**Analytics Project – III**

**Bank churn analysis using Python**

**2205 MSA 6703 6E1 503W PRA 43156 GC**

**Project Report**

**Venkata Sai Suneeth Ravilla**

**Academic Supervisor:**

**Dr. John Chen**

**Summer 2020**

**Bowling Green State University, Ohio**

**Dataset Background:**

Churn is generally defined as a customer who stops using a product or service for a given period of time. Customer churn is a major problem of customers leaving your products/subscription and moving to another service. Banking is one of the highly competitive sector where customer relations is of the utmost importance for any bank. Marketing costs to acquire new customers are high. Therefore, it is important to retain customers so that the initial investment is not wasted.

**Variables:**

RowNumber, CustomerID, Surname, Geography, CreditScore, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited

**Goal:**

Aim of this analysis is to develop a model to predict the probability of a customer is likely to discontinue and preventing customer churn.

**Approach:**

In this report, we are first analyzing the data background how the patterns of the target variable lie and fitting the model to predict the exit customers.

**Major Findings:**

From the analysis in the report, based on the models the customer can be estimated either by using a Decision tree or by a random forest with better accuracy. But by comparing some key parameters we identified that Random forest gives better results. Hence the customer can be estimated using Random Forest might give good results.

Below report gives detailed analysis and approach followed for estimating customer.

**Data Preprocessing**

Row number is not required for our analysis. Customer Id is not required Surname has high cardinality and I will remove it from the modelling part. Geography has 3 categories. Germany, Spain, France. so we need to encode this. Gender is categorical and has two categories and we need to encode this. Number of products is categorical. Removing the observations who’s age is greater than 60. Also dropping the customers who has credit score less than 400.

**Missing data:**

Aim in this task is to analysis the data whether the data has any missing values to make sure the data has the values completely for the analysis.

No missing values are found in the given dataset. This ensures that there is no need to pass any mean or median values for the empty cells.

**Data Transformation:**

Given above that preprocessing helps in identifying irrelevant data and missing values, data transformation helps to transform the data into appropriate justification for the analysis.

**Normalization:**

This technique is performed in preprocessing which provides linear transformation on original range of the data

**Categorical data:**

The categorical data present in the dataset has various levels which belongs to specific sets and these need to be scaled for better fit of the model.

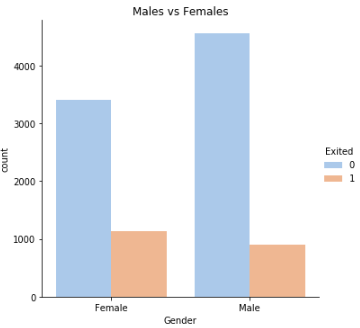
**Data Configuration:**

Upon checking the number of customers who has accounts in bank is 7614 and number of customers who exited the bank is 1903. Performing over sampling process to make sure the predicted model does not have any sampling error. Partitioning the data randomly into 80% train set and 20% test set. Scaling the complete data using StandardScaler function. We need to get the best algorithm which is giving us the best output as predictability. So, we need to try out with various models and then we can select the best ones which will give us the best performance based on accuracy.

**Data Exploration:**

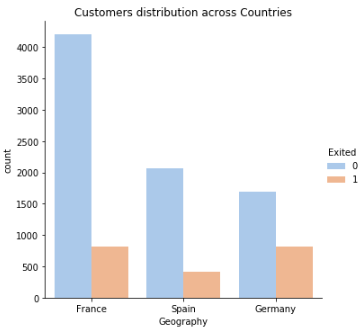
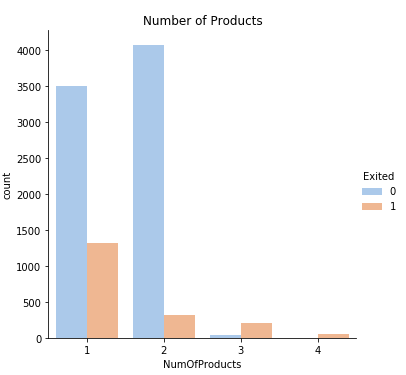
In data there are three countries, France, Spain and Germany Here, Exited means the customer has left the bank France has most number of customers in the data Even though France and Germany has same amount of churn, Germany has most churn when compared with total number of customers with other two countries. The proportion of churned customers is inversely related to the population of customers, indirectly the bank possibly having a problem in the areas where it has fewer clients.





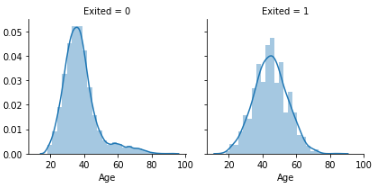
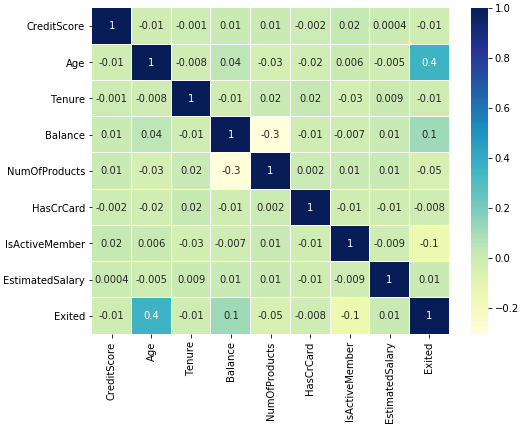
Male customers are high compared with female. Comparing the churn data among male and female customers, female customers are most likely to get churn than male customers.

Customer who is currently having back account but does not perform any transactions from some period of time is considered as in-active and vice versa. Comparing number of customers who has exited, Customers who are not active has exited more in number. Comparing non exited vs active member (Blue bar) we can see that more than 40% of the customers are in-active state which means there are likely to get churned.

Products means number of services some customer possess in a bank. Customer enrolling at least 2 products are high in number and their exit level is very low. But customer enrolling to only one service has high chance of getting churned.

We can see the Age distribution when people exited from bank. Most of the customers started to churn from 40’s. At the same time, we can see more number of customers present in the back from Age 30 to 40 It seems younger customers tend to stick with the company more compared to older customers.

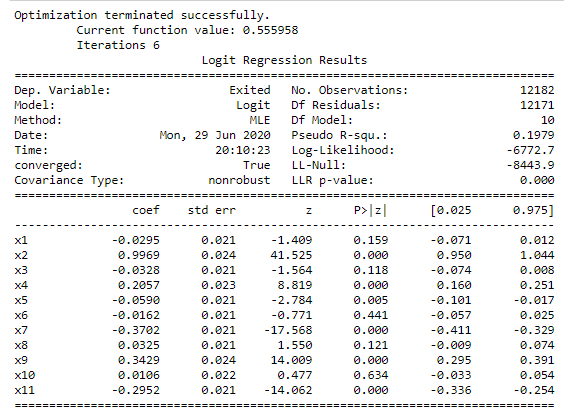
From correlation matrix, there is no much correlation among the variables which shows very low correlation. The highest positive correlation is 0.4 that is between Age and Exited. The highest negative correlation is -0.3 that is between Balance and number of products.

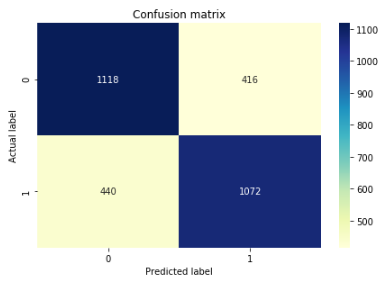


**Logistic Regression model:**

Logistic model is built to predict the response variable “Exited”. Almost all the variables are significant to the model expect variable x6 and x10 where x6 represents whether customer is active or not and x10 represents the country Spain. The fitted model is then used for prediction and from the confusion matrix we can say that the model has predicted with an accuracy of 71%

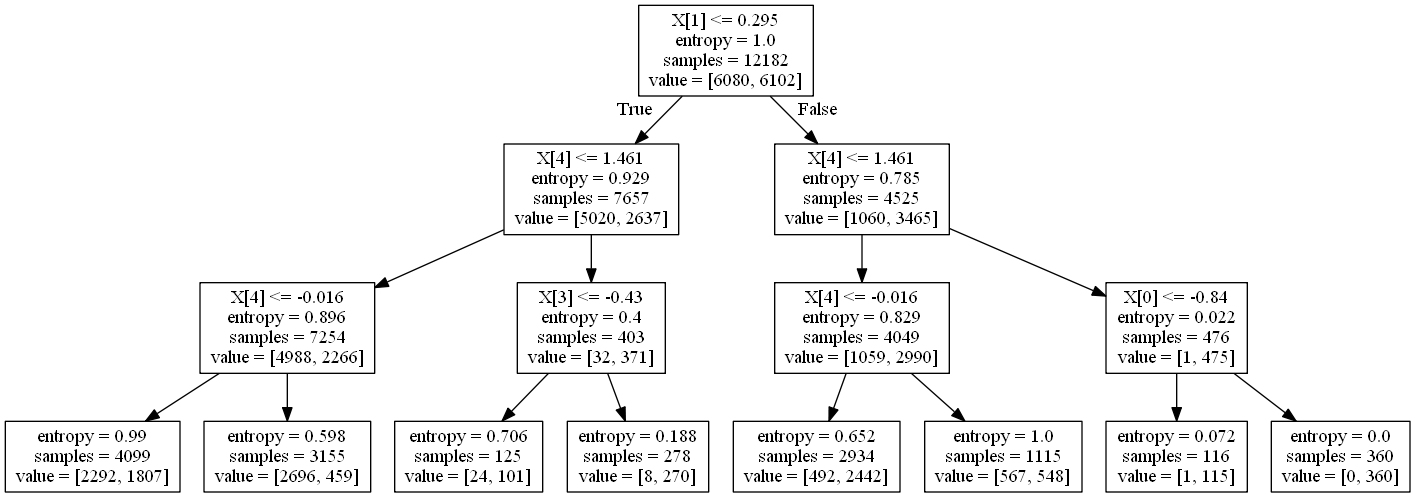
Out of 3046 customers, the model has predicted 1118 customers are likely to be exited from the bank. By checking the ROC, the model indicates that 72% of the time a randomly selected pair of subjects will be correctly predicted by the model.

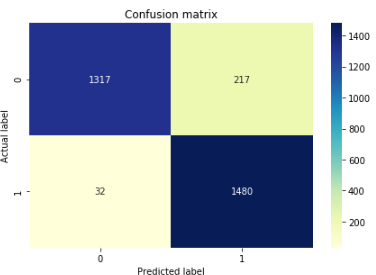




**Decision Tree Model**

Decision tree is built to predict the response variable “Exited”. The model starts with X1 variable splitting if the value is less than 0.295. Here X1 is credit score the value is less because it is scaled in pre-processing The fitted model is then used for prediction and from the confusion matrix we can say that the model has predicted with an accuracy of 91% Out of 3046 customers, the model has predicted 1317 customers are likely to be exited from the bank. Which is nearly half number of customers in the test set. By checking the ROC, the model indicates that 92% of the time a randomly selected pair of subjects will be correctly predicted by the model.

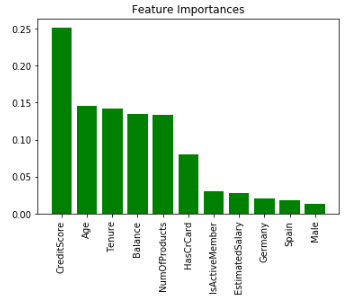
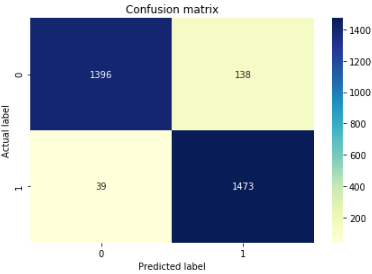




**Random Forest Model:**

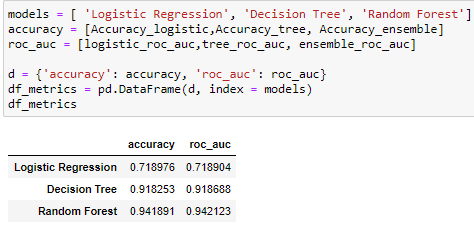
Random forest model is built to predict the response variable “Exited” and also to check the feature importance. From the feature importance graph, we can see that Credit score is the most contributing feature to the model then followed with Age, Tenure, Balance and no of products.

The fitted model is then used for prediction and from the confusion matrix we can say that the model has predicted with an accuracy of 94% Out of 3046 customers, the model has predicted 1396 customers are likely to be exited from the bank which is more than the prediction of decision tree model. By checking the ROC, the model indicates that 94% of the time a randomly selected pair of subjects will be correctly predicted by the model.

**Conclusion:**

* In this analysis of bank customer’s data, the models were built to predict how likely a customer is going to be exited from the bank.
* In exploratory analysis, we found that the ratio of customers churned in Germany is higher than other two countries, also the we found out that the female customer are the most likely to churn and the customer using only one bank service (product) are likely to get churned.



* Almost 40% of the customers are inactive in the bank and the ratio of exiting the bank for inactive members is high.
* After building several model we ended up with Random Forest which performed better than others. The accuracy of Random forest and decision tree is higher but Random forest is best with 94%. It has classified 1473 as good customers and 1396 customers are likely to be exited from the bank. Using most efficient metric ROC the model indicates that 94% of the time the model will predict correctly.