Predicting Churn for Bank Customers

Bhargava Gaggainpali  
DSC680 - Applied Data Science  
[bgaggainpali/bgaggainpali\_DSC680 (github.com)](https://github.com/bgaggainpali/bgaggainpali_DSC680)

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

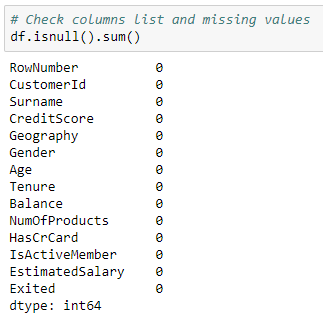
HasCrCard Customer has Credit card

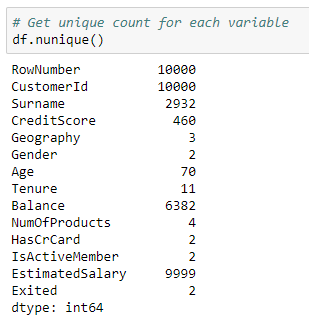
IsActiveMember Customer Active Member

EstimatedSalary Customer Salary

Exited Customer Churn value

Initial Observations





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

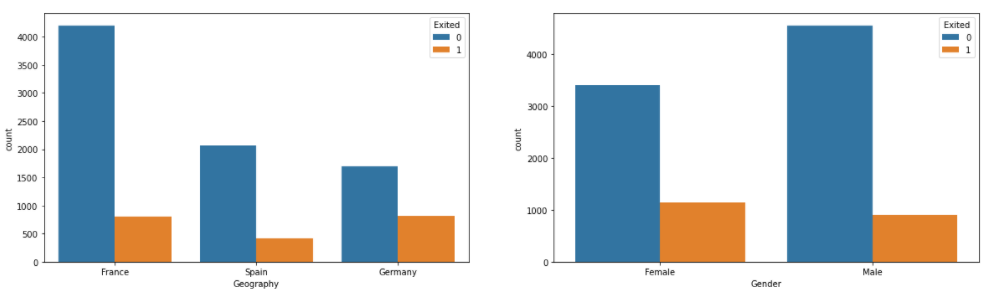
Balance

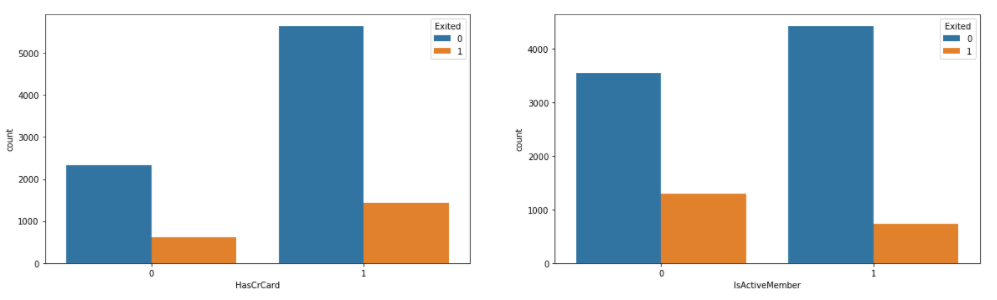
NumOfProducts

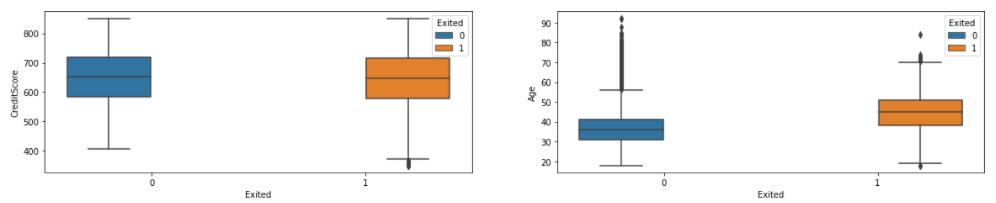
EstimatedSalary

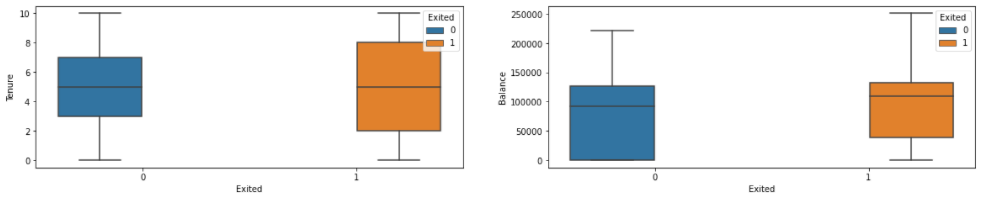
Exploratory Data Analysis

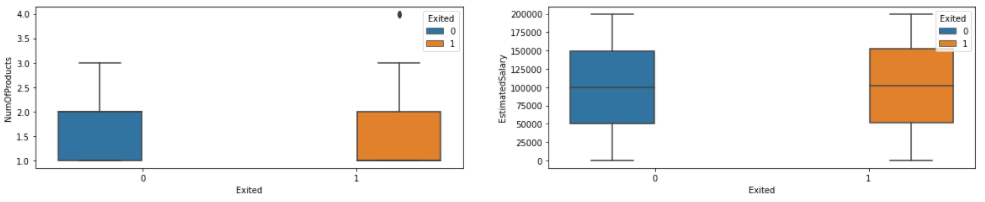
After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression to consider variables like Balance, Age, Gender, Tenure, CreditScore, HasCrCard and EstimatedSalary. I saw surprising stats when I used 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.

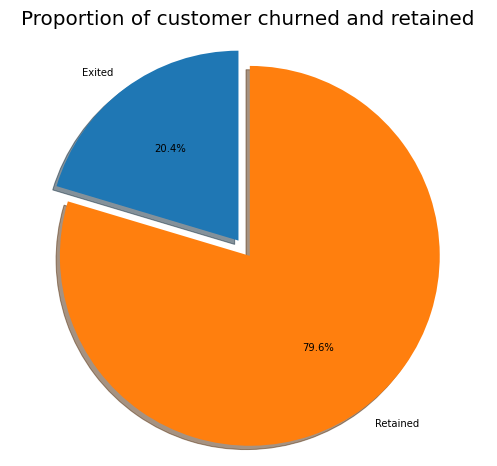


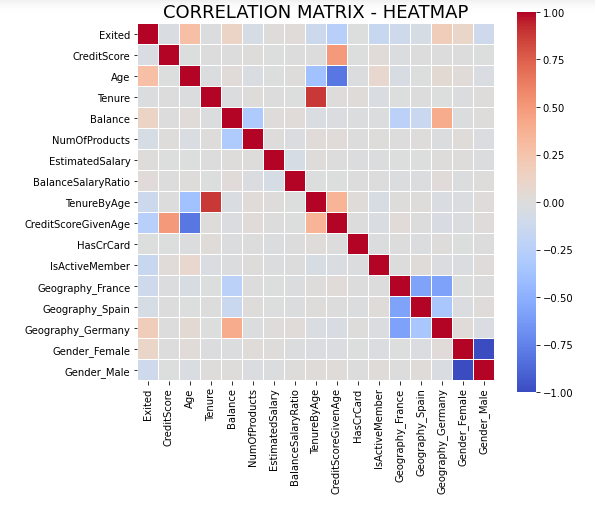










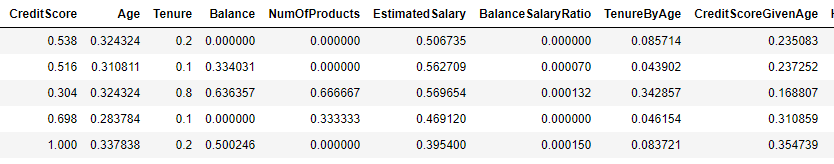


Few observations based on the above plots.

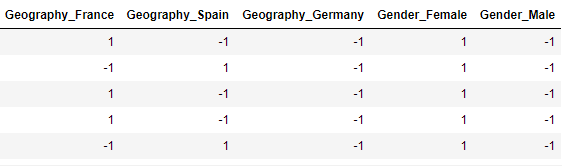
1. Majority of the data is from persons from France. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank.
2. The proportion of female customers churning is also greater than that of male customers
3. Majority of the customers that 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

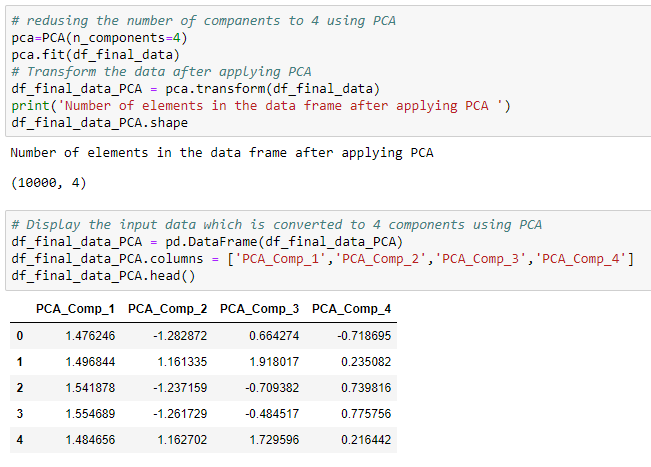
Applied MinMax Scaler to scale the numeric data variables.



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

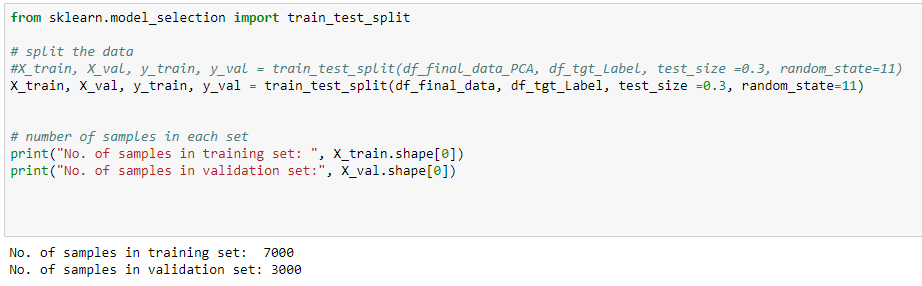


Applied PCA to reduce the number of Dimensions or Variables.



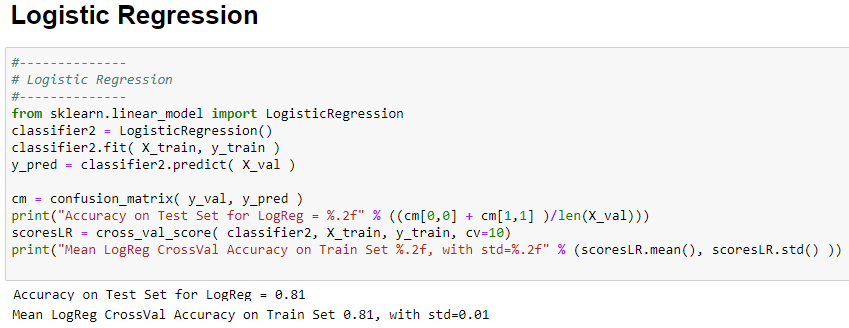
Model Development

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.

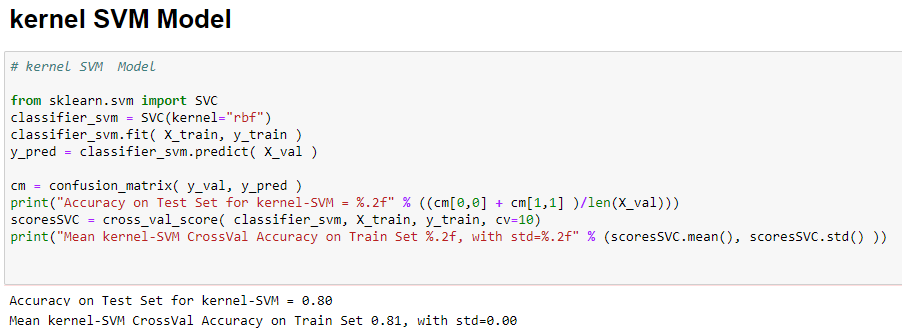


Below are four models

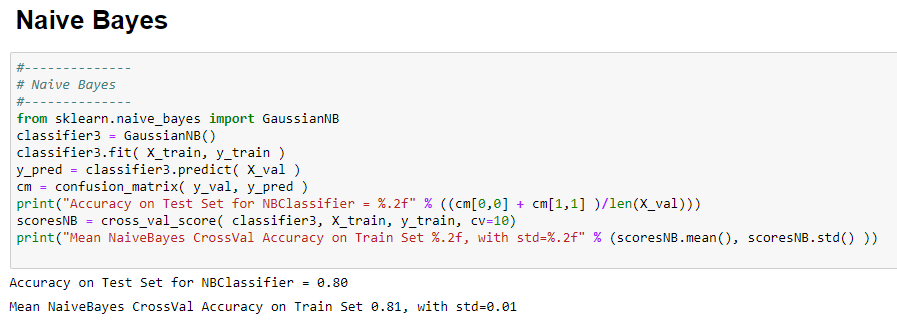
* Logistic Regression



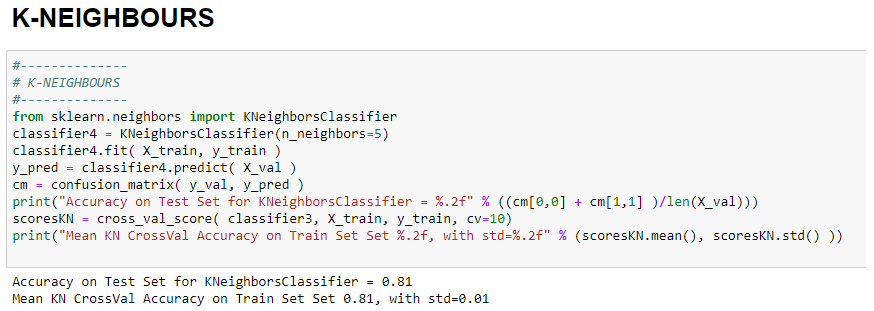
* Kernel SVM Model



* Naïve Bayes



* KNeighborsClassifier

****

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.

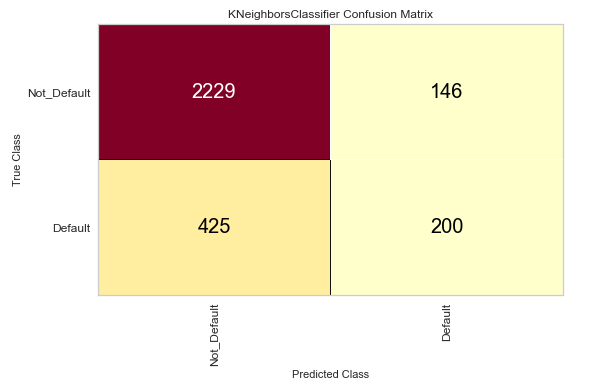


Chart shows the good precision and recall values and f1score 0.886 for not-Default cases indicates that model is performing as expected.

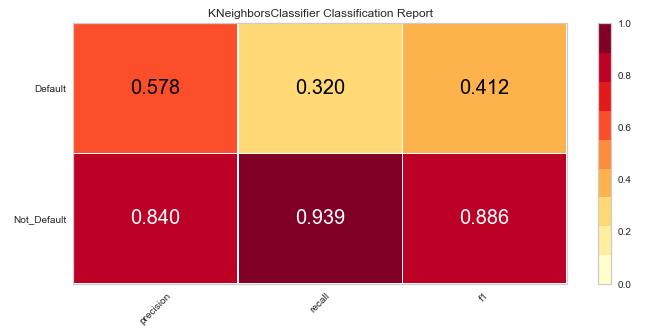
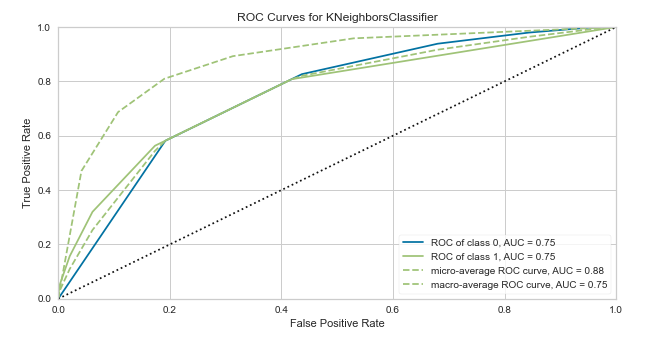
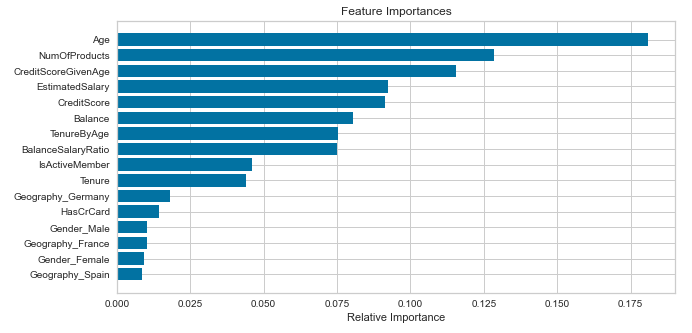
****

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](https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/)

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)](https://www.researchgate.net/publication/342424673_Prediction_of_Customer_Churn_in_Banking_Industry)

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](https://www.qualtrics.com/blog/customer-churn-banking/)

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

[TGI0223.pdf (unl.pt)](https://run.unl.pt/bitstream/10362/71584/1/TGI0223.pdf)