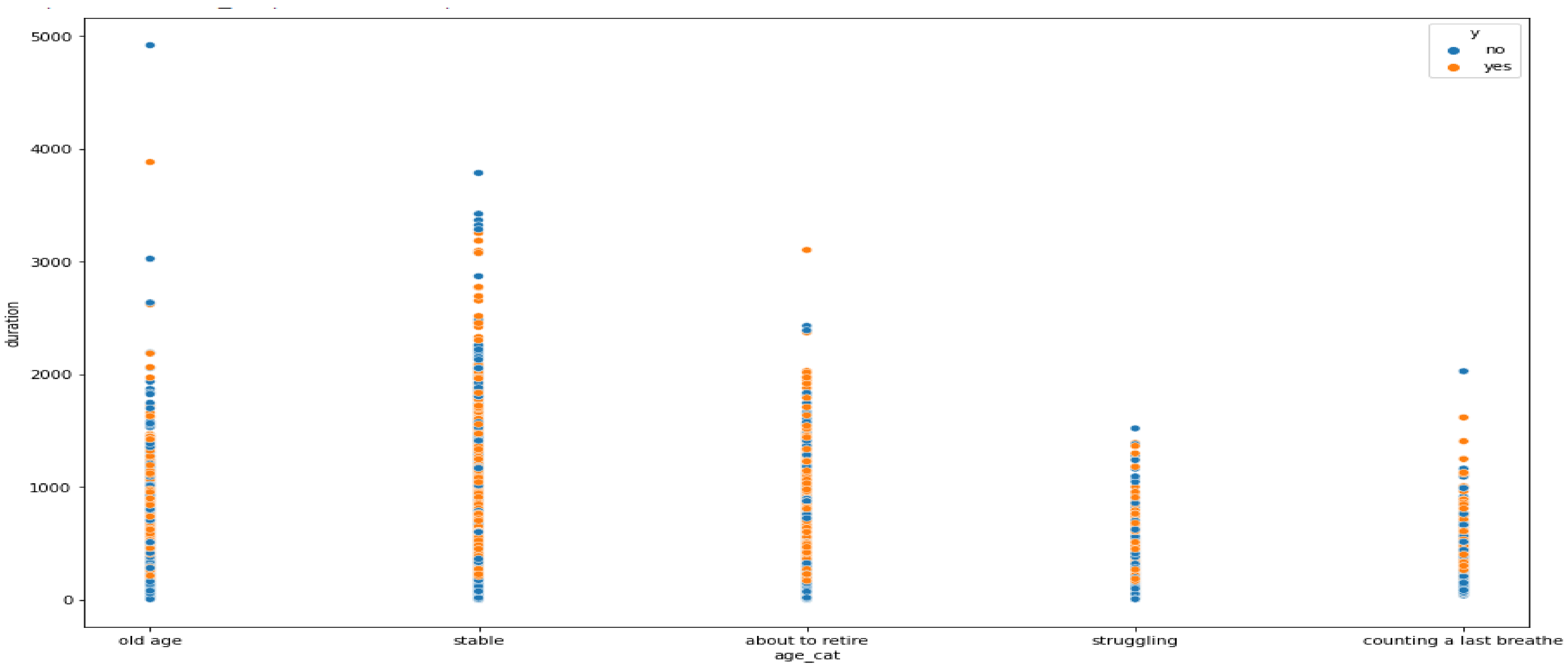

```python
# converting age column to categorical column by assinging categories
def convert_age(age):
    if age < 25 :
        return 'struggling'
    elif age < 48 :
        return 'stable'
    elif age < 57 :
        return 'about to retire'
    elif age < 72:
        return 'old age'
    else:
        return 'counting a last breathe'
```

- According to the age we have categories age column into 5 categories .
- When age of the clients is below 25 they belong to struggling category , when the age of clients is 25-47 and 48-57 they belong to category stable and about to retire respectively.
- When the age of the client is 57-72 they belong to old age category after that all are greater than 72 age they belong to counting last breathe category.
- So from the above categories we have seen that when the category of client is stable then there is high possibility that those client agree to subscribe for term deposit after that about to retire and old age categories have more possibility.

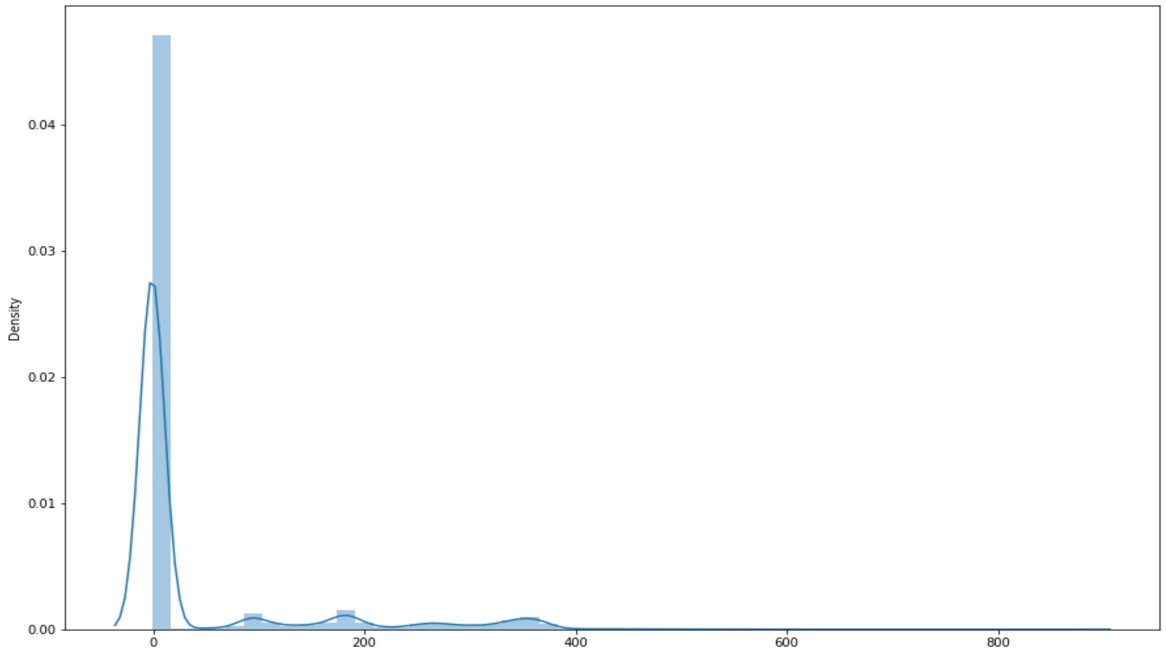




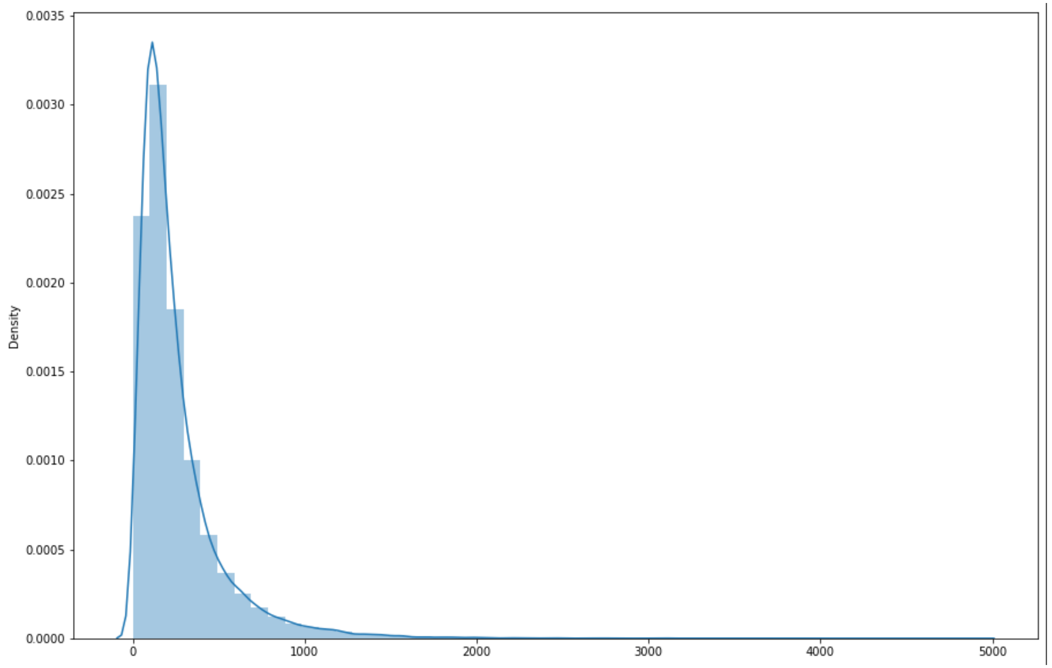
- Client who has not having personal loan have high possibility that they agree to subscribe for term deposit but when clients having personal loan they rarely agree to subscribe for term deposit.
- There are somewhat equal possibility that when the client having housing loan or do not having personal loan agree to subscribe for term deposit so we can say that housing loan does not affect much to predict .



- When the age category is old age and stable then communication duration with these age category are higher that's why there is high possibility that old age and stable category clients having high possibility to subscribe for term deposit.
- When the age categories are struggling and counting last breathe then communication with those clients are less that's why there is very less possibilty that those categories agree to subscribe for term deposit.



pdays



duration

```
#droppping colums because we have extracted new features from that columns
bank_df.drop(columns=['age','pdays','duration','job'],axis=1,inplace=True)
```

- In pdays columns most the values are zero or below than 0 there are very less value who are above zero and these column contains very big outliers that effects the accuracy of our model also pdays does not affects on prediction of our model Soo we have dropped these column from our dataset.
- In the distribution plot of duration Confidance interval tents to zero most of the values are 0 in that columns also in problem statement it is given that we have to drop that column to train our model hence we dropped that columns from our dataset.
- Also age and job columns have been also dropped because we extracted other feature by using that columns.

CORRELATION HEATMAP

Frequency count encoding for month column

```
[167]  bank_df.month.value_counts().to_dict()
```

```
{'apr': 2932,
 'aug': 6247,
 'dec': 214,
 'feb': 2649,
 'jan': 1403,
 'jul': 6895,
 'jun': 5341,
 'mar': 477,
 'may': 13766,
 'nov': 3970,
 'oct': 738,
 'sep': 579}
```

```
[168]  # And now let's replace each label in months by its count

       # first we make a dictionary that maps each label to the counts
       bank_df_frequency_map = bank_df.month.value_counts().to_dict()
```

```
# Creating dummy variable for categorical variables- season, month, weekofdays, year, holidays, functional day
marital = pd.get_dummies(bank_df['marital'],prefix='marital')
contact = pd.get_dummies(bank_df['contact'], prefix='contact')
poutcome = pd.get_dummies(bank_df['poutcome'], prefix = 'poutcome')
age_cat = pd.get_dummies(bank_df['age_cat'],prefix = 'age_cat')
new_jobs = pd.get_dummies(bank_df['new_jobs'],prefix = 'new_job')
education = pd.get_dummies(bank_df['education'],prefix = 'educaton')
```

```
bank_df = pd.concat([bank_df,marital,contact,poutcome,age_cat,new_jobs,education],axis=1)
```

```
#seprating our dependent and independent features
y=(bank_df['y'])
x=bank_df.drop(columns=['y'],axis=1)
```

- We use frequency count encoding for month variable i.e we assign number to each month according to there counts.
- For other variable we have used on hot encoding method are create dummy variable
- After that we Seperating our dependent variable and independent variables to train model.

SMOTE Oversampling for handling class imbalance

```
[282]  #Dependent variable business treatment - Smote oversampling
```

```
[283]  from imblearn.over_sampling import SMOTE
       sampler=SMOTE()
       X ,y = sampler.fit_resample(x,y)
```

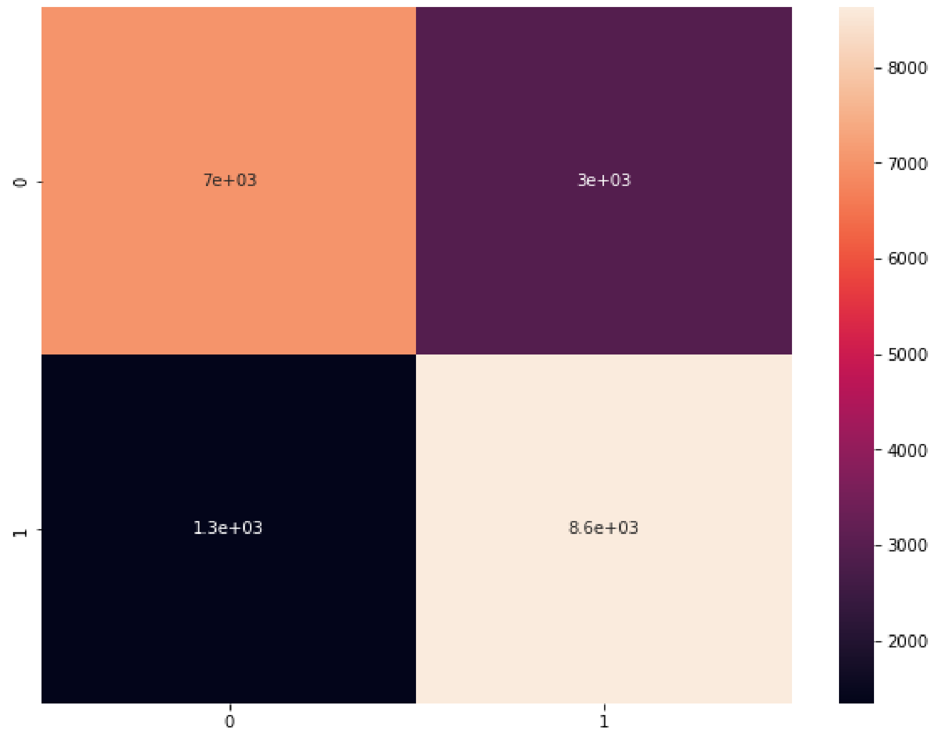
```
[284]  #Original length and Resampled Length
       print('Original Dataset length',len(x))
       print('Resampled Dataset length',len(X))

Original Dataset length 45211
Resampled Dataset length 79844
```

```
[189]  #loading required libraries and performing train test split by 75-25 ratio
       from sklearn.model_selection import train_test_split
       from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
       from sklearn.model_selection import cross_val_score,ShuffleSplit,cross_val_predict
       from sklearn.metrics import accuracy_score, roc_auc_score, roc_curve, log_loss, precision_score,
       from sklearn import metrics

       #Splitting the dataset into the Training set and Test set
       X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=123, stra
       print('train features shape:',X_train.shape)
       print('test features shape:',X_test.shape)
       print('train label shape:',y_train.shape)
       print('test label shape:',y_test.shape)

train features shape: (59883, 31)
test features shape: (19961, 31)
train label shape: (59883,)
test label shape: (19961,)
```

- Applying SMOTE sampling to handle class imbalance.
- Fit it to our dependent and independent features.
- Then applying train test split and split the 75% of data to train and 25% of data to test.

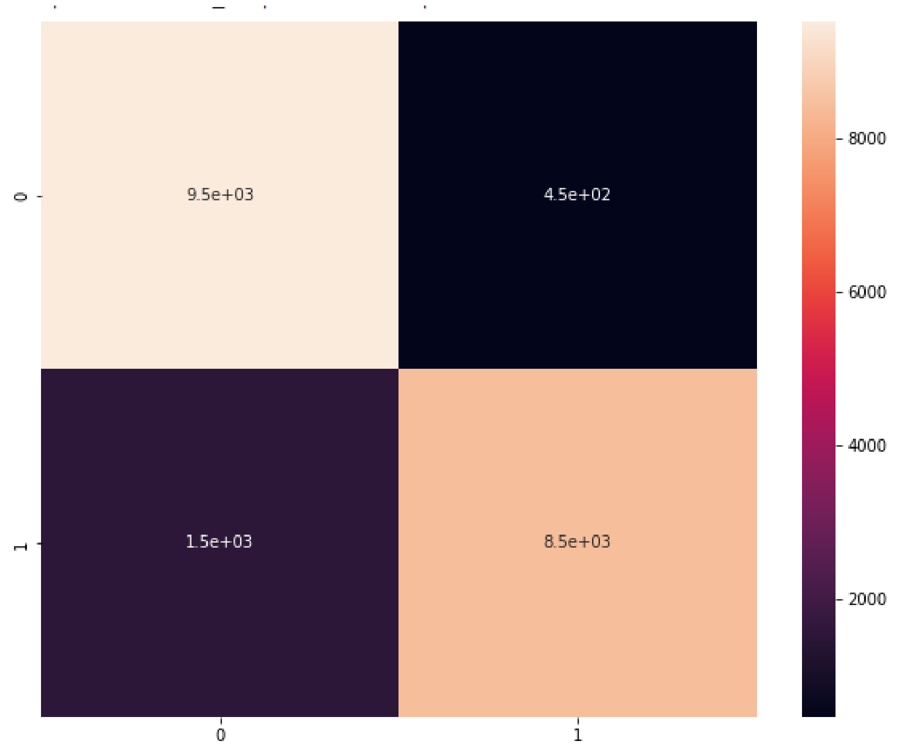
K – NEAREST NEIGHBOUR CLASSIFIER

```
Cross_validation score [0.76329632 0.77523587 0.77423395 0.77722111 0.7759686 ]
KNN Test accuracy Score 0.7844797354841941
              precision    recall  f1-score   support

           0       0.84      0.70      0.77      9981
           1       0.74      0.87      0.80      9980

    accuracy                           0.78     19961
   macro avg       0.79      0.78      0.78     19961
weighted avg       0.79      0.78      0.78     19961

array([[7025, 2956],
       [1346, 8634]])
```



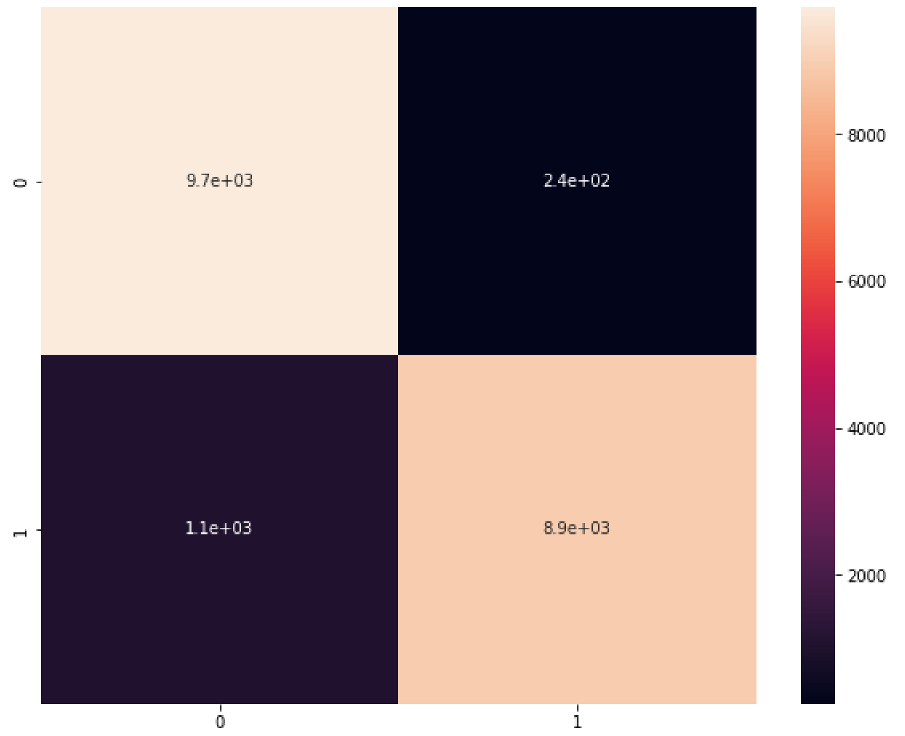
RANDOM FOREST CLASSIFIER

```
Cross_validation score [0.90047591 0.89863906 0.89696919 0.89512358 0.89863059]
RandomForest Test accuracy Score 0.9009067681979861
              precision    recall  f1-score   support

           0       0.86      0.95      0.91      9981
           1       0.95      0.85      0.90      9980

    accuracy                           0.90     19961
   macro avg       0.91      0.90      0.90     19961
weighted avg       0.91      0.90      0.90     19961

array([[9529,  452],
       [1526, 8454]])
```



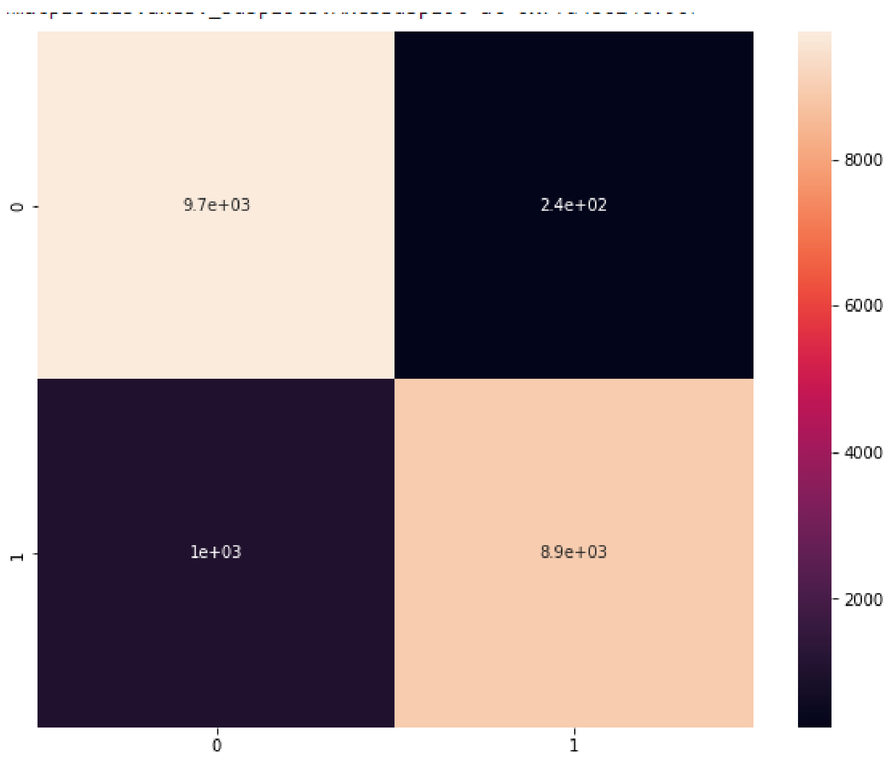
XG BOOST CLASSIFIER

```
Cross_validation score [0.93228688 0.930784   0.9308675  0.92718771 0.93495324]
xgb Test accuracy Score 0.9348730023545915
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      9981
           1       0.97      0.89      0.93      9980

    accuracy                           0.93     19961
   macro avg       0.94      0.93      0.93     19961
weighted avg       0.94      0.93      0.93     19961

array([[9736,  245],
       [1055, 8925]])
```

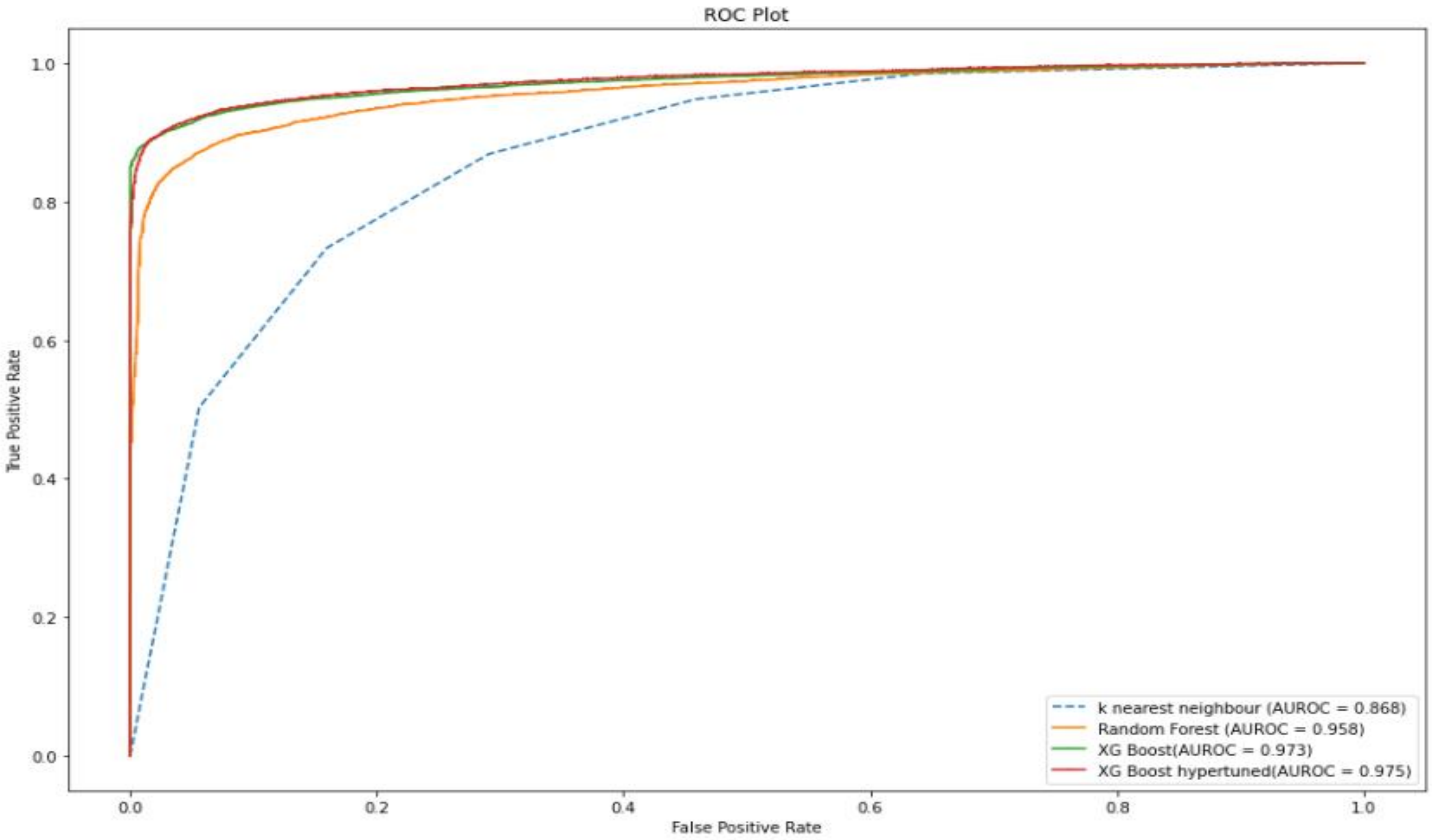


HYPERPARAMETER TUNING OF XG BOOST CLASSIFIER

```
Cross_validation score [0.9308675  0.93061702 0.93153544 0.92844021 0.93461924]
xgb Test accuracy Score 0.9356244677120384
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      9981
           1       0.97      0.90      0.93      9980

    accuracy                           0.94     19961
   macro avg       0.94      0.94      0.94     19961
weighted avg       0.94      0.94      0.94     19961

array([[9740,  241],
       [1044, 8936]])
```



AUC-ROC SCORE OF ALL MODELS



MODEL AND ITS ACCURACY

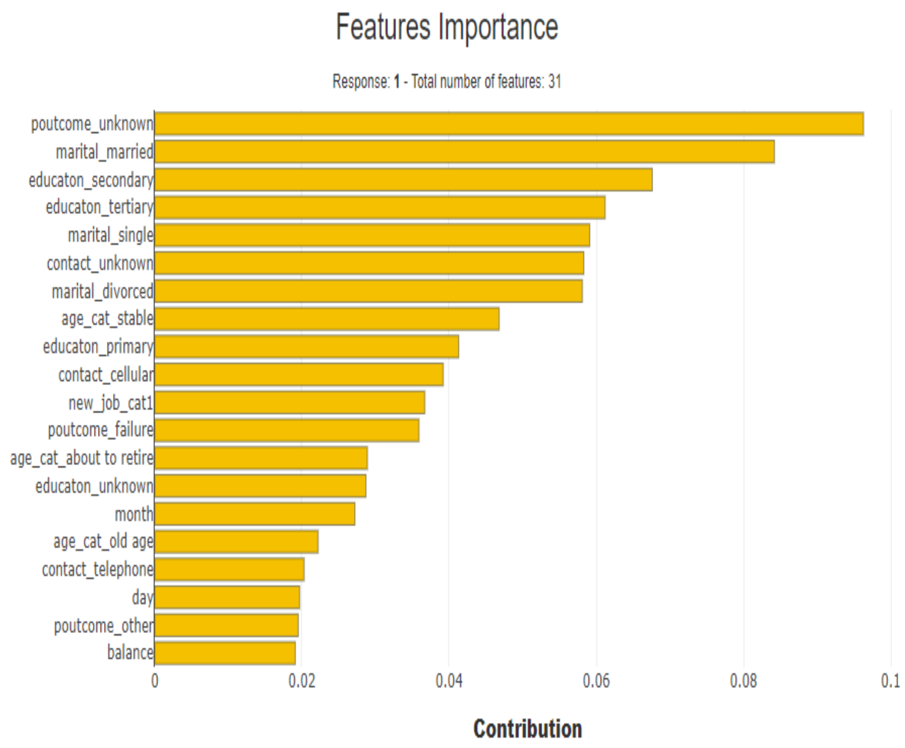
SERIAL NO.	MODEL	ACCURACY	AUC_ROC CURVE
1.	K-NEAREST NEIGHBOUR	0.79	0.868
2.	RANDOM FOREST CLASSIFIER	0.90	0.958
3.	XG BOOST CLASSIFIER	0.93	0.973
4.	HYPERPARAMETER TUNING OF XG BOOST CLASSIFIER	0.94	0.975

- XG Boost classifier perform best for predicting target variable also we we hypertuned XG boost model it will increases
- The accuracy of prediction by 1 % .
- XG Boost hypertunes classifier gives the accuracy of 94% and auc-roc score 97.5 %.

SHAPASH MODEL EXPLANATORY

	y	proba	feature_1	value_1	contribution_1	feature_2	value_2	contribution_2	feature_3	value_3	contribution_3	feature_4	value_4	contribution_4
46659	1	0.999093	educaton_secondary	0.0	1.03521	educaton_tertiary	0.0	0.987838	poutcome_success	1.0	-0.982098	educaton_primary	0.0	0.938685
9640	0	0.985386	contact_unknown	1.0	2.890928	age_cat_old age	1.0	1.263951	poutcome_unknown	1.0	1.206027	marital_married	1.0	1.07472
5490	0	0.983699	contact_unknown	1.0	2.468846	marital_married	1.0	1.372475	poutcome_unknown	1.0	1.11631	educaton_tertiary	1.0	0.948965
4130	0	0.961427	contact_unknown	1.0	2.233424	marital_married	1.0	1.347285	educaton_primary	1.0	1.173288	poutcome_unknown	1.0	1.100887
70998	1	0.193081	marital_married	1.0	-1.283402	educaton_secondary	1.0	-1.158298	poutcome_unknown	1.0	-1.073155	age_cat_stable	1.0	-0.592026
13347	0	0.966695	age_cat_old age	1.0	1.610339	marital_married	1.0	1.319019	poutcome_unknown	1.0	1.134885	loan	1.0	0.779314
65079	1	0.999010	poutcome_unknown	0.0	1.658317	educaton_secondary	0.0	1.097342	educaton_tertiary	0.0	1.000353	educaton_primary	0.0	0.954748
52952	1	0.976650	age_cat_stable	0.0	1.616673	marital_married	1.0	-0.960048	educaton_secondary	1.0	-0.947516	age_cat_about to retire	0.0	0.923824

- From the above model explanatory tool we have seen that poutcome Unknown is the most important feature while predicting our target variable also from the table we can see that when the poutcome is 0 then it contribute in the negative way and increases the probability of predicting 0.
- Marital married is the second most important feature for predicting target variables from the table we can see that when the marital married then it will affect positively and increases the probability of predicting 1.
- Also age cat stable variable affect positively on the target variable when the age of clients is stable then it will increases the probability of predicting 1 that means it higher the probability that client will subscribe for term deposit.
- Also education secondary affects positively on the target variable when the client education is secondary then it increases the probability that client will agree to subscribe for term deposit.



CONCLUSION

- From the above project we can conclude that XG boost classifier is the best fit classification model for predicting weather the client agree to subscribe for personal loan or not.

- When we Hypertuned these XG Boost classifier the accuracy of the model increases by 1 % So it predicts 94% prediction correctly.

- There are some important feature for predicting our target variable we use Shapash model explanatory to explore that features.

- We visualize 20 feature which are most important while predicting target variable.

- From that feature we conclude that clients age , education ,job and and marital status and outcome of previous campaign are the most important feature for predicting that weather client agree to subscribe for term deposit or not that’s why bank prefer these information to start for new campaign and to target customer.