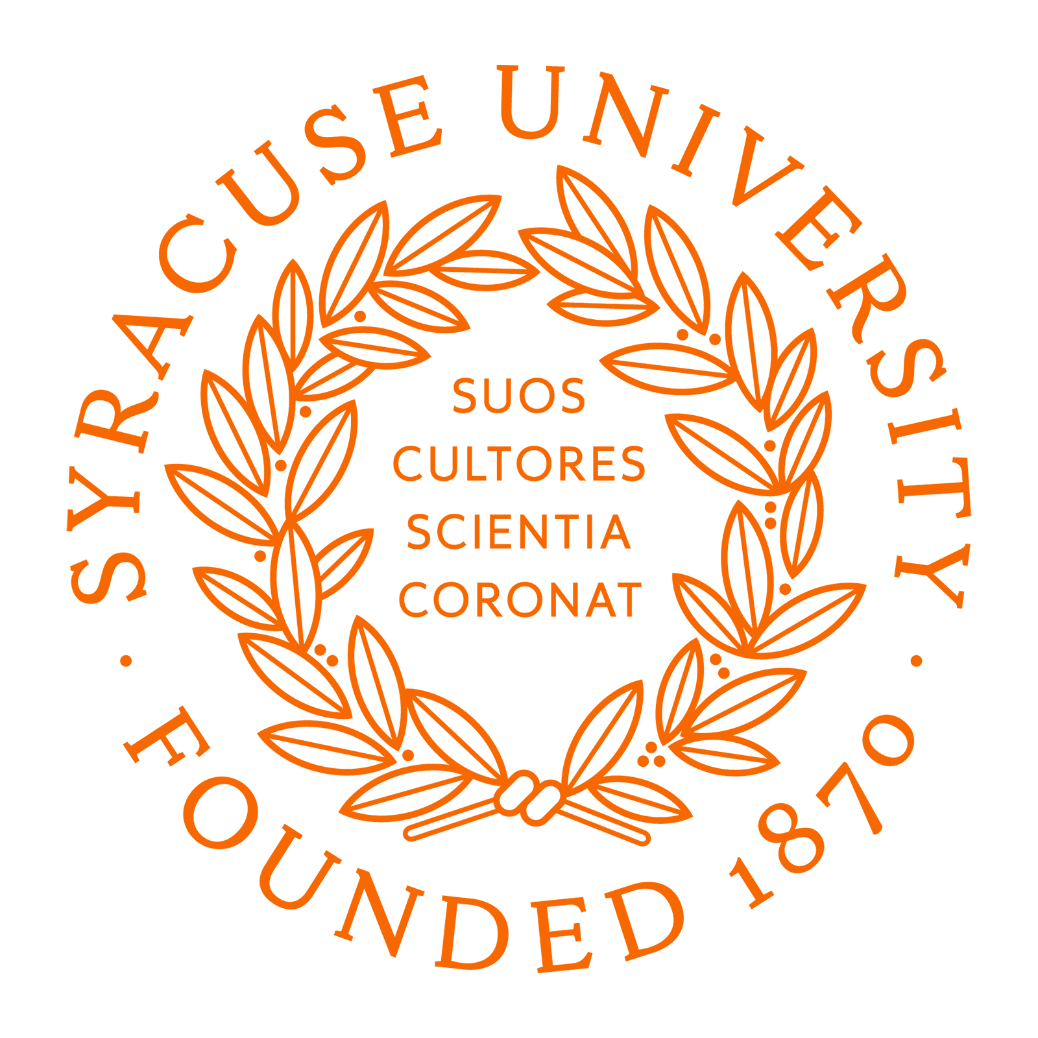
**FINAL REFLECTION PAPER**



**PAI 793**

**PREDICTIVE ANALYTICS**

**Name : Sonamgyalpo Sherpa**

**SUID – 398497433**

**Bank Customer Churn Prediction**

**INTRODUCTION**

Banking is a popular business field which has experience undesired change through the decades. Today many banks, however, have no competitive advantage with a large number of customers, particularly in resolving one of the most recognized problems – customer churn.

Whilst maintaining current customers and thereby growing their lifetime value is something that everybody knows is significant, the banks will do nothing with customer churn if they don't see it coming in the first place. This is where the right time to predict churn is important particularly when there is no clear feedback from customers. The early and precise forecast helps CRM and customer service departments to communicate with the customer creatively and proactively. Indeed, 11 per cent of the churn can be stopped by merely meeting the client early enough.

**Why Predictive Analytics?**

The first step to address customer churn is to optimize the Customers’ Experience by using data sets to obtain a global view of the customer decision journey. New technologies have given consumers the power to be informed and to compare prices, complain loudly and decide.

In this regard, companies should understand customers’ expectations from the very beginning to gain competitive advantage and reduce the probabilities of cancellation. This is where, predictive modelling techniques can come in handy to solve the business problem. With a huge number of banks in the modern era, advanced Machine Learning Algorithms can be crucial in deciding the success of a bank against in competitors and retaining customers.

**Business Problem**

Binary Classification is the Machine learning task. Regarding the business issue; a bank is facing massive customer attrition due to latest market trends and high increase in competitors. They wish to classify the customers into two categories, if they stay or leave the bank. Moreover, the customers who are more likely to opt out from the services of the bank, can be targeted with personalized marketing.

**Data Set (10000 rows )**

|  |  |
| --- | --- |
| Column | Description |
| Row Number | Row Numbers from 1 to 10000 |
| CustomerID | Unique Ids for bank customer identification |
| Surname | Customer's last name |
| CreditScore | Credit score of the customer |
| Geography | The country from which the customer belongs |
| Gender | Male or Female |
| Age | Age of the customer |
| Tenure | Number of years for which the customer has been with the bank |
| Balance | Bank balance of the customer |
| NumProducts | Number of bank products the customer is utilising |
| HasCrCard | Binary Flag for whether the customer holds a credit card with the bank or not |
| IsActiveMember | Binary Flag for whether the customer is an active member with the bank or not |
| Estimated Salary | Estimated salary of the customer in Dollars |
| Exited (Response Variable) | Binary flag 1 if the customer closed account with bank and 0 if the customer is retained |

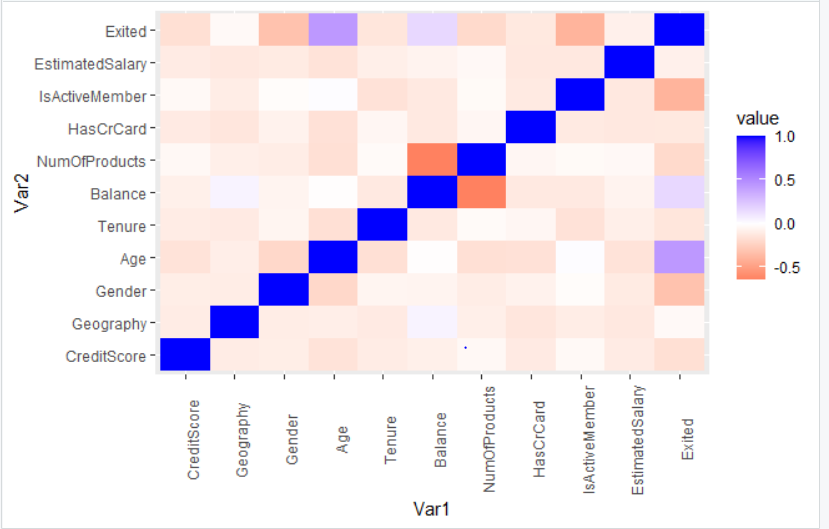
This data set contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer. This data set is obtained from Kaggle - <https://www.kaggle.com/shrutimechlearn/churn-modelling>

**Variables**

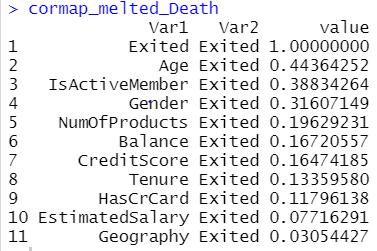
The first three variables – “*Row Number, CustomerID, Surname”* were not taken into consideration while building the models, since they had no significance to the response variable.

The variables “*Geography and Estimated Salary”* were excluded from the models, since they had very low correlation to the Response variable ( less than 0.1) and did not indicate an increase in the predictive accuracy or AUC of the test data set.

**Correlation Matrix**



The above correlation matrix indicates no high negative or positive correlation between the any of the independent variables. There may seem to be high correlation between *NumofProducts and Bank balance* of customers but including these two variables in the models did not improve the predictive performance.



The above correlation table against the response variables *Exit*, indicates *Age* has the highest correlation and the last two variables have been excluded from the models.

**Standardization and Cross – Validation Approach**

All models have been implemented using K-fold Cross Validation, with the value of K being 5 or 10 (Since, this is the optimal value). Since, KNN algorithm considers distances, the variables have been scaled using the preprocess feature in caret package. Scaling the other variables in the models provided no significant difference in the result.

\*\*The predictor variables used in the models is of type ‘factor’ or ‘numerical’ and the data has been partitioned in a 75:25 proportion with replacement = True.

**Models Implemented**

1. Logistic Regression
2. K-Nearest Neighbors
3. Random Forest
4. Decision Tree
5. Bagging
6. Boosting -

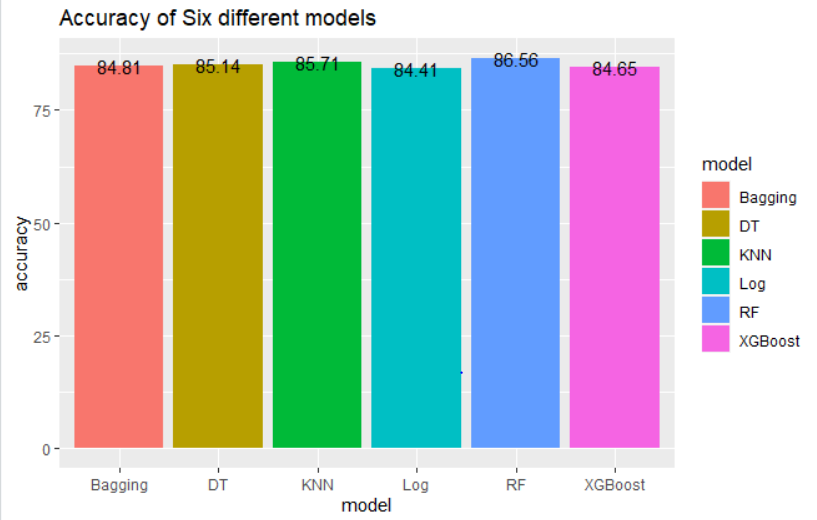
These models are crucial in resolving a classification problem. Logistic Regression, Decision Trees provided more interpretability indicating customers who brought 3 bank services or those who were elder customers were more likely to churn, while random forests and ensemble methods provided higher prediction accuracy but lower interpretability.

**Evaluation Method**

The model has a baseline of 79% for predicting customers who stay and 20% for customers who churn. Therefore, models need to perform better than this value. To evaluate the models, Classification Accuracy and AUC-ROC have been used to predict the response in the test set. Two methods have been chosen to evaluate because one provides the best accuracy at a fixed cut-point while the latter applies across many cut-points.

Also, to detect the customers who are likely to churn, the specificity value from the confusion matrix has been considered. All algorithms are carried out using the caret package in R.

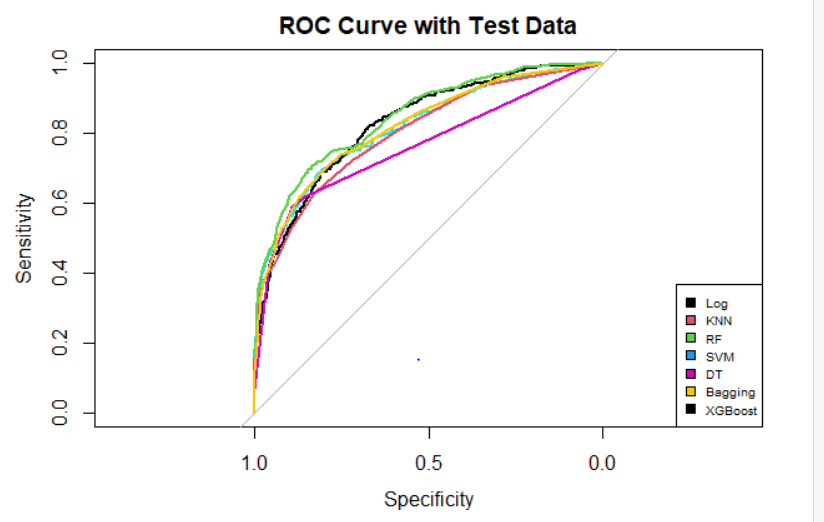
**Accuracy – Calculated from Confusion Matrix of test set**



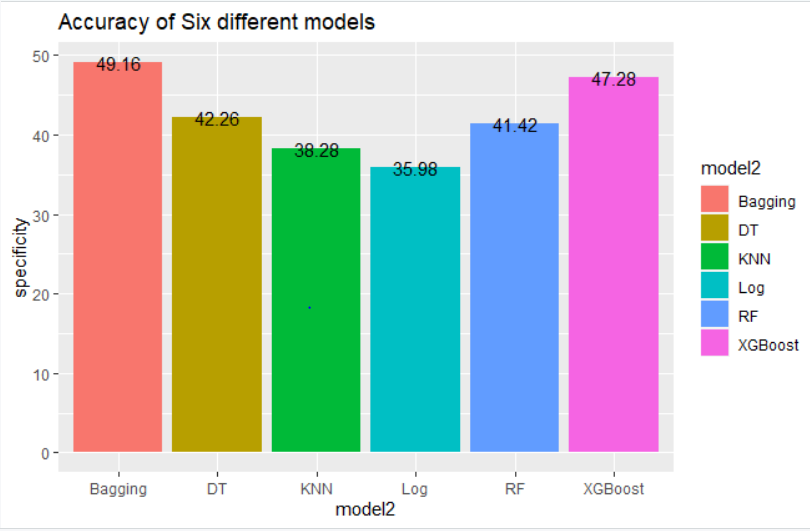
The Random Forest Model has the highest Accuracy of 86.56%.

Logistic Regression has the lowest accuracy of 84.41%.

**AUC – ROC**



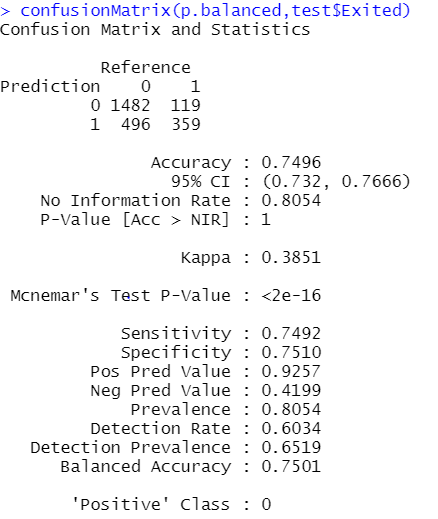
Random Forest(Green line) had the highest Area under curve. 84.52, while Decision Trees had the lowest.

**Specificity ( To identify customer churn)** 

Bagging Model Has the highest specificity, 49.16%. As seen before, it’s accuracy is high. Therefore, the bagging model can be good in identifying customer churn and retention, but there is the issue of missing out on 50% of customers who are likely to churn. To address this issue, we use the XGBoost Model with a balanced dataset.

**Considering Imbalanced Dataset ( To better identify customers)**

If the dataset is considered imbalanced, then the churn variable can be equalized to the dominant variable ( i.e. customers who will stay). By doing this, imbalances in the dataset are considered and it can lead to a decrease in False positive rate



From the above result of a balanced XGBoost Model, we can notice that the specificity has drastically increased to 75.10%, as compared to the previous models highest i.e. 49.16%. Therefore, this model, although low in classification accuracy, can be utilized to better predict a customer who is likely to churn.

**Summary**

Therefore, from this analysis we can learn that balancing the dataset can lead to a better result for customer churn, but at the cost of trading off accuracy. In other words, if we take the balanced dataset we will be more likely to identify the customers who will churn. If not, we attain a better classification accuracy but are not able to identify customer churn as efficiently.

All models perform better than the baseline models for predicting customers who will churn and those who will not. To further improve the classification result, a larger dataset can be useful.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |