STAT 515 Final Project Report

I. Introduction

This dataset which is on Customer Churn or also called as attrition is about how customer stops using a particular service or a company's products. This can be measured at different stages, often evaluated for a specific time period like, monthly, quarterly, or yearly bason on actual usage or failure to renew the product. When customers start purchasing the product, that product's growth rate increases. In the same way, customers might even discontinue using a product either because they do not need the product or because they switched to an alternate company's product, or they are not happy with their experience with the product, or they can no longer afford the product. This is the reason churn analysis is essential to help solve business problems and make decisions which would lead to increase in customer satisfaction. We do bank customer churning, in which, initially we perform exploratory analysis and visualize factors that contribute to churning. This in turn helped us to build a model to predict whether or not a customer will churn. This is a classification problem; hence, churned customers are more important for the bank.

II. About the dataset

Source of this dataset is from Kaggle website. This dataset comprises of 10,000 customers data with 14 different factors contributing to churn. The deciding factor in this dataset is 'Exited', in which '0' stands for NO and '1' stands for YES. The dataset also consists of other instances such as Customer ID, Surname, Credit score, Geography, tenure, gender, balance, age, number of products, has credit card, is an active member and estimated salary.

III. Exploratory Analysis

The plots below show the exploratory analysis of the dataset, this is done before fitting statistical model. Analysis is the most essential part for data exploration as it gives in-depth knowledge about the dataset. Different visualization techniques are applied based on whether the data is categorical or numerical. We plotted density plot, box plot, and histogram as part of summary statistics on some columns.

Density plot:

Shows the smoothed distribution of continuous or time period data points along numeric variables. The peaks of density plot displays where the values of the data are concentrated over an interval of time. A density plot better displays the distribution shape over a histogram because the peaks in a density plot are not affected by the number of bins used as in a histogram. As in a histogram the more the number of bins the shape of distribution is good enough.

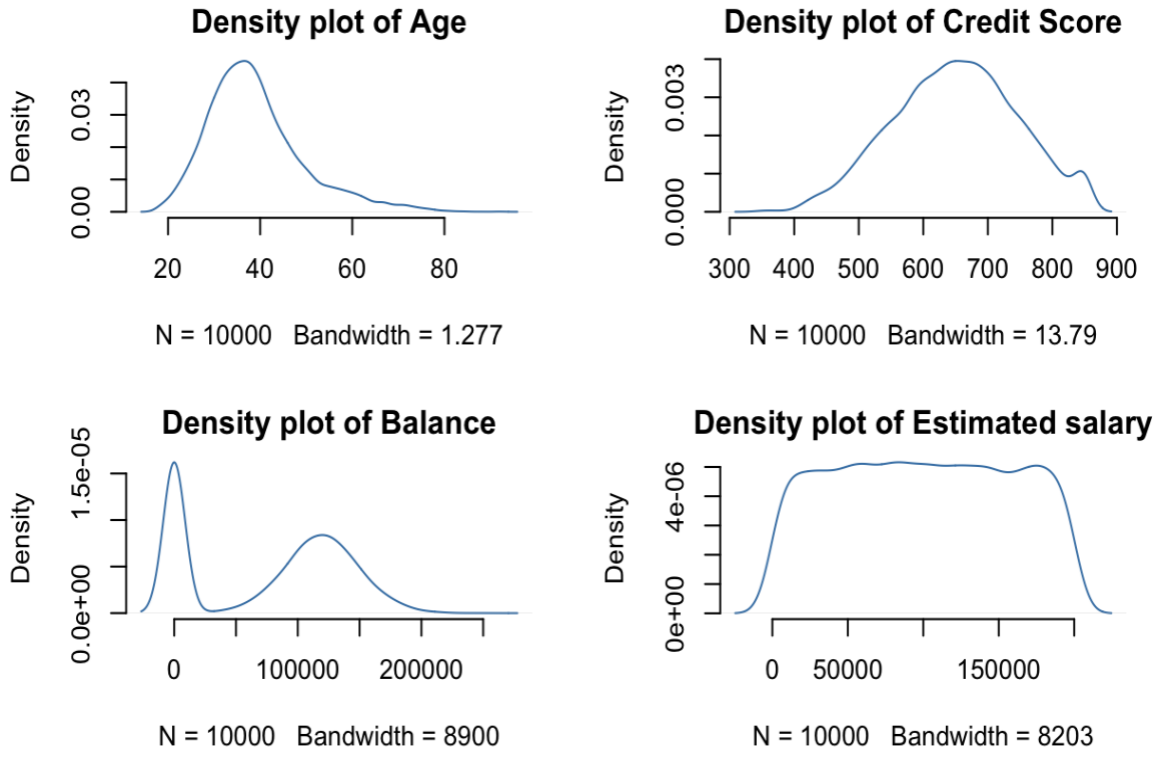


Fig.1

Skewed distribution is a condition where data in the dataset is more dominant to one side of the graph when compared to the other side. Density plot is considered for 10,000 instances and the bandwidth indicates how close the density can be to match the distribution. From the above density plot it can be inferred that, the attribute Age is right skewed which means the mean of the data is less than that of the median, credit score is left skewed where its mean is greater than its median, the attribute balance is in bimodal shape with two peaks which means it has its distributions over two centers and estimated salary is uniformly skewed where in all observations are equally spread across the dataset.

Box Plot:

Box plot is used to observe relation between features. A box plot divides the data into four sections each consisting of approximately 25% of data. These plots are used to easily identify the outliers in a dataset.

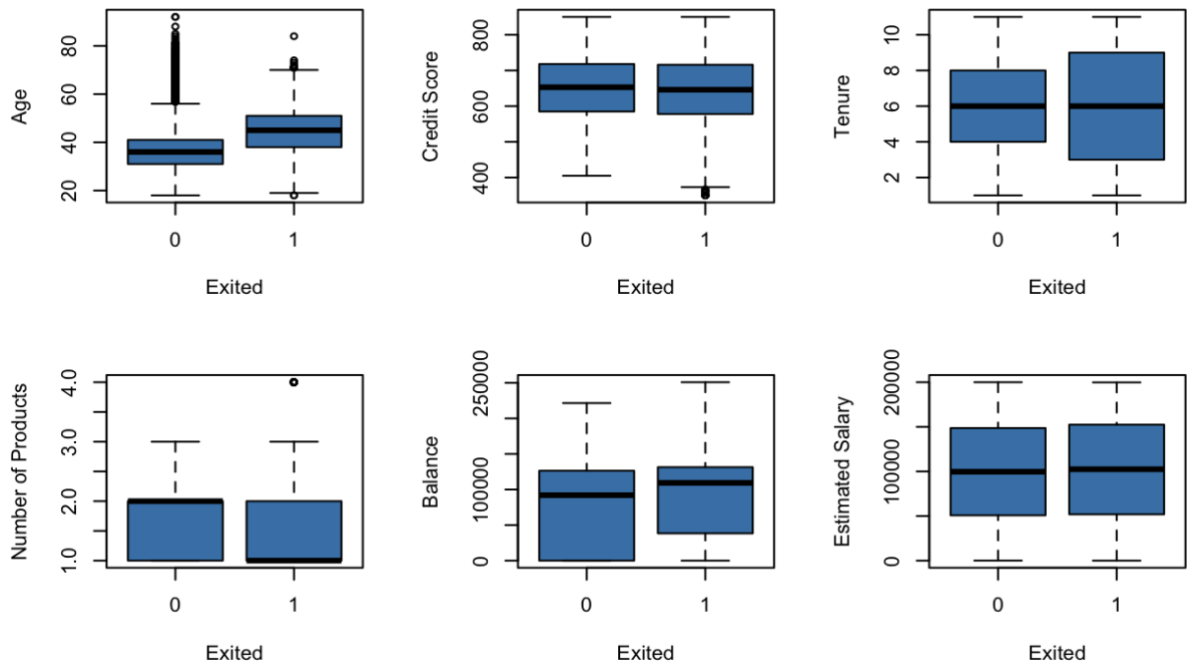


Fig.2

It is clear from the above box plots that there are outliers in the attributes age, credit score, number of products. Outliers are noisy data which would have heavy impact on the distribution if they were present, hence, box plots are constructed to check for outliers. After performing the analysis there were some intriguing results found. Almost all the plots are normally distributed expect number of products, balance. The attribute number of products is skewed left for churned data and skewed right for unchurned data. Data for balance is skewed left. In case of age groups where the customers with age 40 or more are more likely to churn. This indicates that there is a change in preference with age. The credit score is higher than 600 for churned customers, balance of churned customers is nearly $1 million and the estimated salary for churned customers is nearly $1.5 million. Hence, it can be concluded that salary does not contribute to churning.

Histogram:

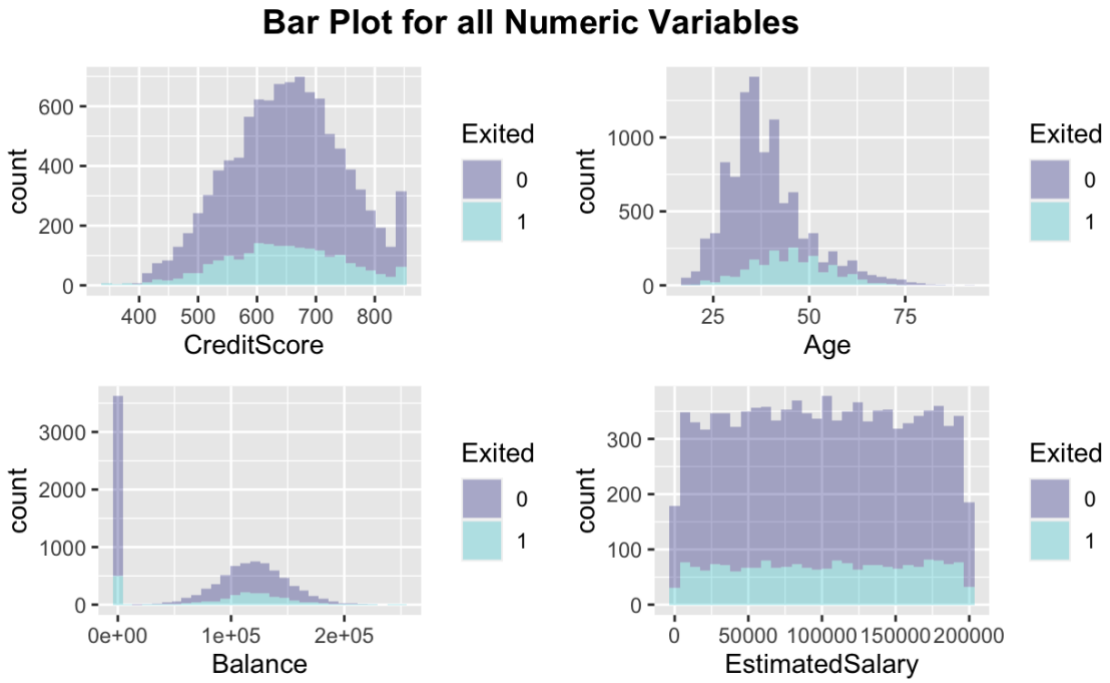


Fig.6

From the above plots which are plotted for the numerical variables, the following observations have been made.

1. Credit Scores seems like to be left skewed.
2. Most of the customers are between the age 28-40.
3. Customer Balance seems to be symmetrically distributed.
4. There is not much variation in the estimated salary since all the values are lying between 300k to 400k.

Correlation Matrix:

There's no multicollinearity implies We don't see any high correlation between the continuous variables.

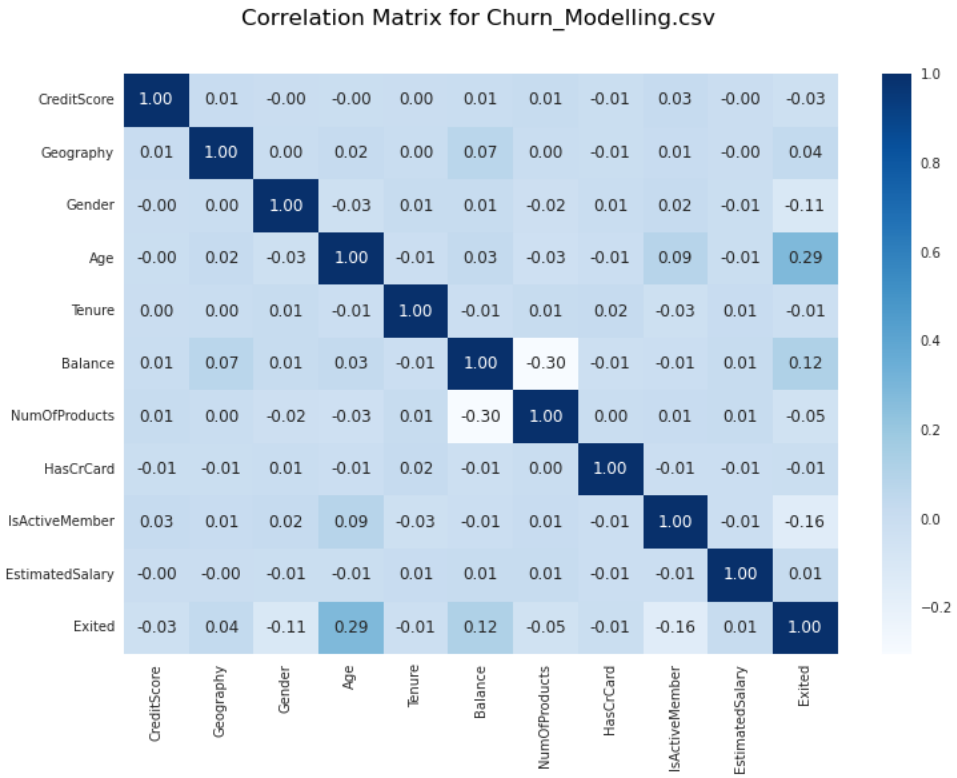


Fig.4

It can be observed that the predicted variable is mainly affected by age as its correlation score is greater than the rest of the attributes, i.e., 0.29.

We plotted a bar graph for the predicted variable, ‘Exited’, to observe the number of churned and unchurned customers.

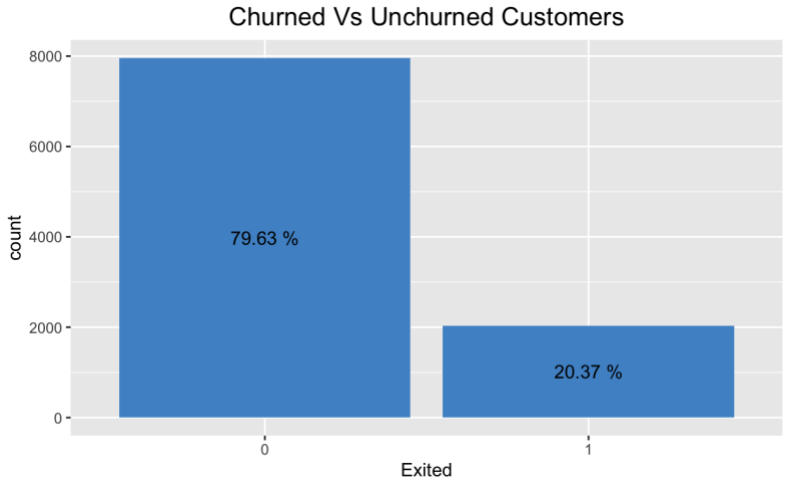


Fig.5

The above graphs are comparison of customers churned and remained. Here in this analysis, the target variable is "Exited". From the above chart we can depict that roughly 80 percent of the customers are remained while remaining 20 percent of the customers churned. This is clearly an unbalances data, and we need to treat that another time. Here exited which is the target variable is stored as numeric. So, it is converted to the levels related to remain and churn.

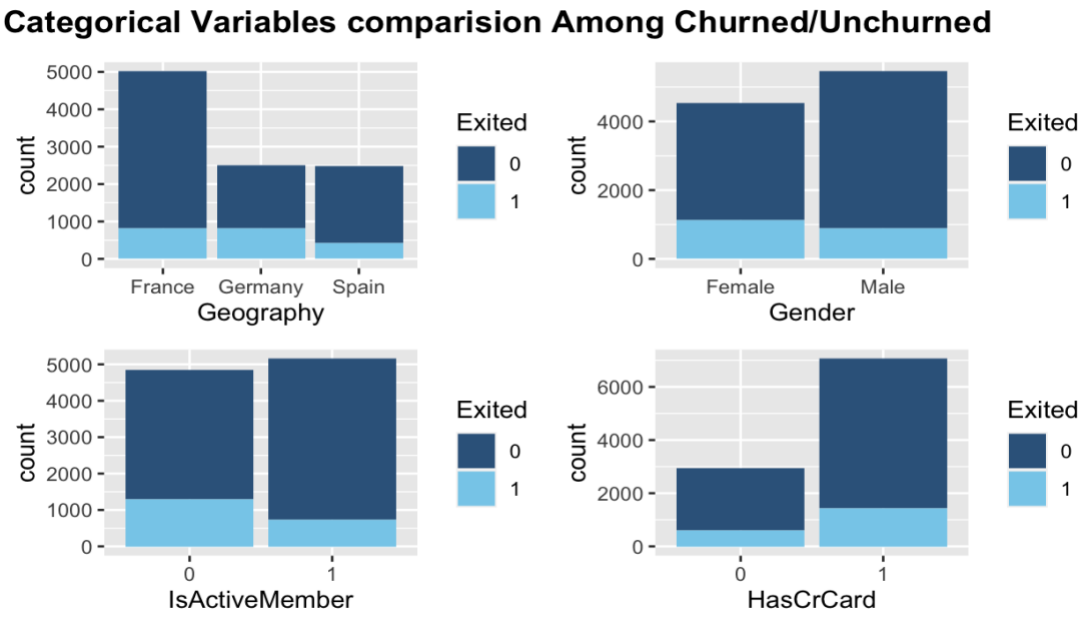


Fig.7

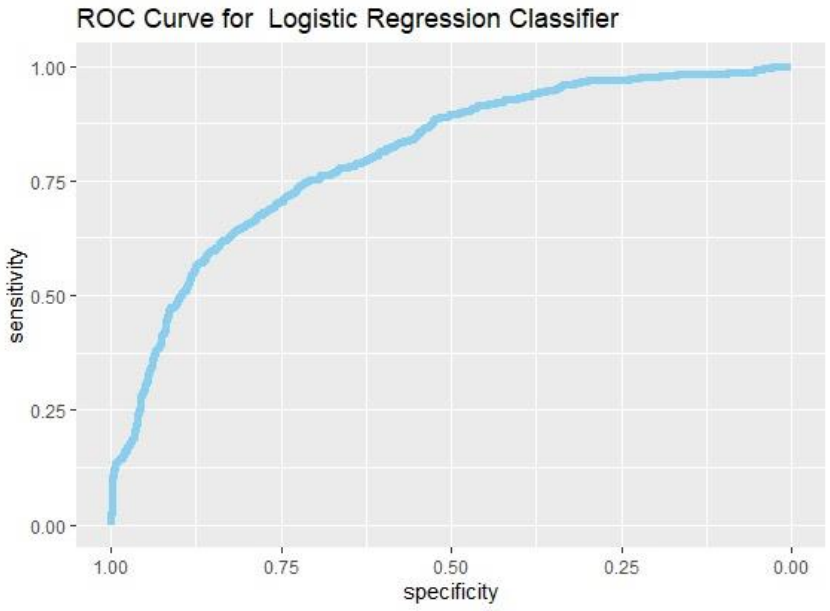
From the above plots we can infer that,

1. The total number of customers retained from the bank has recorded to be highest for France and the number of customers exited is highest for Germany which means the bank must focus more on the customers form Germany followed by France.
2. The Proportion of the female customers churn is greater than the proportion of male customers. [1]
3. Customers who have credit card churned more when compared to the customer who doesn't have a credit card.
4. Customers who are inactive has churns compared to the one who are active.

IV. Regression Models:

1. Logistic Regression

The dataset predicts churned and unchurned customers using binary variables with two outcomes 1 and 0. While fitting a logistic regression model, we took a look at statistical effects of balancing data. To fit the model, balanced or unbalanced data is used for fitting logistic regression models which show effect on conclusions drawn from the model.




```
[1] "Confusion Matrix for Logistic Regression
                Actual
Predicted         0         1
        0         2293      96
        1         463       148
```

Fig.8

The Logistic Regression is a regression model in which the response variable (dependent variable) has categorical values such as True/False or 0/1. It actually measures the probability of a binary response as the value of response variable based on the mathematical equation relating it with the predictor variables. [2]

Here using logistic regression, initially we built a model to discover the impacting depending on factors .

Then we try fitting a new model only with the impacting factors and calculate the accuracy of the model.

Here ROC curve is used to show the performance of the model.

2. Lasso Regression

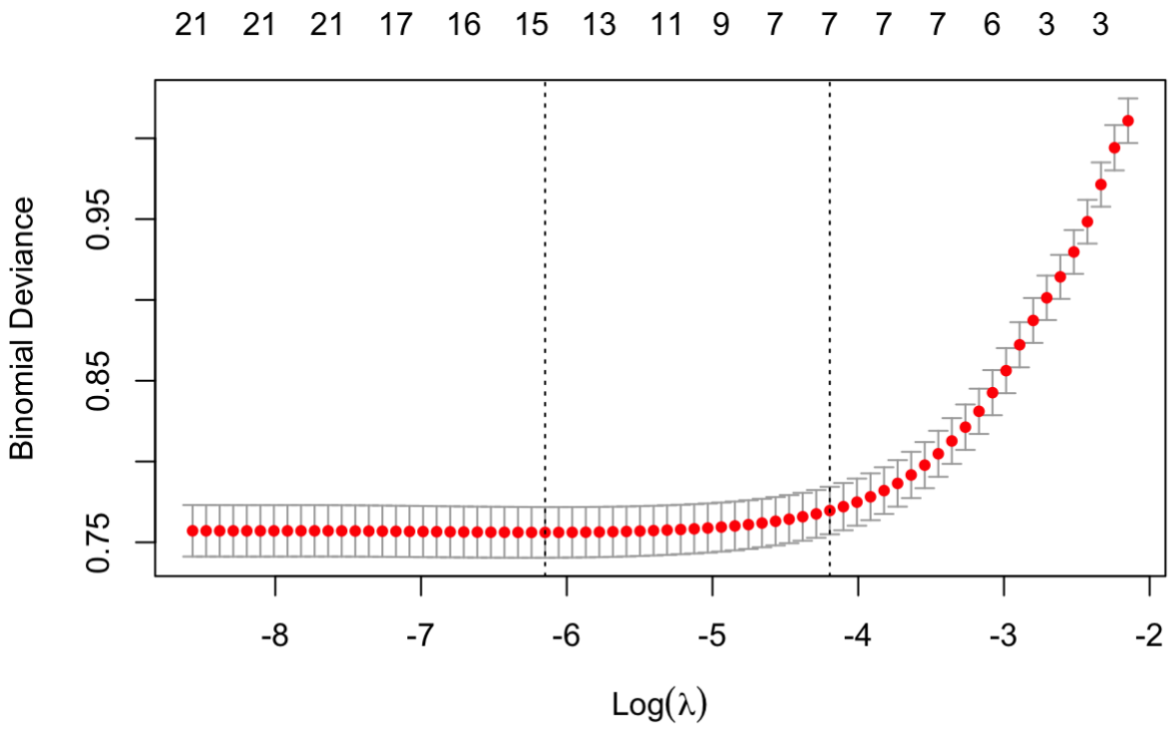


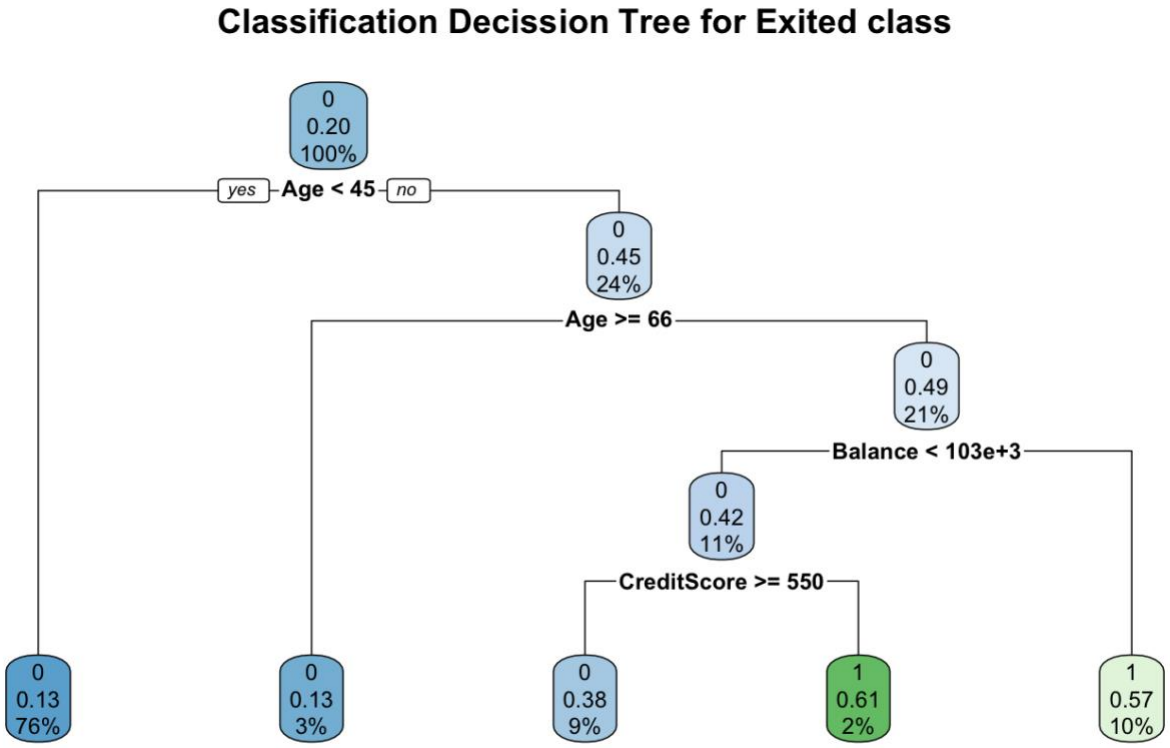
Fig.9

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. [3]

This technique is widely used when the data is highly colinear or correlated , and all the insignificant attributes are turned to 0's and 1's.

A slightly biased and low variance model is built using this technique.

Classification using Decision Tree:




```
[1] "Confusion Matrix for Decision Tree"
          Actual
Predicted        0        1
        0        2243     441
        1        146      170
```

```
[1] "Decision Tree Accuracy 0.804333333333333"
```

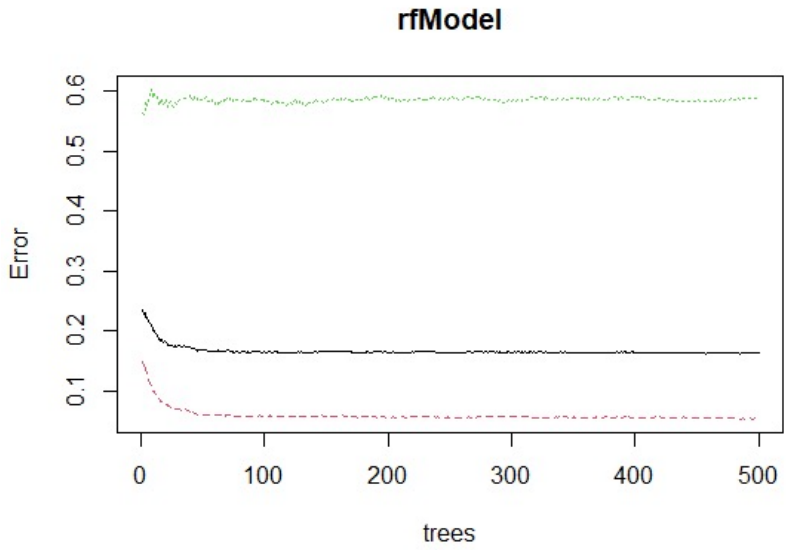
Fig.10

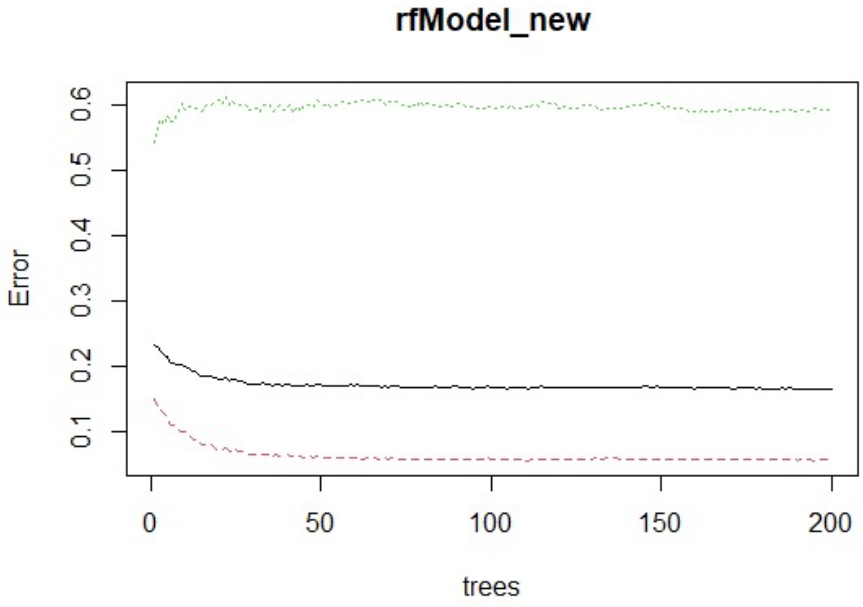
Considering the most significant variables for the target value, which is exited, which are Age balance and credit score, decision tree classification has been done.

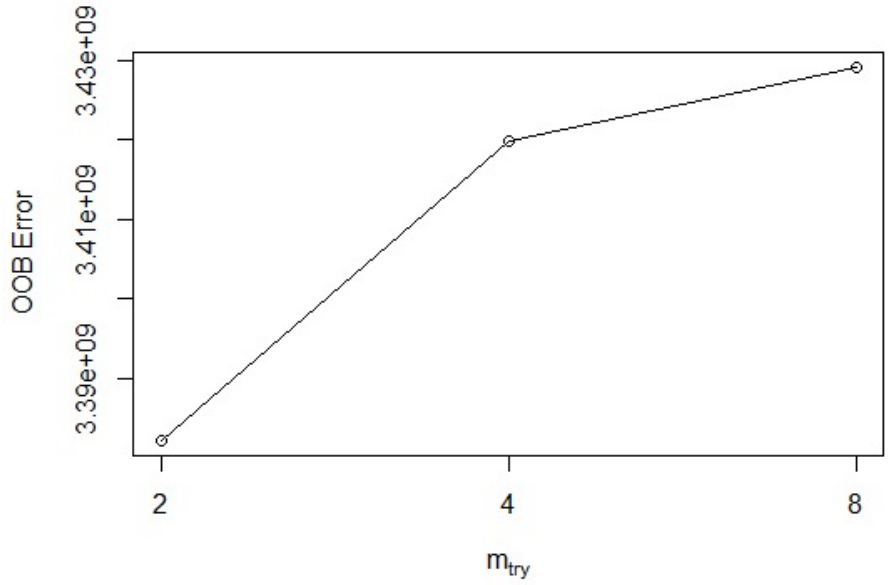
1. Out of these two variables, Age is the most important variable and considered to be best attribute for splitting to predict the customer churn or not churn.
2. If a customer whose age is less than 45 are more likely to churn which is 76%.

3. If a customer whose age is less than 65 are also likely to churn whose proportion is around 3%.
4. On the other hand, Customer whose age is greater than 66 and whose balance is less than 130k, and whose credit score is less than 550 are likely to churn whose proportion is around 9%.

Random Forest:








```
[1] "Confusion Matrix for Random Forest
                Actual
Predicted         0         1
        0         2231      158
        1         350       61
```

Fig.11

In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. [4]

An error estimate is made for the cases which were not used while building the tree. [5] That is called an OOB (Out-of-bag) error estimate which is mentioned as a percentage.

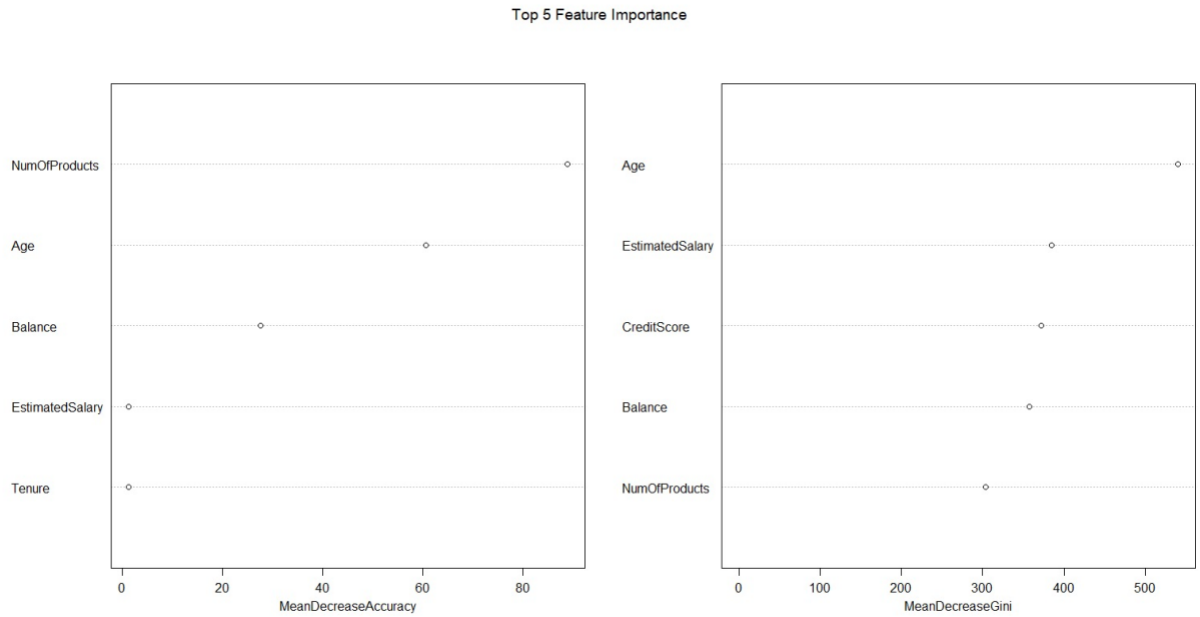
The R package "randomForest" is used to create random forests. [6]

To overcome the pitfalls of Decision tree model and also to increase the accuracy , we come up with Random Forest Model.

The above plot helps us to understand the count of trees, we can observe that, the error rate decreases and becomes constant after 200 trees.

This plot helps us to choose best value for mtry. As OOB error rate is least for mtry=2, we bulit our final model with it.

V. Conclusion



[1]	Model	Accuracy
1	Logistic Regression	81.36%
2	Decision Tree	80.43%
3	Random Forest	83.36%

Fig.12

From the above analysis, we can see that Logistic Regression, Lasso Regression, Decision Tree and Random Forest were used for customer churn analysis for this dataset .

Throughout the analysis, we have learned several important things:

- Features such as tenure, Age, Balance, Estimated Salary and Number of products play a role in customer churn.
- There does not seem to be a relationship between gender, location, and churn.

References

[1] *How Machine Learning Can Help with Customer Retention | by ...*, https://towardsdatascience.com/how-machine-learning-can-help-with-customer-retention-6b5bf654e822.
13
[2] *R - Logistic Regression - Tutorialspoint*, https://www.tutorialspoint.com/r/r_logistic_regression.htm.
13
[3] *What is LASSO Regression Definition, Examples and Techniques*, https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/.
13
[4] *[100% Working Code] - R Random Forest - r tutorial - By 10 ...*, https://www.wikitechy.com/tutorials/r-programming/r-random-forest.
13
[5] *An error estimate is made for the cases which were not ...*, https://www.coursehero.com/file/p4tko070/An-error-estimate-is-made-for-the-cases-which-were-not-used-while-building-the/.
13
[6] *R Random Forest in R Programming language Tutorial 20 ...*, https://www.wisdomjobs.com/e-university/r-programming-language-tutorial-1579/r-random-forest-18326.html.
13
[7] Predict customer churn – logistic regression, decision tree and Random Forest. DataScience+. (n.d.). Retrieved December 13, 2021, from https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/