Project no :- 1

Bank customer churn

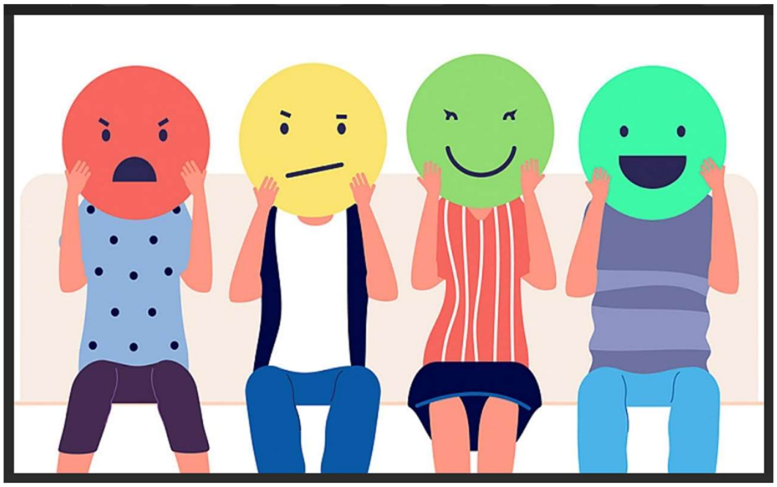
(Binary class classification problem)

By

Ankit Gurudas Dhanore

Content

- Problem statement
- Objective
- Introduction
- Data summary
- Importing libraries
- Preprocessing
- EDA
- Feature engineering
- Modelling
- Model comparison
- Challenges faced



Problem Statement

- To develop a bank customer churn prediction model that can accurately identify customers who are likely to end their relationship with a bank in the near future. The challenge is to identify the key factors that drive customer churn and develop a scalable and interpretable model that can provide actionable insights to help banks retain their customers.

Objective

- The objective of creating a bank customer churn prediction model is to accurately identify customers who are likely to churn in the near future and provide actionable insights to help banks develop effective retention strategies. As retaining old custommers is easy than to gain new customers By achieving this objective, banks can retain their customers, improve customer satisfaction, and boost their revenue and profitability.

Introduction

- Bank customer churn refers to the phenomenon of customers ending their relationship with a bank by closing their accounts or reducing the number of services they use. Customer churn can have a significant impact on a bank's revenue, profitability, and customer retention rates.
- There are several reasons why bank customers may choose to churn, such as dissatisfaction with the bank's services, better offers from competing banks, or changes in the customer's financial situation. Understanding the reasons behind customer churn is essential for banks to develop effective strategies to retain their customers.
- Some of the common strategies that banks use to reduce customer churn include improving customer service, offering personalized and relevant products and services, and providing incentives for customers to stay with the bank. By implementing effective retention strategies, banks can minimize the impact of customer churn on their business and improve their overall customer satisfaction and loyalty.

Importing Libraries

- Here we import all the libraries which are required for the project like sklearn for modelling and preprocessing , matplotlib and seaborn for data visualization etc.
- Then import the data set and see some basic information regarding the dataset like no of rows and columns and their data type.

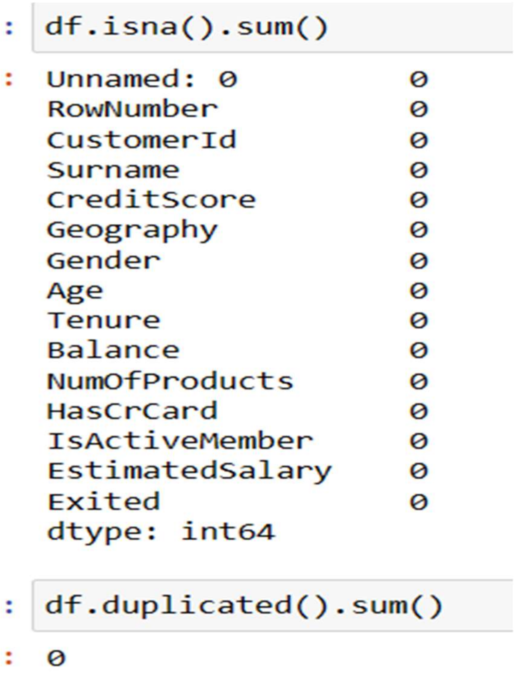
Data summary

- The data set is from online website.
- This data contains 10000 rows and 15 columns
- **RowNumber** :- Corresponds to the record (row) number and has no effect on the output.
- **CustomerId** :- Contains random values and has no effect on customer leaving the bank.
- **Surname** :- The surname of a customer has no impact on their decision to leave the bank.
- **CreditScore** :- Can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
- **Geography** :- A customer’s location can affect their decision to leave the bank.
- **Gender** :- It’s interesting to explore whether gender plays a role in a customer leaving the bank.
- **Age** :- This is certainly relevant, since older customers are less likely to leave their bank than younger ones.
- **Tenure** :- Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
- **Balance** :- Also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
- **NumOfProducts** :- Refers to the number of products that a customer has purchased through the bank.

- **HasCrCard** :- Denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
- **IsActiveMember** :- Active customers are less likely to leave the bank.
- **EstimatedSalary** :- As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- **Exited** :- Whether or not the customer left the bank.

Preprocessing

- First we checked for missing values but there are 0 null values present in the dataset.
- Then we checked for duplicate values but there also 0 duplicate values are present.
- we droped the features which are not required for the model that includes rownumber, customerid, surname, and unnamed:0 .
- Since there are no null or duplicate values are present in the dataset we jump to next step which is EDA.

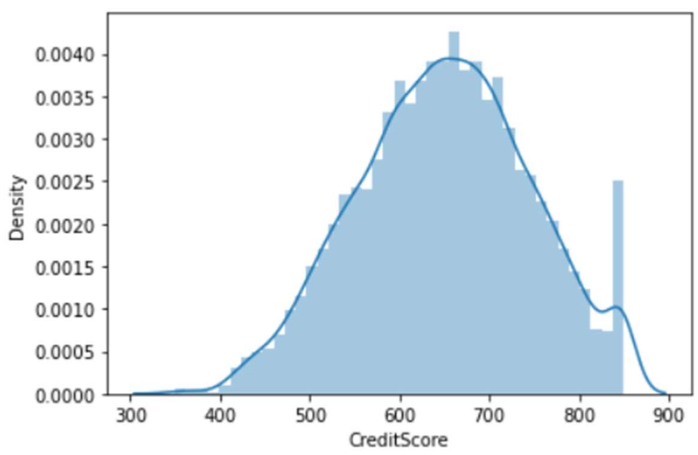


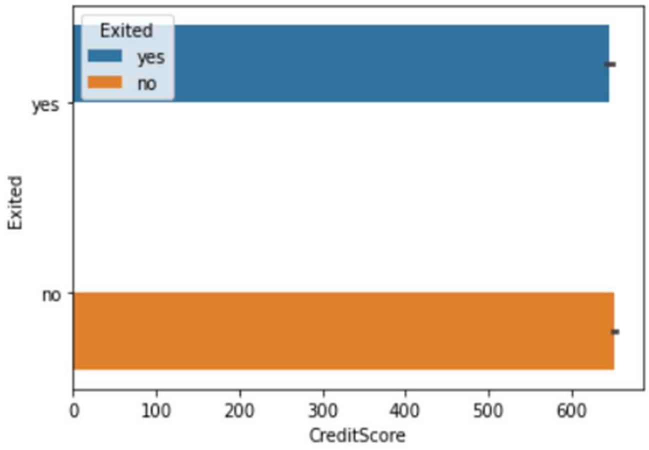
```
: df.isna().sum()

: Unnamed: 0          0
  RowNumber           0
  CustomerId          0
  Surname             0
  CreditScore         0
  Geography           0
  Gender              0
  Age                 0
  Tenure              0
  Balance             0
  NumOfProducts       0
  HasCrCard           0
  IsActiveMember      0
  EstimatedSalary     0
  Exited              0
  dtype: int64

: df.duplicated().sum()

: 0
```

EDA (exploratory data analysis)

❖ For Credit score

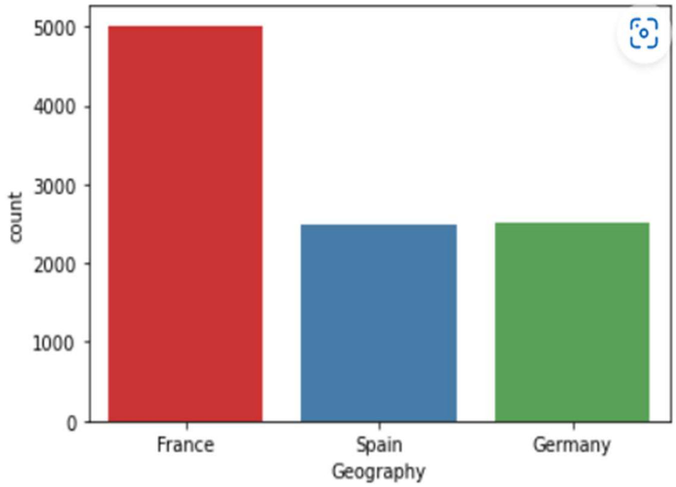
- Here we can see that the distribution curve of the credit score is nearly normal distribution.
- By comparing to exited feature it can be concluded that the credit score of customers is nearly equal which is around 650.

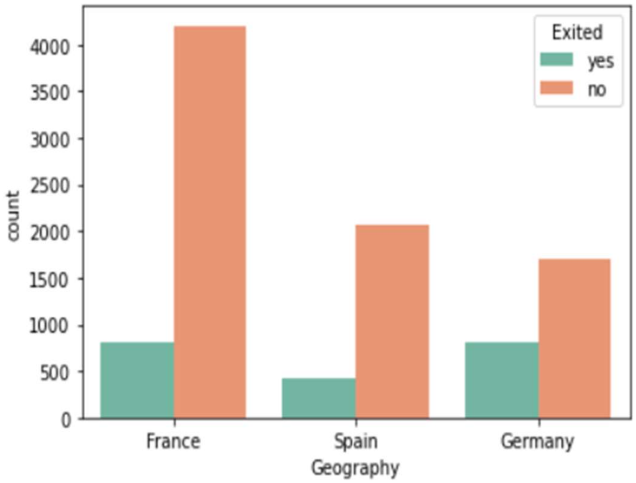




❖ For Geography

- From the below tables we can see that the maximum no of customers are from france.
- We can see that Germany has least no of customers but exited rate is so high there as compare to others.

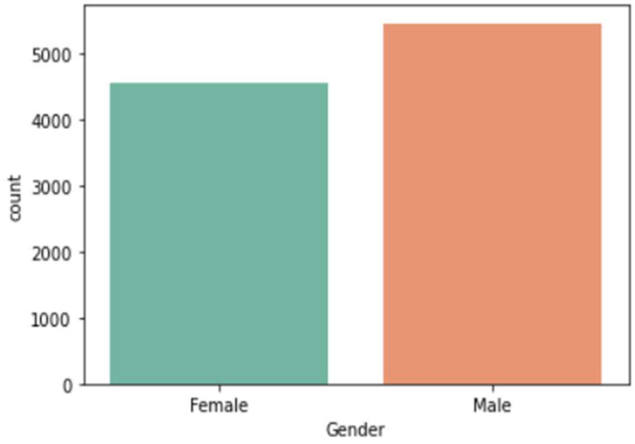


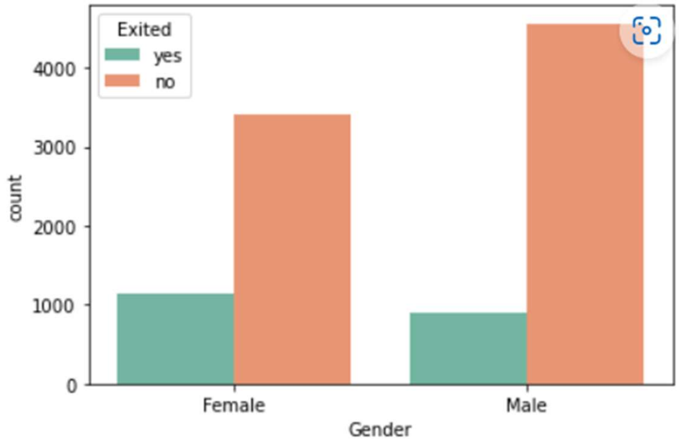


❖ For Gender

- Here we can clearly see that the no of male customers are more than female customers .

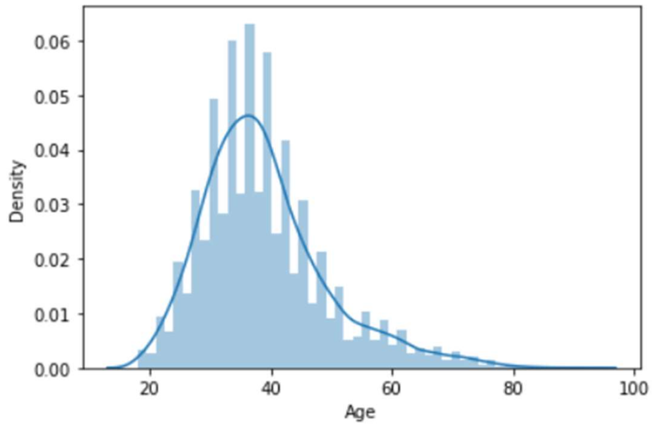
- But it is seen that though female customer are less but their exit rate are higher than male customers.

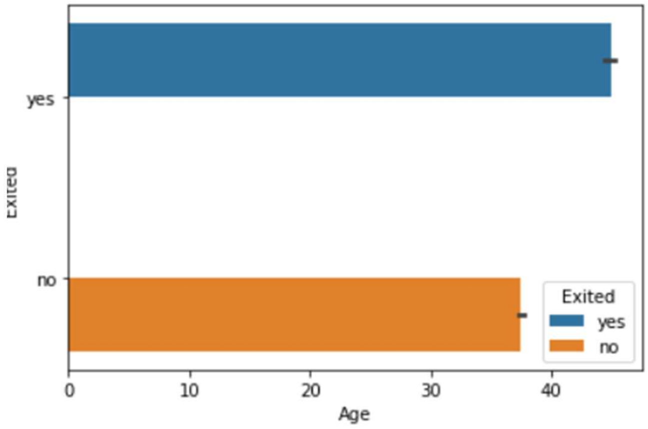




❖ For Age

- Here in age its distribution is little skewed but it is expected to be skewed as bank account is not opened till certain age and remain open till last.
- It is seen that the customers with average age more than 45 are exited more as compare to customer with average age less than 40.

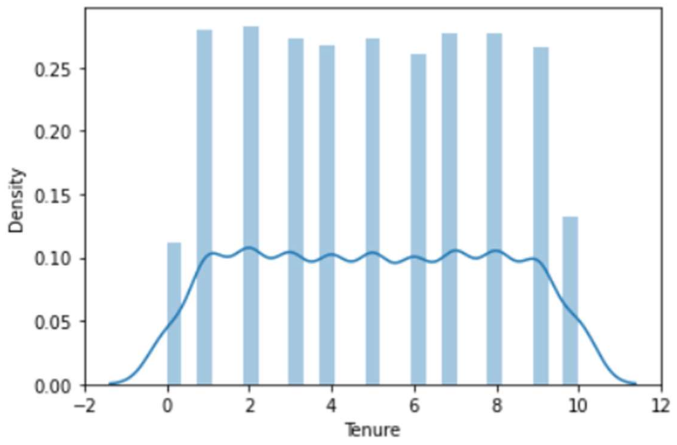


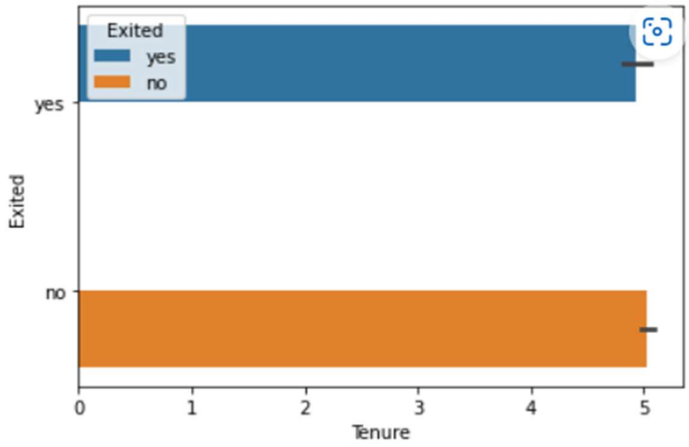


❖ For Tenure

- Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.

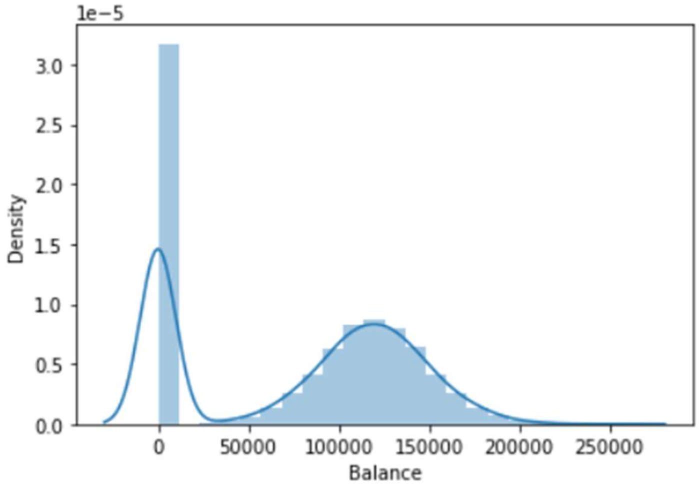
- Tenure of the exited customers and non exited customers are nearly same.

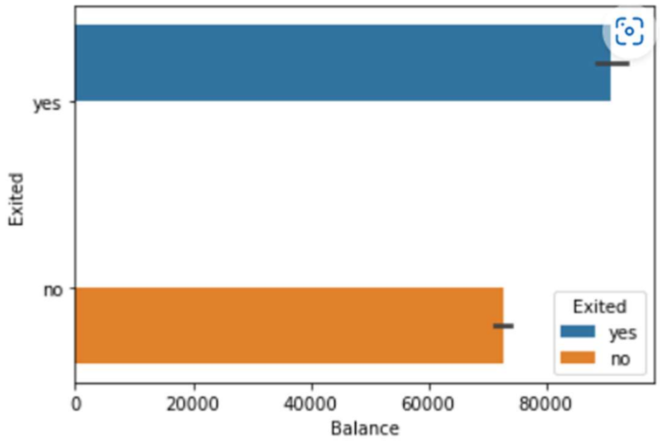




❖ For Balance

- It is clear that the max no of accounts are zero balance account and they don’t have any money in that account.
- From the below table it is seen that the customers with more avg balance are more likely to leave the bank.

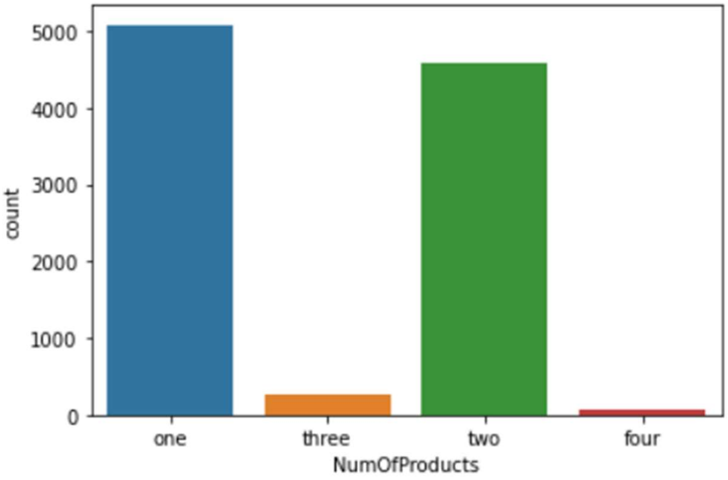


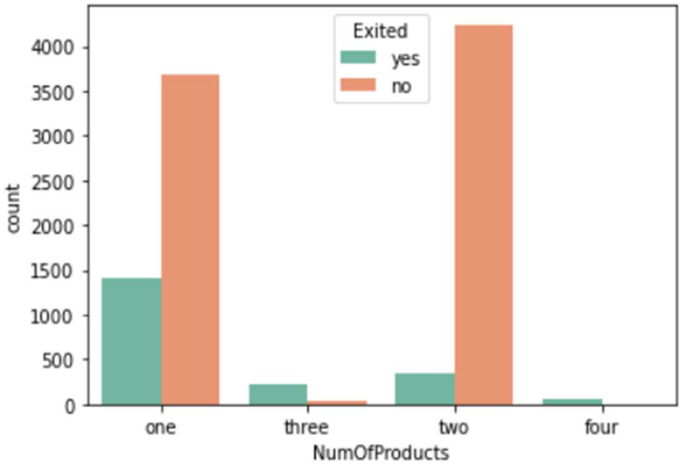


❖ For NumOfProducts

- There are total 4 products of the bank out of which only 1st and 2nd are selling.
- 3rd and 4th products sales are very less as compare to other two.

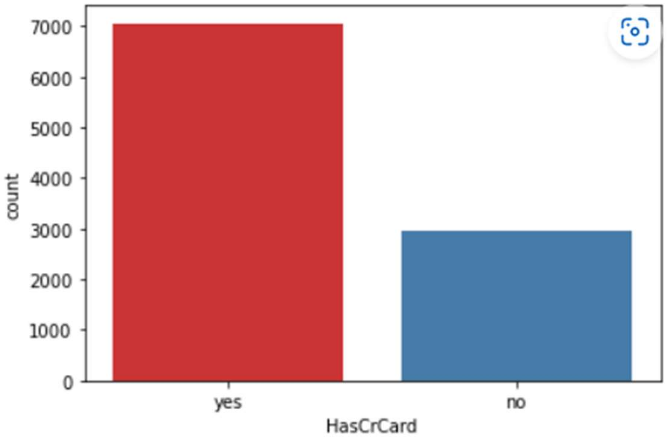
- Only product 2 is performing well in comparison to other 3.

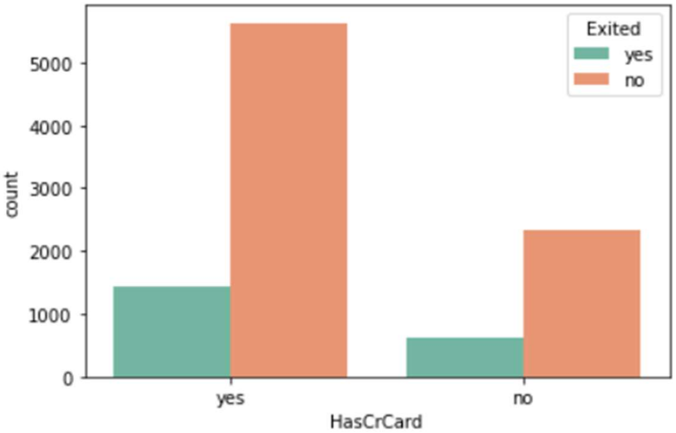




❖ For HasCrCard

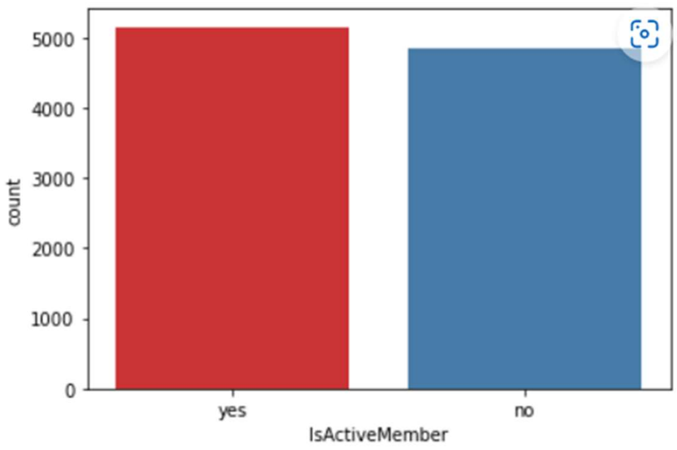
- From the below graph it is clear that the no of customers which has credit crad is greater which don't has credit cards.
- customer which has credit card has greater chance of leaving the bank.

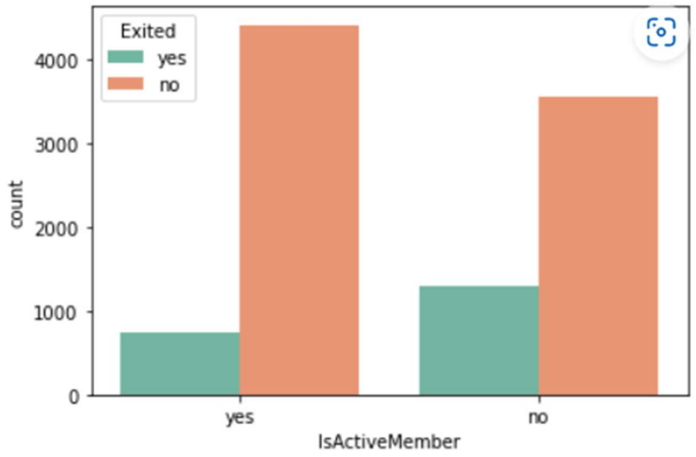




❖ For IsActiveMember

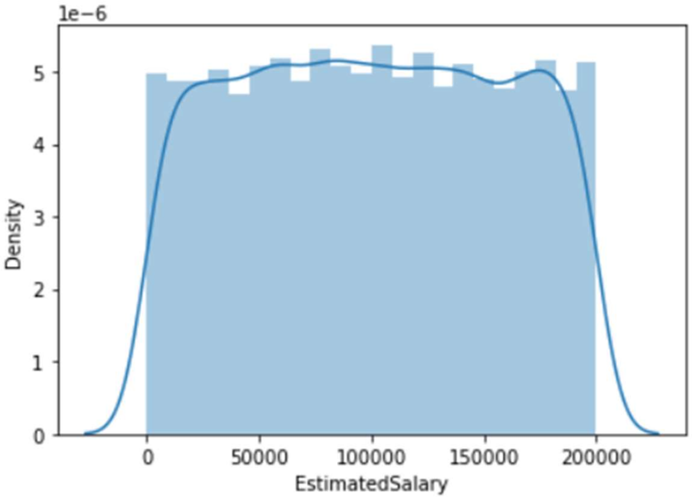
- There are many inactive members in the bank.
- The inactive members are more likel to leave the bank.

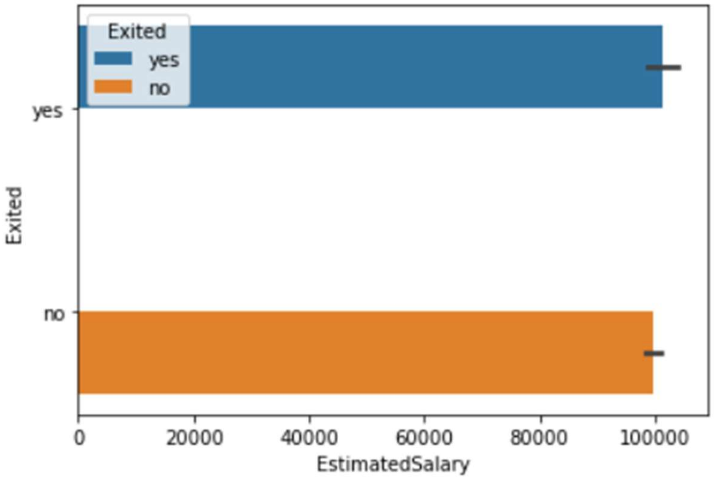




❖ For EstimatedSalary

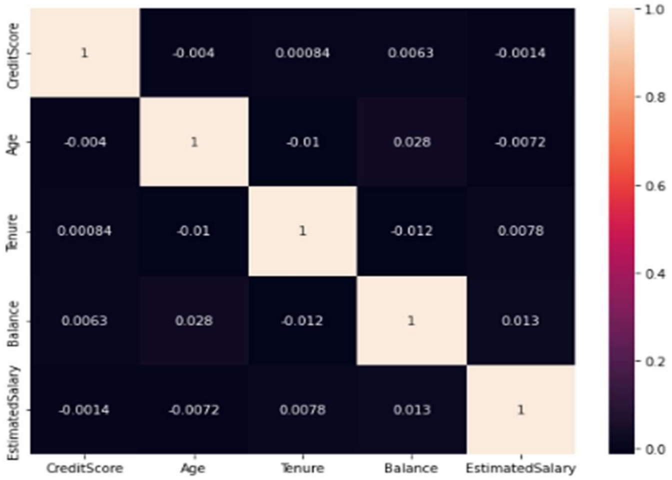
- people with lower salaries are more likely to leave the bank compared to those with higher salaries.

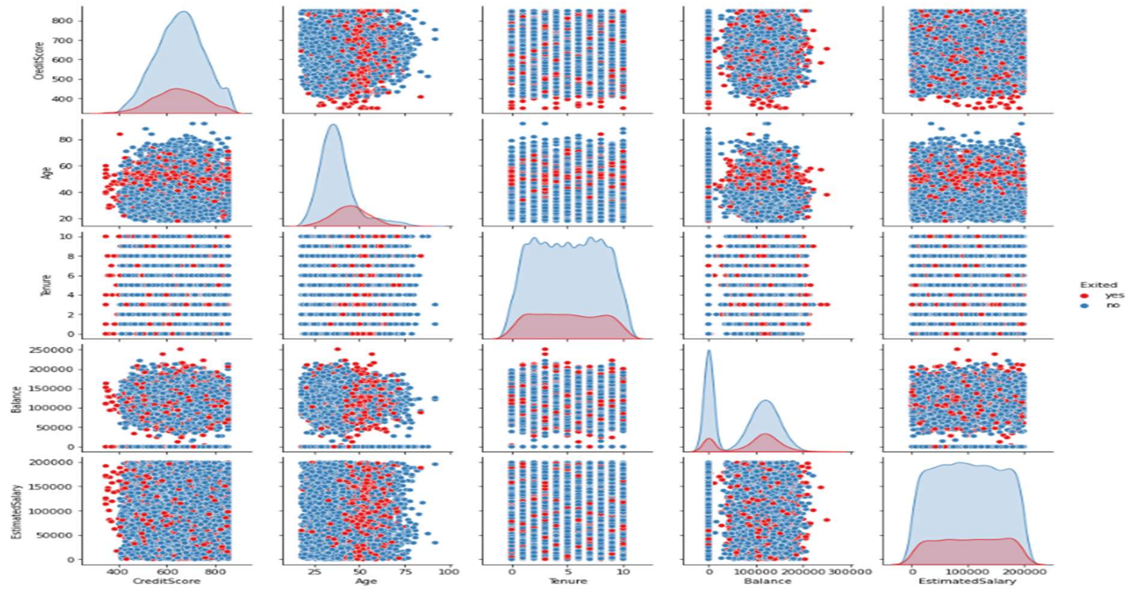




❖ Multivariate analysis

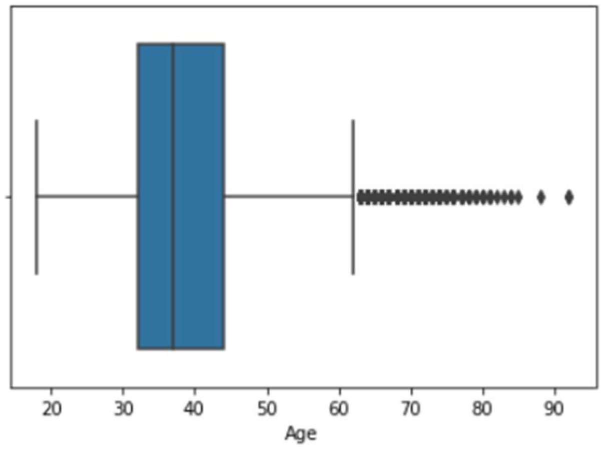
- It shows no high multi colinearity between the features.
- All features are independent to each others.

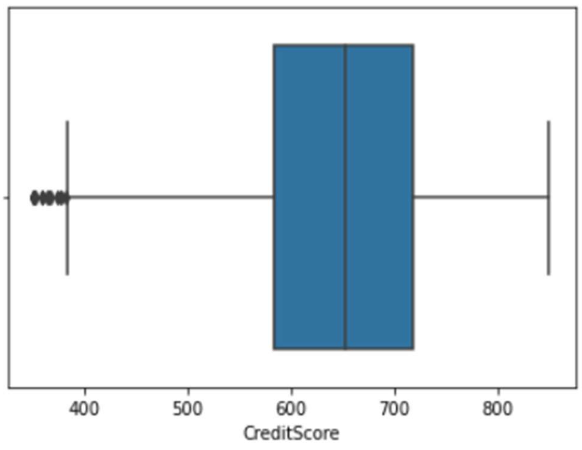




❖ For outlier detection

- We detect outliers using boxplot.
- Only age and credit score has some outlier but it is expected to have outliers. It is neccesery to keep them.





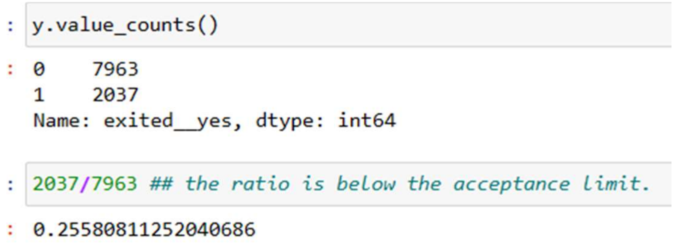
Feature engineering

- First we apply one hot encoding to geography, gender, numofproducts, hascrcards, isactivemember and exited feature as it has nominal type of data.
- After applying one hot encoding we concat all individual dataset to create one final dataset.

- From that new data set we drop old geography, gender, numofproducts, hascrcards, isactivemember and exited features.

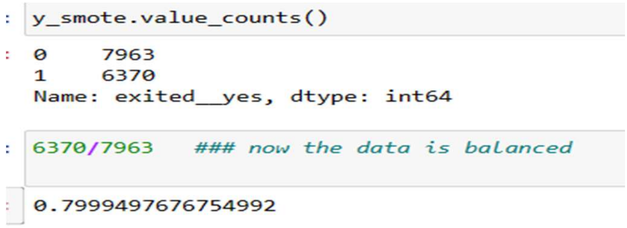
Handling imbalanced data

- We see that our data is imbalanced with th ratio of 0.255 so we apply **smote** (synthetic minority oversampling technique.



```
y.value_counts()

0    7963
1    2037
Name: exited__yes, dtype: int64

2037/7963 ## the ratio is below the acceptance limit.

0.25580811252040686
```

- After applying smote we get ratio of 0.80 which is good for our model .



```
y_smote.value_counts()

0    7963
1    6370
Name: exited__yes, dtype: int64

6370/7963    ### now the data is balanced

0.7999497676754992
```

Feature selection

- There are 3 types of feature selection methods .
  1. Filter method.
  2. Wrapper method.
  3. Embedded method.
- We check for different methods like correlation method in filter method, forward propagation in wrapper method, and extra tree classifier in embedded method.

Modelling

- Here we checked for cross validation score for evaluation of models.

❖ Random forest classifier

- Cross validation score :- 0.87

❖ Adaboost classifier

- Cross validation score :- 0.85

❖ Decision tree classifier

- Cross validation score :- 0.82

❖ Kneighbors classifier

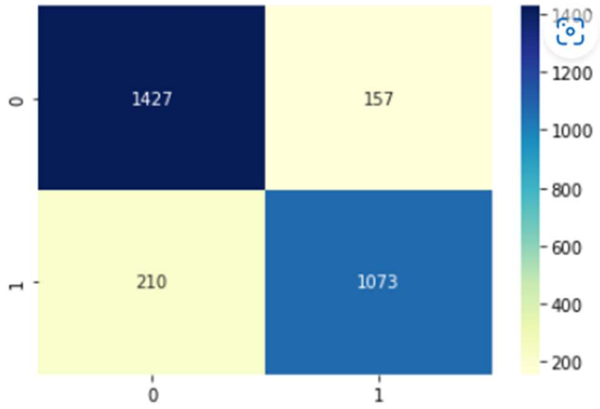
- Cross validation score :- 0.68

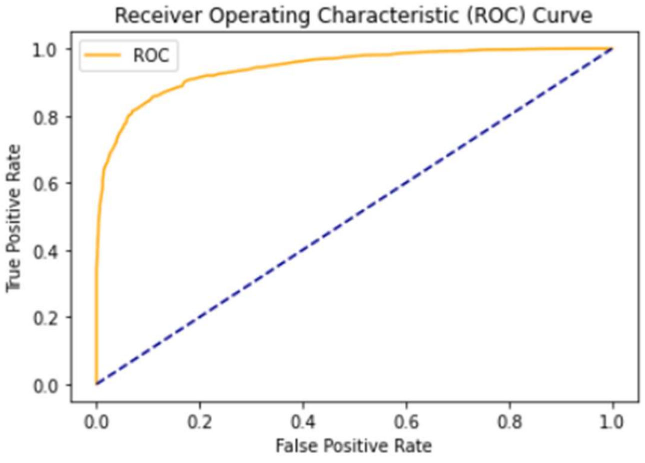
❖ Logistic regression

- Cross validation score :- 0.66

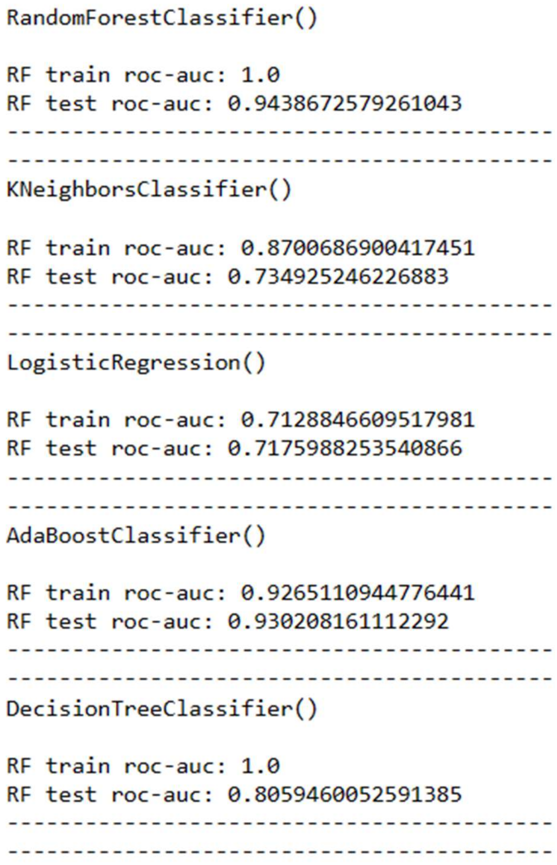
Model comparison

- It is clear that the Random forest is working best so we will go with Random forest algorithm.
  - Accuracy score :- 087
  - Precision :- 087
  - Recall :-0.87
  - F1 score :- 0.87
  - Confusion matrix :-
- We also checked for auc curve.





- We also calculate the roc auc score for model evaluation here also random forest works best.



```
RandomForestClassifier()

RF train roc-auc: 1.0
RF test roc-auc: 0.9438672579261043
------------------------------------------
------------------------------------------
KNeighborsClassifier()

RF train roc-auc: 0.8700686900417451
RF test roc-auc: 0.734925246226883
------------------------------------------
------------------------------------------
LogisticRegression()

RF train roc-auc: 0.7128846609517981
RF test roc-auc: 0.7175988253540866
------------------------------------------
------------------------------------------
AdaBoostClassifier()

RF train roc-auc: 0.9265110944776441
RF test roc-auc: 0.930208161112292
------------------------------------------
------------------------------------------
DecisionTreeClassifier()

RF train roc-auc: 1.0
RF test roc-auc: 0.8059460052591385
------------------------------------------
------------------------------------------
```

Challenges faced

- Most difficult task was to do preprocessing as the data was very noisy and a bit tricky to simplify.
- Doing feature engineering was difficult.

Thank you