Predicting Churn for Bank Customers

Bhargava Gaggainpali

DSC680 - Applied Data Science

bgaggainpali/bgaggainpali_DSC680 (github.com)

Business problem & Hypothesis

In financial industry, banks are playing important role in challenging times like now, with COVID pandemic across the globe. People are losing jobs and financial institutions are facing more Customer churn in Bank Accounts. The increase in Customer Churn rate will result in significant financial loss to commercial banks. It is very critical for lending institutions like banks to have a prediction model to be able to predict customers churn to better serve the customers and reduce the churn.

I have selected the topic, as I was interested in knowing the variables which influence the Bank Account Customer Churn key factors. As I explore more about the domain, I understand that it's not same set of rules which is being used across domain and each different banks and credit unions are based on different features and calculations when predicting the churn and the factors which influence them.

Solution Method

I see this problem as a classification issue, where we should try to understand and able to predict the customers, who have high Bank Customer Churn chances. Planning to use supervised machine learning algorithm to work on the classification problem to be trained with algorithms like:

1. Logistic Regression
2. Decision Tree
3. Random Forest

Start with loading data into a data frame and then understand the data, then perform Exploratory Data Analysis (EDA) on the data set. EDA involves doing Univariate and Bivariate Analysis, identify missing values and outliers and fill the gaps with appropriate values. In the next step, building model with starting from logistic regression and observe the accuracy of the model. When the accuracy of the of the model is not high, then planning to use Decision Tree and Random Forest to achieve higher accuracy.

Technical approach involves understanding the data by drawing multiple charts to observe the target variable with respect to each of the variable. Build a heat map to understand the relationship between variables. Build model using the different algorithms and observe the accuracy of the model, evaluate the accuracy of the model by building confusion matrix.

Data

I have identified Churn_Modelling.csv as source for my work, below is the Kaggle link. There are 10,000 observations in the dataset, each row in the dataset represents a Bank Customer Account. Given is the list of variables in the dataset.

Source File:

https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers?select=Churn_Modelling.csv

Variable	**Description**
RowNumber	Sequence Number
CustomerId	Customer Account Number
Surname	Customer Name
CreditScore	Credit Score

Geography	Location Country
Gender	Male / Female
Age	Customer Age
Tenure	Period of time in Years as Customer to the Bank
Balance	Balance amount in the Bank
NumOfProducts	Number of Products availed by the Customer
HasCrCard	Customer has Credit card
IsActiveMember	Customer Active Member
EstimatedSalary	Customer Salary
Exited	Customer Churn value

Initial Observations

```
# Check columns list and missing values
df.isnull().sum()
```

```
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
```

```
# Get unique count for each variable
df.nunique()
```

```
RowNumber          10000
CustomerId         10000
Surname             2932
CreditScore          460
Geography              3
Gender                 2
Age                   70
Tenure                11
Balance             6382
NumOfProducts          4
HasCrCard              2
IsActiveMember         2
EstimatedSalary     9999
Exited                 2
dtype: int64
```

Categorical Features: Based on the data, below are categorical variables.

Geography

Gender

Exited 1 = Exited, 0 = Not Exited

Ordinal Features: Based on the data with inherent hierarchy, below are ordinal variables.

HasCrCard

IsActiveMember

Numerical Features: Based on the numerical data, below are numerical variables.

CreditScore

Age

Tenure

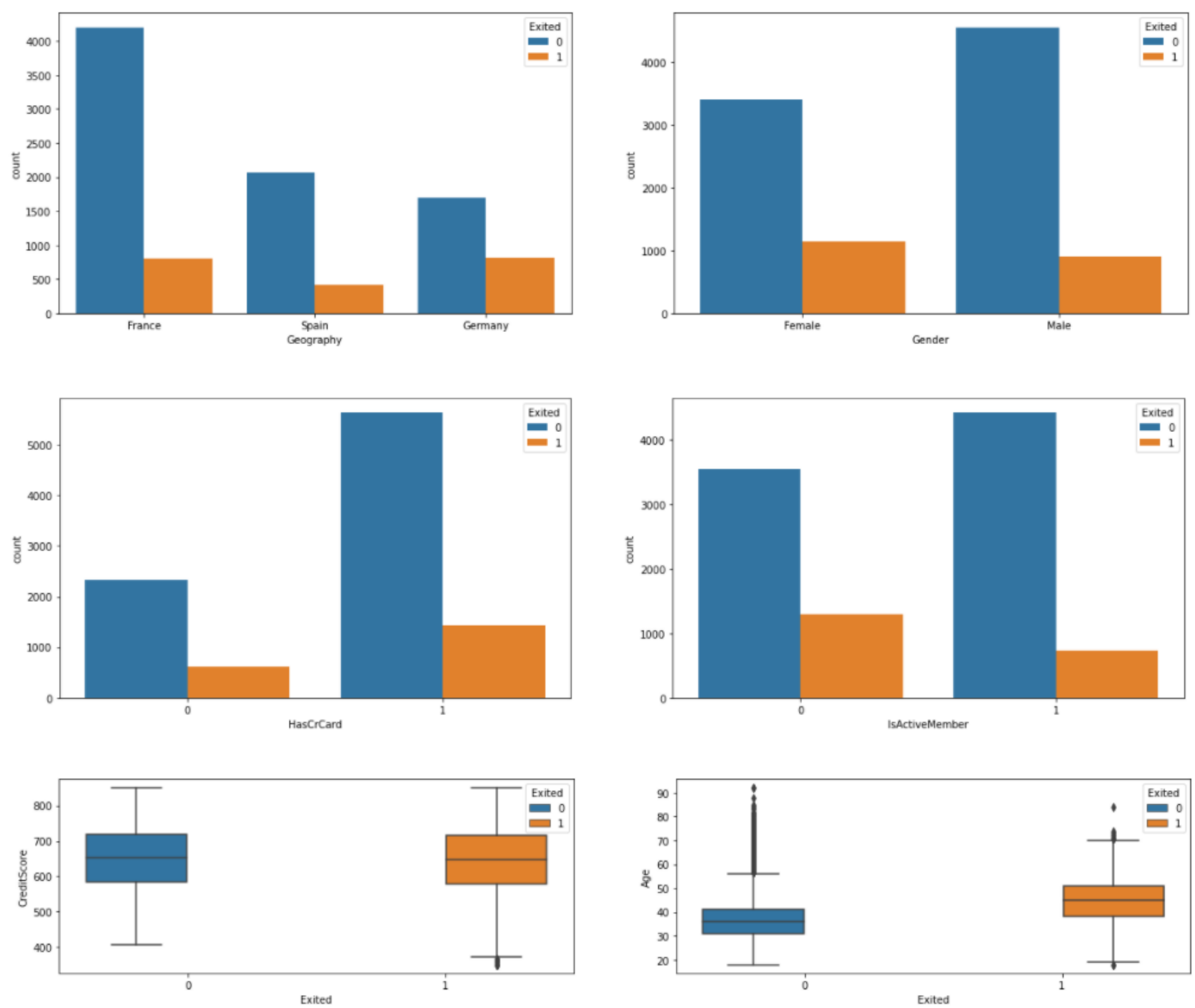
Balance

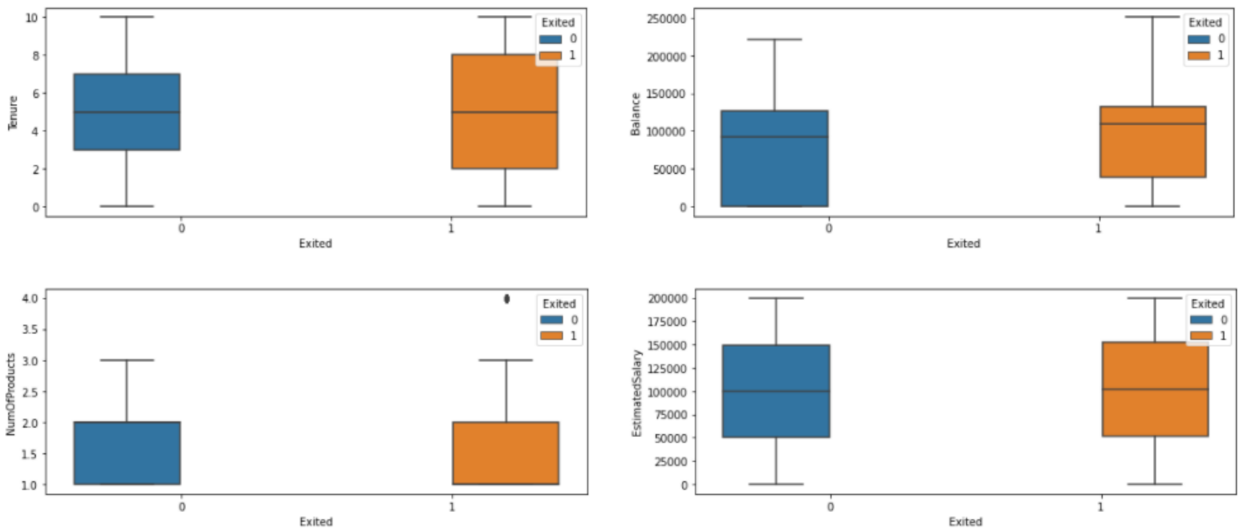
NumOfProducts

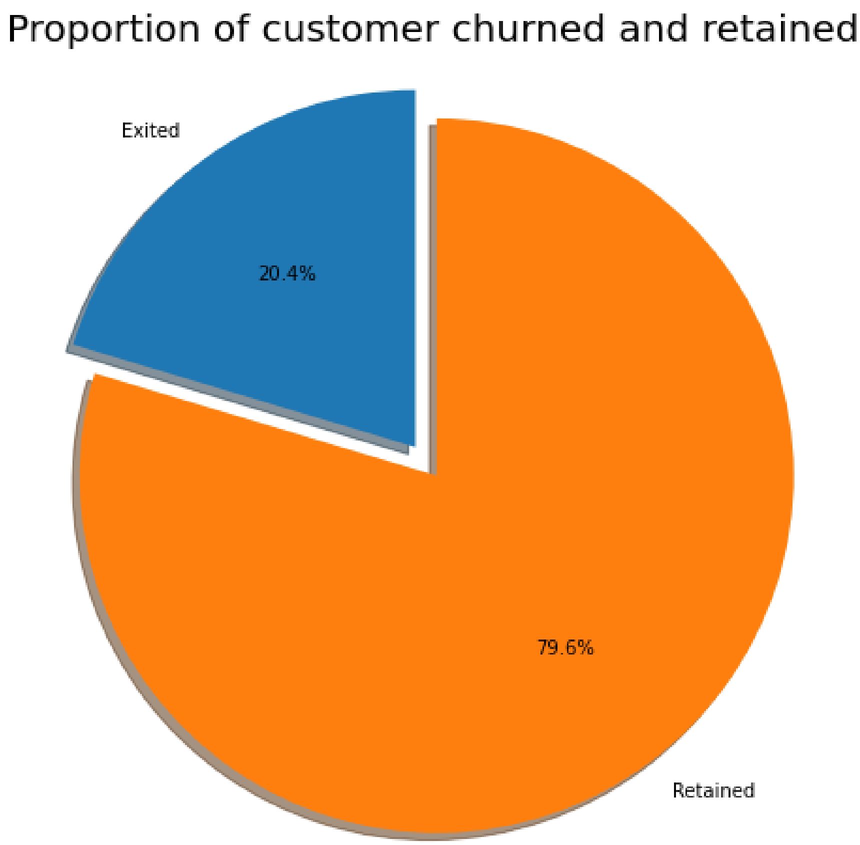
EstimatedSalary

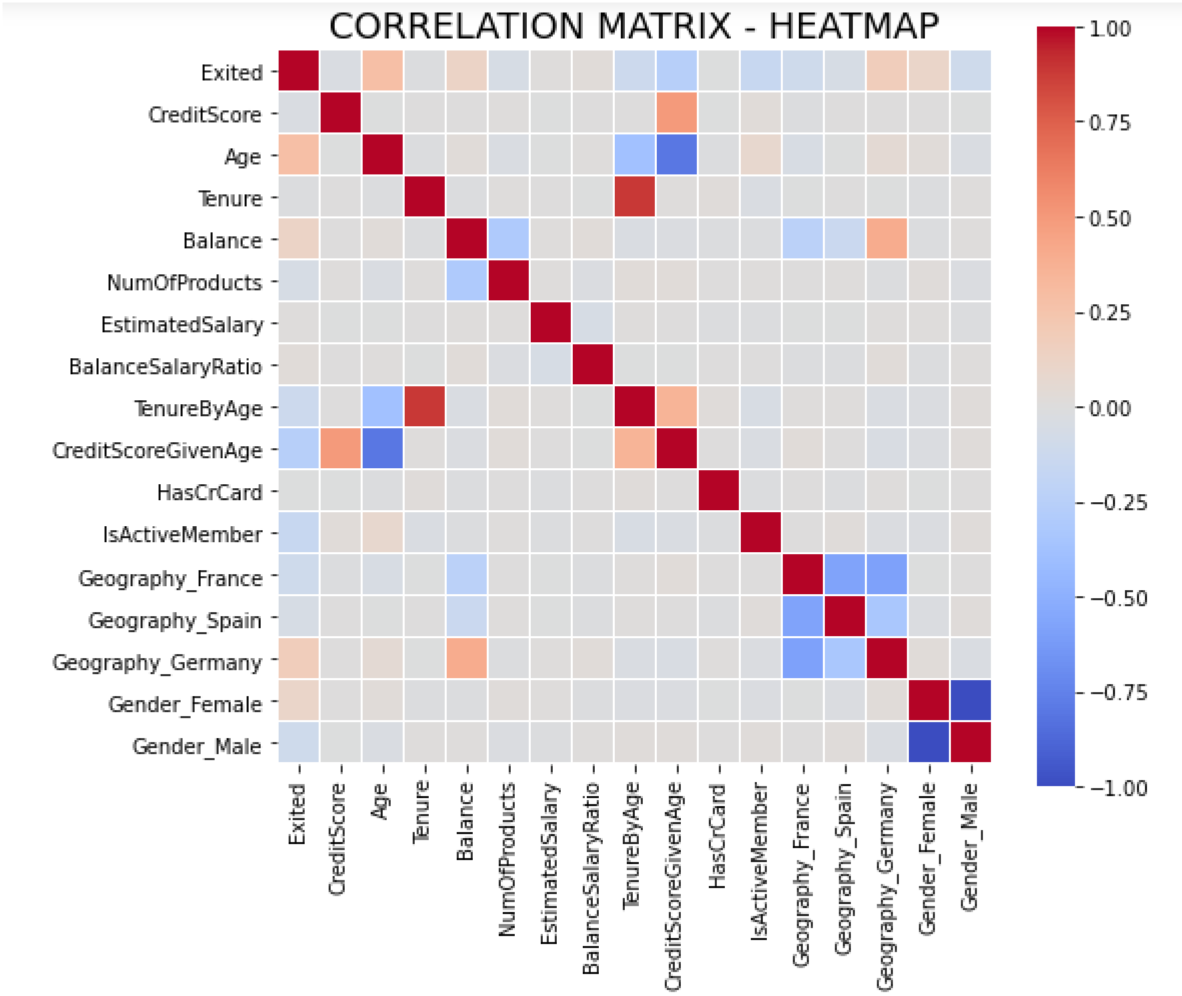
Exploratory Data Analysis

After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression to consider variables like Balance, Age, Gender, Tenure, CreditScore, HasCrCard and EstimatedSalary. I saw surprising stats when I used visualizations to give clear idea on how each factor has its effect on the Customer Churn. I had to increase my research questions to explore and include more variables, than initially prepared. Its based on the initial analysis using visualization.




Tenure
Exited
Balance
Exited
NumOfProducts
Exited
EstimatedSalary
Exited
Exited
0
1


Proportion of customer churned and retained
Exited
20.4%
79.6%
Retained



Few observations based on the above plots.

1. Majority of the data is from persons from France. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank.
2. The proportion of female customers churning is also greater than that of male customers
3. Majority of the customers that churned are those with credit cards.
4. Inactive members have a greater churn.

5. There is no significant difference in the credit score distribution between retained and churned customers.
6. The older customers are churning at more than the younger ones alluding to a difference in service preference in the age categories.
7. With regard to the tenure, the clients on either extreme end (spent little time with the bank or a lot of time with the bank) are more likely to churn compared to those that are of average tenure.
8. Bank is losing customers with significant bank balances which is likely to hit their available capital for lending.
9. Neither the product nor the salary has a significant effect on the likelihood to churn.

Data Preparation

Applied MinMax Scaler to scale the numeric data variables.

CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	BalanceSalaryRatio	TenureByAge	CreditScoreGivenAge
0.538	0.324324	0.2	0.000000	0.000000	0.506735	0.000000	0.085714	0.235083
0.516	0.310811	0.1	0.334031	0.000000	0.562709	0.000070	0.043902	0.237252
0.304	0.324324	0.8	0.636357	0.666667	0.569654	0.000132	0.342857	0.168807
0.698	0.283784	0.1	0.000000	0.333333	0.469120	0.000000	0.046154	0.310859
1.000	0.337838	0.2	0.500246	0.000000	0.395400	0.000150	0.083721	0.354739

Applied One Hot Encoding to convert Categorical data to Numerical data variables.

Geography_France	Geography_Spain	Geography_Germany	Gender_Female	Gender_Male
1	-1	-1	1	-1
-1	1	-1	1	-1
1	-1	-1	1	-1
1	-1	-1	1	-1
-1	1	-1	1	-1

Applied PCA to reduce the number of Dimensions or Variables.

```
# redusing the number of companents to 4 using PCA
pca=PCA(n_components=4)
pca.fit(df_final_data)
# Transform the data after applying PCA
df_final_data_PCA = pca.transform(df_final_data)
print('Number of elements in the data frame after applying PCA ')
df_final_data_PCA.shape
```

```
Number of elements in the data frame after applying PCA

(10000, 4)
```

```
# Display the input data which is converted to 4 components using PCA
df_final_data_PCA = pd.DataFrame(df_final_data_PCA)
df_final_data_PCA.columns = ['PCA_Comp_1','PCA_Comp_2','PCA_Comp_3','PCA_Comp_4']
df_final_data_PCA.head()
```

	PCA_Comp_1	PCA_Comp_2	PCA_Comp_3	PCA_Comp_4
0	1.476246	-1.282872	0.664274	-0.718695
1	1.496844	1.161335	1.918017	0.235082
2	1.541878	-1.237159	-0.709382	0.739816
3	1.554689	-1.261729	-0.484517	0.775756
4	1.484656	1.162702	1.729596	0.216442

Model Development

As part of the current project, four models were developed after data preparation steps. Data is split in the ratio of 70:30 for train and test, i.e. 70% of the data is fed to the model to understand the patterns and remembering the outcome, later 30% of the data is used to validate the prediction results.

```
from sklearn.model_selection import train_test_split

# split the data
#X_train, X_val, y_train, y_val = train_test_split(df_final_data_PCA, df_tgt_Label, test_size =0.3, random_state=11)
X_train, X_val, y_train, y_val = train_test_split(df_final_data, df_tgt_Label, test_size =0.3, random_state=11)


# number of samples in each set
print("No. of samples in training set: ", X_train.shape[0])
print("No. of samples in validation set:", X_val.shape[0])
```

```
No. of samples in training set:  7000
No. of samples in validation set: 3000
```

Below are four models

- ***Logistic Regression***

Logistic Regression

```
#--------------
# Logistic Regression
#--------------
from sklearn.linear_model import LogisticRegression
classifier2 = LogisticRegression()
classifier2.fit( X_train, y_train )
y_pred = classifier2.predict( X_val )

cm = confusion_matrix( y_val, y_pred )
print("Accuracy on Test Set for LogReg = %.2f" % ((cm[0,0] + cm[1,1] )/len(X_val)))
scoresLR = cross_val_score( classifier2, X_train, y_train, cv=10)
print("Mean LogReg CrossVal Accuracy on Train Set %.2f, with std=%.2f" % (scoresLR.mean(), scoresLR.std() ))
```

```
Accuracy on Test Set for LogReg = 0.81
Mean LogReg CrossVal Accuracy on Train Set 0.81, with std=0.01
```

- ***Kernel SVM Model***

kernel SVM Model

```
# kernel SVM  Model

from sklearn.svm import SVC
classifier_svm = SVC(kernel="rbf")
classifier_svm.fit( X_train, y_train )
y_pred = classifier_svm.predict( X_val )

cm = confusion_matrix( y_val, y_pred )
print("Accuracy on Test Set for kernel-SVM = %.2f" % ((cm[0,0] + cm[1,1] )/len(X_val)))
scoresSVC = cross_val_score( classifier_svm, X_train, y_train, cv=10)
print("Mean kernel-SVM CrossVal Accuracy on Train Set %.2f, with std=%.2f" % (scoresSVC.mean(), scoresSVC.std() ))
```

```
Accuracy on Test Set for kernel-SVM = 0.80
Mean kernel-SVM CrossVal Accuracy on Train Set 0.81, with std=0.00
```

- ***Naïve Bayes***

Naive Bayes

```
#--------------
# Naive Bayes
#--------------
from sklearn.naive_bayes import GaussianNB
classifier3 = GaussianNB()
classifier3.fit( X_train, y_train )
y_pred = classifier3.predict( X_val )
cm = confusion_matrix( y_val, y_pred )
print("Accuracy on Test Set for NBClassifier = %.2f" % ((cm[0,0] + cm[1,1] )/len(X_val)))
scoresNB = cross_val_score( classifier3, X_train, y_train, cv=10)
print("Mean NaiveBayes CrossVal Accuracy on Train Set %.2f, with std=%.2f" % (scoresNB.mean(), scoresNB.std() ))
```

```
Accuracy on Test Set for NBClassifier = 0.80
Mean NaiveBayes CrossVal Accuracy on Train Set 0.81, with std=0.01
```

- ***KNeighborsClassifier***

K-NEIGHBOURS

```
#--------------
# K-NEIGHBOURS
#--------------
from sklearn.neighbors import KNeighborsClassifier
classifier4 = KNeighborsClassifier(n_neighbors=5)
classifier4.fit( X_train, y_train )
y_pred = classifier4.predict( X_val )
cm = confusion_matrix( y_val, y_pred )
print("Accuracy on Test Set for KNeighborsClassifier = %.2f" % ((cm[0,0] + cm[1,1] )/len(X_val)))
scoresKN = cross_val_score( classifier3, X_train, y_train, cv=10)
print("Mean KN CrossVal Accuracy on Train Set Set %.2f, with std=%.2f" % (scoresKN.mean(), scoresKN.std() ))
```

```
Accuracy on Test Set for KNeighborsClassifier = 0.81
Mean KN CrossVal Accuracy on Train Set Set 0.81, with std=0.01
```

Testing and Evaluation

After completing Model building using different algorithms, evaluate the accuracy of the model by building confusion matrix. As part of this project confusion matrix is built for each of the models, below is confusion matrix built on KNeighborsClassifier Model.

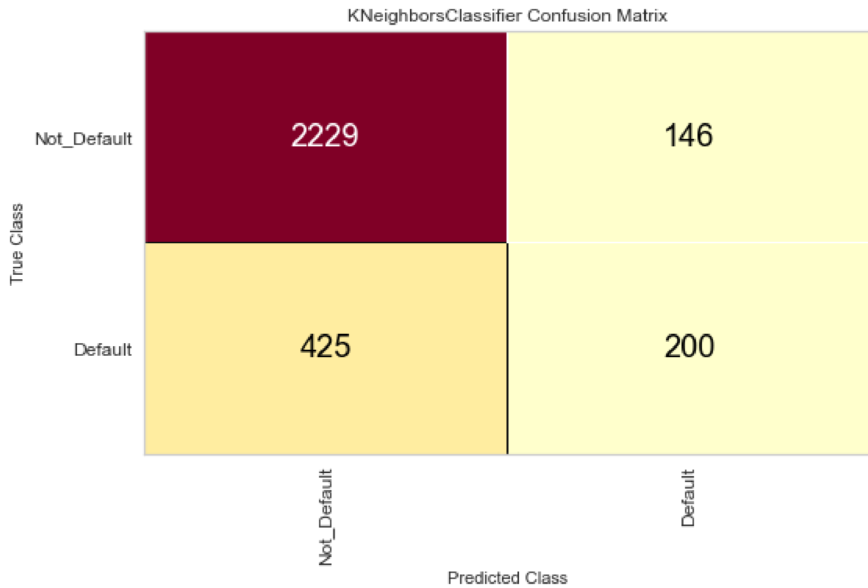


Chart shows the good precision and recall values and f1score 0.886 for not-Default cases indicates that model is performing as expected.

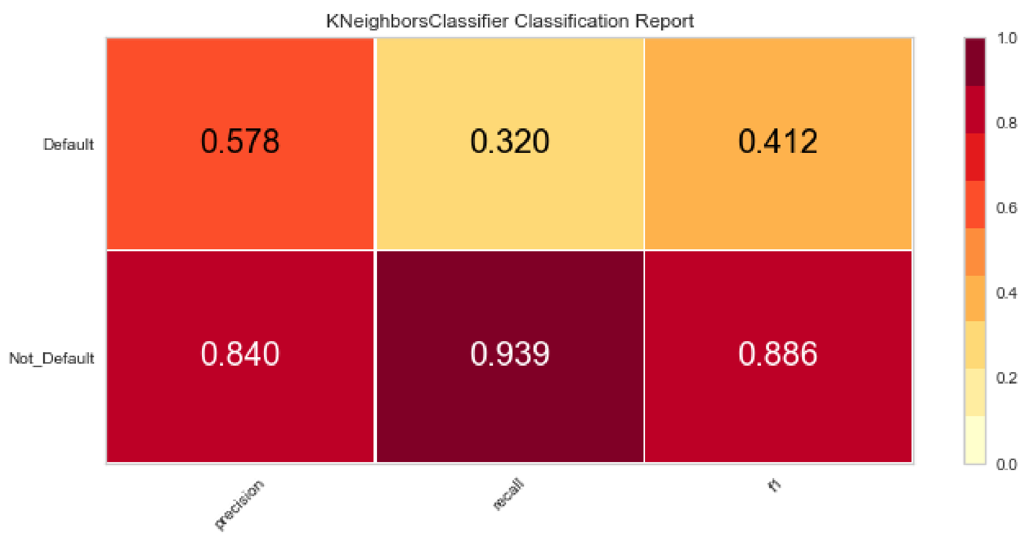
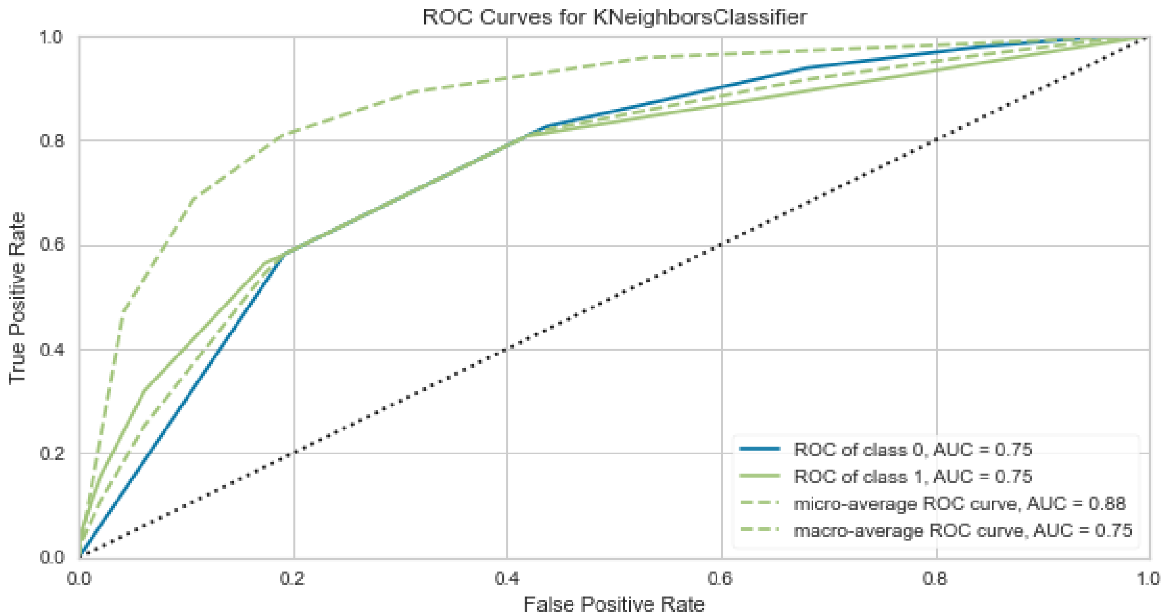
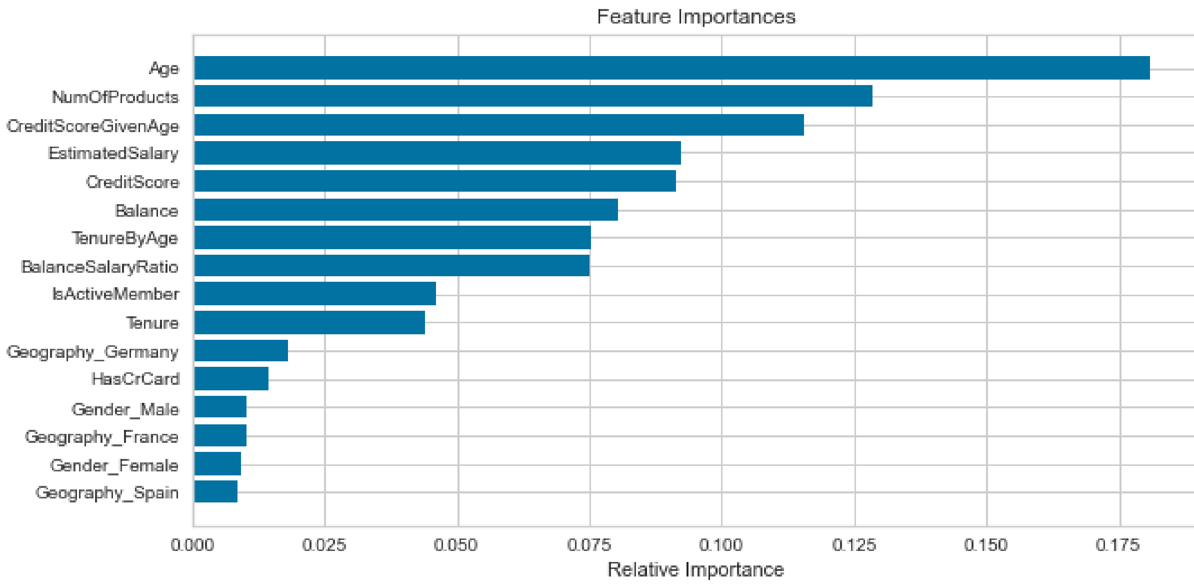


Chart shows the AUC (Area Under Curve) is at a value of 0.75 indicates that model is performing as expected.



Conclusion

Given chart shows that Age and Number of Products are most important factors in deciding the Bank Customer Churn followed by Credit Score and Estimated Salary.



Future Analysis Questions

I would like to continue my analysis and try to explore further to find answers to the given questions.

1. With the COVID situation in place, are these features still valid to predict the Bank Customer Churn.
2. As with the different type of Churns like Telecom Customer Churn, Credit Card Customer Churn, would these features stay in common across industry.
3. Will the feature importance change with respect to geographic location?

Reference:

1. 1. Soumya Sethuraman 2019 - Why Customers Leave & What Can Banks Do?

Why Customers Leave & What Can Banks Do? | Tiger Analytics

2. Sina Esmaeilpour Charandabi - Kent State University - 2020 - Prediction of Customer Churn in Banking Industry

(PDF) Prediction of Customer Churn in Banking Industry (researchgate.net)

3. Abbas Keramati, Hajar Ghaneei & Seyed Mohammad Mirmohammadi – 2016 - Developing a prediction model for customer churn from electronic banking services using data mining

https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6

4. Diana Kaemingk - August 29, 2018 - Reducing customer churn for banks and financial institutions

Reducing customer churn for banks and financial institutions // Qualtrics

5. Nelson Belém da Costa Rosa - November 2018 - Gauging and Foreseeing Customer Churn in the Banking Industry

TGI0223.pdf (unl.pt)