**FINAL PROJECT REPORT**

**INTRODUCTION**

We’ve selected ‘Churn Modelling’ dataset of customers holding a credit card, where the target variables are ‘Exited’ and ‘hasCreditCard’. And implemented descriptive statistics and hypothesis testing to determine the minimum salary required for the credit card. To predict our target variables, we’ve implemented classification methods such as Linear and Logistic Regression, SVM, Naïve Bayes, Decision Tree, RandomForest, Gradient Boosting and XGBoost. Also, we’ve implemented clustering methods such as KNN and K-Means along with Regularization Techniques to determine the variance and overfitting of our data.

**MEHTODS**

1. **Hypothesis Testing:** This helps to study data for Churn Modelling as well as determine the minimum criteria required.
2. **Linear and Logistic Regression:** To predict the target variable for the Churn Modelling
3. **Decision Tree and regularization techniques:** To find the variance and fitting of data, and to predict the target variable
4. **K-means and KNN clustering:** To create clusters for the customer’s data to determine the credit history.
5. **SVM, Naïve Bayes, Random Forest:** Classification methods used to determine target variables ‘Exited’ and ‘hasCreditCard’.
6. **Gradient Boosting and XGBoost:** It is used to overfit our training dataset quickly.

**ANALYSIS**

**PART 1: DESCRIPTIVE STATISTICS AND HYPOTHESIS**

The summary of the data is as shown below:

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

1. From the Scatterplots, the plots show minimal correlation between the attributes of data.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

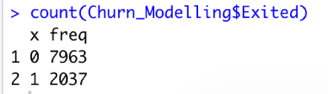
1. Notches in the boxplots do not coincide, we assume that their true medians differ.

A close up of a piece of paper

Description automatically generatedA screenshot of a cell phone

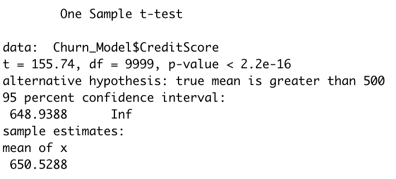
Description automatically generated

1. The Density plot below shows that the observed frequencies of customers who closed the account are 79632 and customers who retained their accounts are 2037.

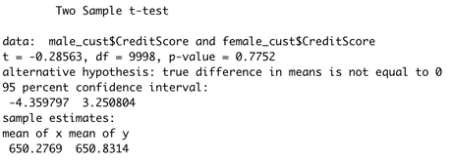
A screenshot of a cell phone

Description automatically generated

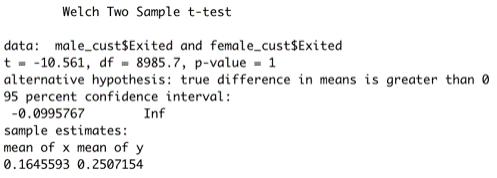
1. **HYPOTHESIS TESTING**:
2. **One Sample t-test:** From the below analysis, we know that the credit score is not greater than 500.



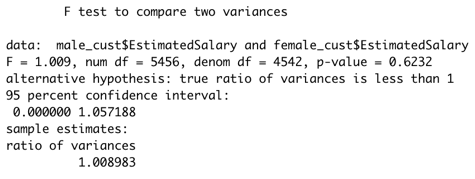
1. **Two Sample t-test:** From the below analysis, we observe that the credit scores are same for both the males and the females.



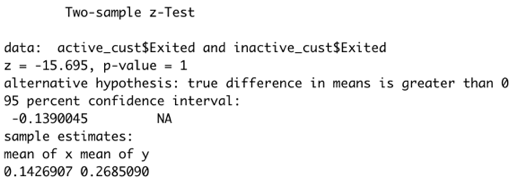
1. **Paired t-test:** After analysis we conclude that more males do retain banks accounts in comparison to the females.



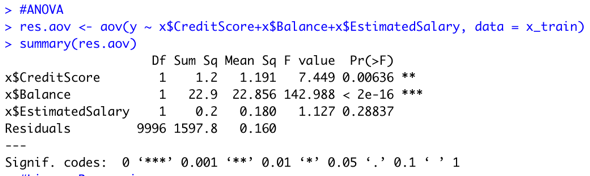
1. **F-test:** Hence from this analysis we infer that the variance of Estimated Salary in males and females is the same.



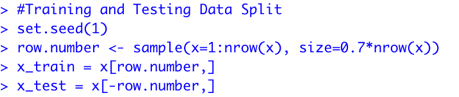
1. **Z-test:** From the below analysis, we observe that the mean of Exited customers more than the ones who chose to stay.



1. **ANOVA:** The p-value is less than the 0.05, we can conclude that there are significant differences between CreditScore and Balance.



1. **Linear Model:** In our linear model we consider three attributes; CreditScore, Balance and EstimatedSalary and from the analysis, we reject the Null Hypothesis and conclude that there is no relationship between the three variables. The accuracy for the Linear Model is 78%.

A screenshot of a cell phone

Description automatically generated

1. **Logistics Regression:** We predicted whether the customer continues with the credit card services or not. The model achieves 78.93 accuracy.

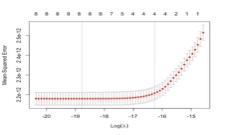
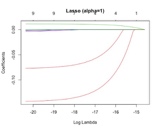
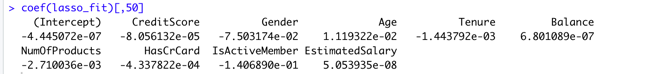
A close up of text on a white background

Description automatically generatedA screenshot of a cell phone

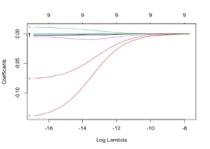
Description automatically generated

**PART 2: REGULARIZATION TECHNIQUES**

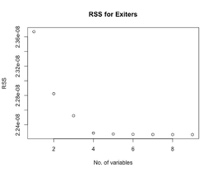
1. **LASSO:** All the variables in the dataset are significant for the prediction model.



1. **RIDGE:** Here the curves follow a similar trend as the LASSO model.



1. **BIGLASSO:** We also used the BigLASSO model also called as the extended LASSO.



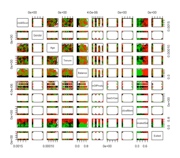
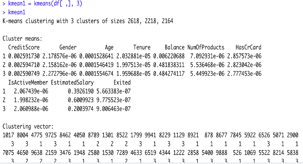
**PART 3: CLUSTERING AND CLASSIFICATION**

**A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated**

1. **K-Means:** With the Euclidian distance method we generated 3 clusters (This number of clusters was derived using the optimal cluster method i.e. nbclust)



A close up of a map

Description automatically generated

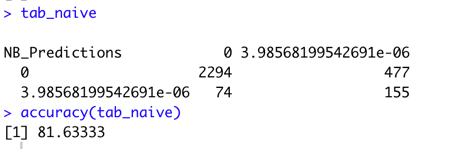
1. **K-Nearest Neighbors clustering:** For this algorithm 13 clusters are created.

A screenshot of a cell phone

Description automatically generated

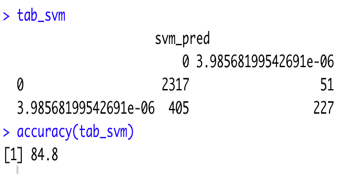
1. **Naïve Bayes:** The ‘Exited’ is compared with the probability of other dependent variables.

A screenshot of text

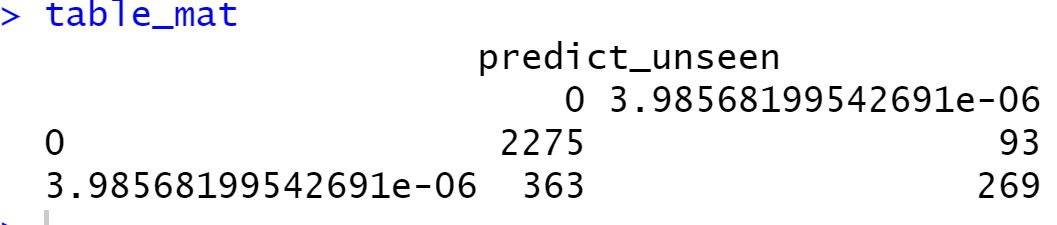
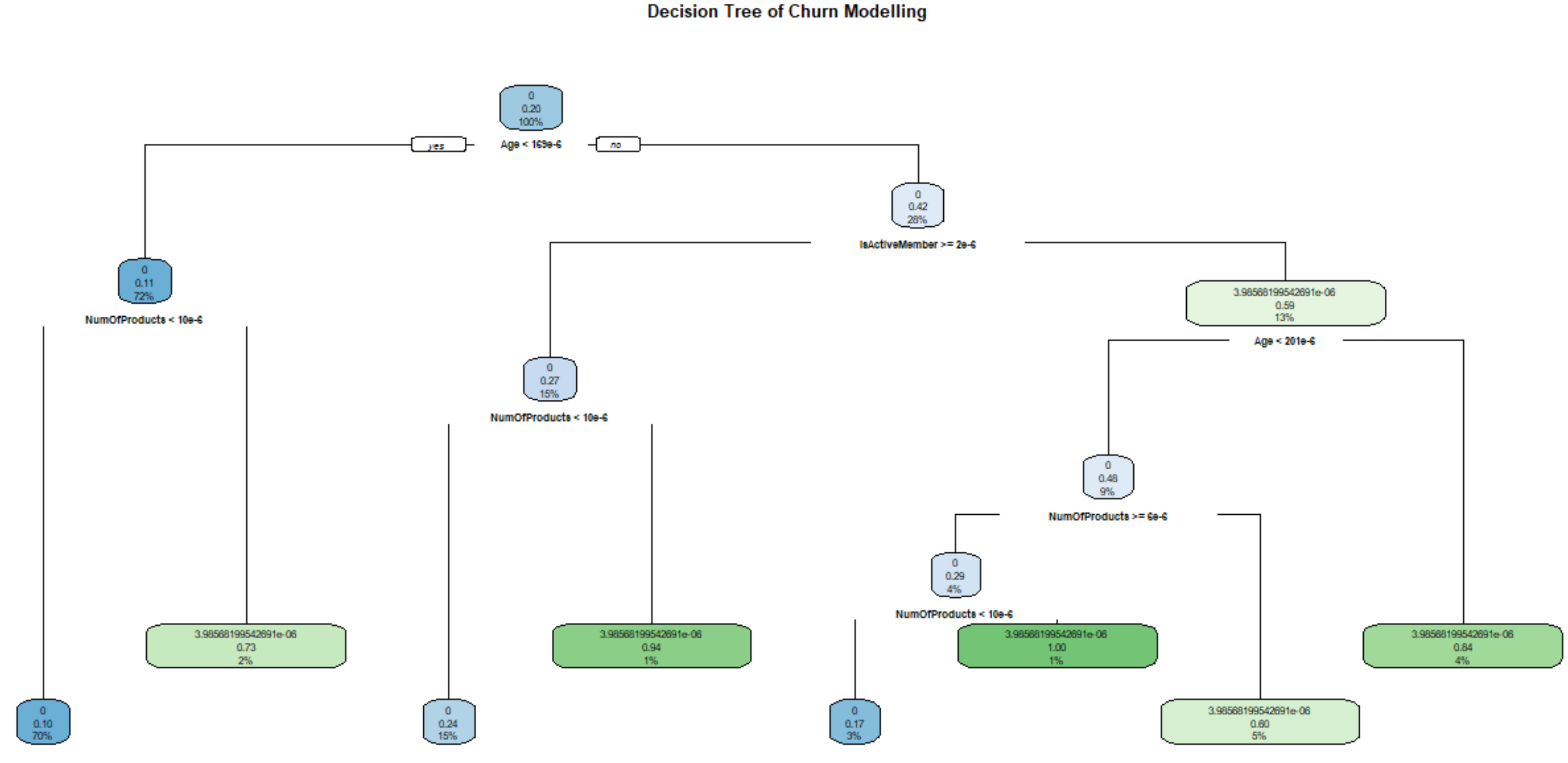
Description automatically generated

1. **Support Vector Machine:** The SVM model is analyzed below:

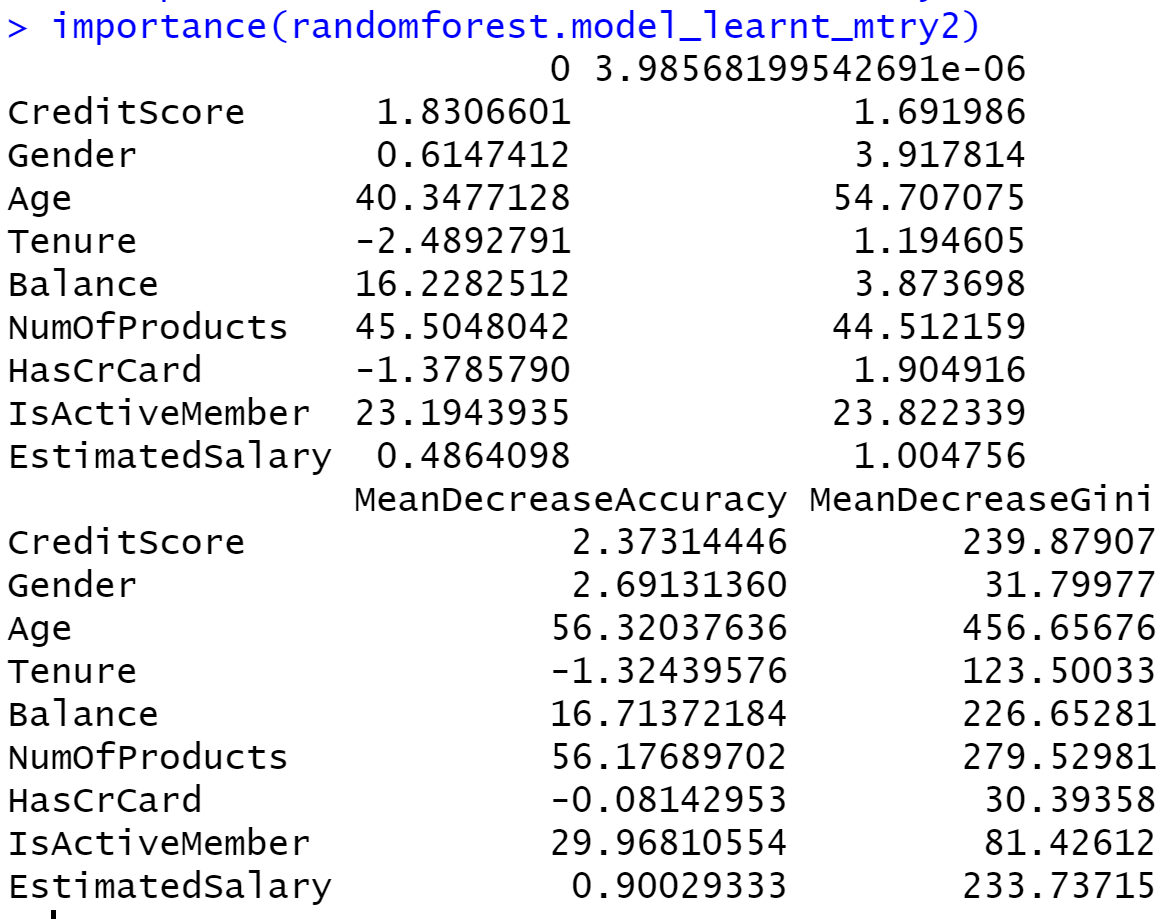
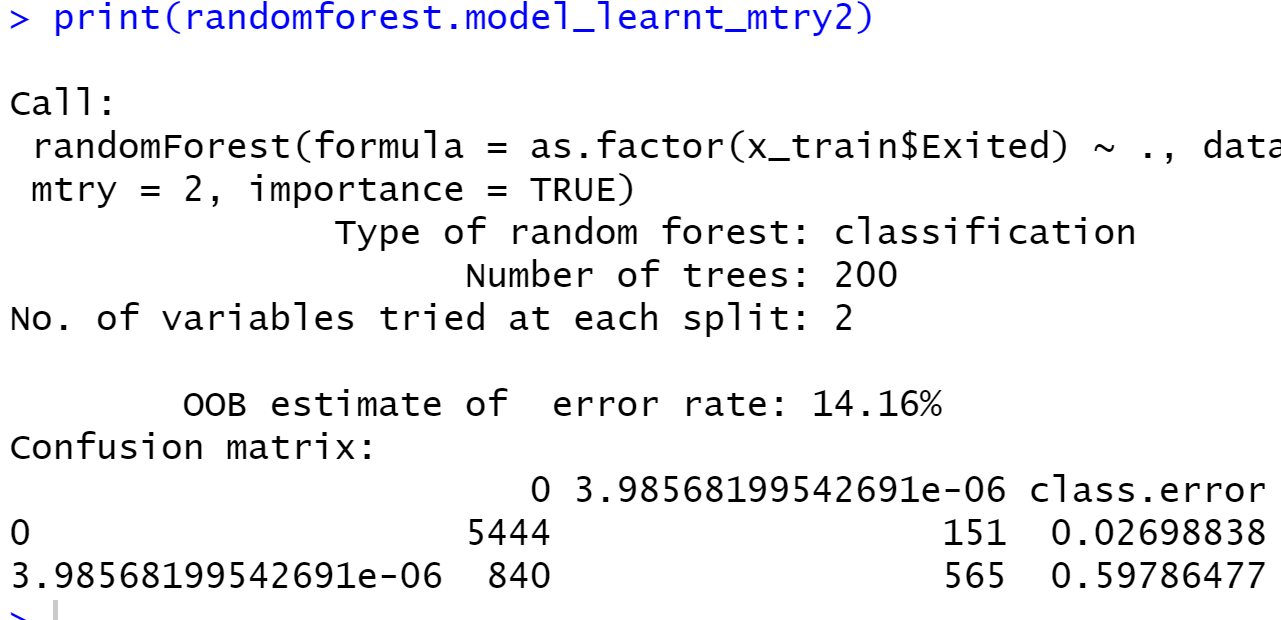
**A screenshot of a social media post

Description automatically generated**

1. **Decision Tree:** Decision tree and confusion matrix is as below. The accuracy is 84.8%



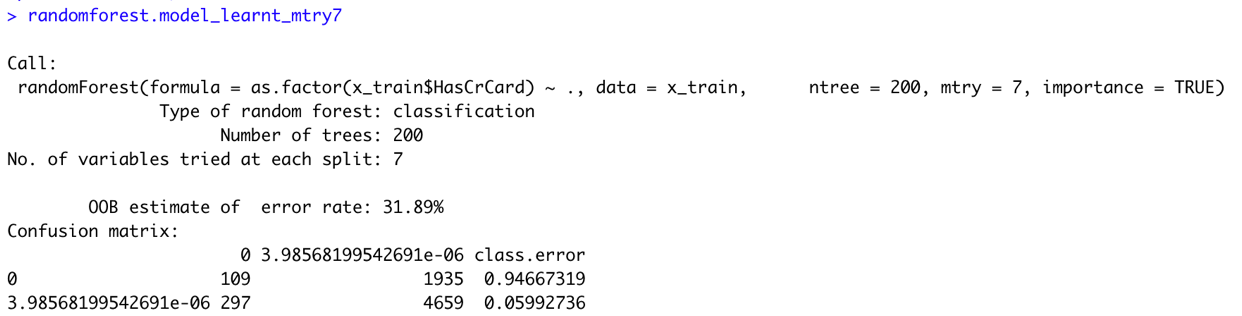
1. **Random Forest Algorithm:** The number of variables at each split is 3 and the accuracy is 85.05%. The important variables are shown below.



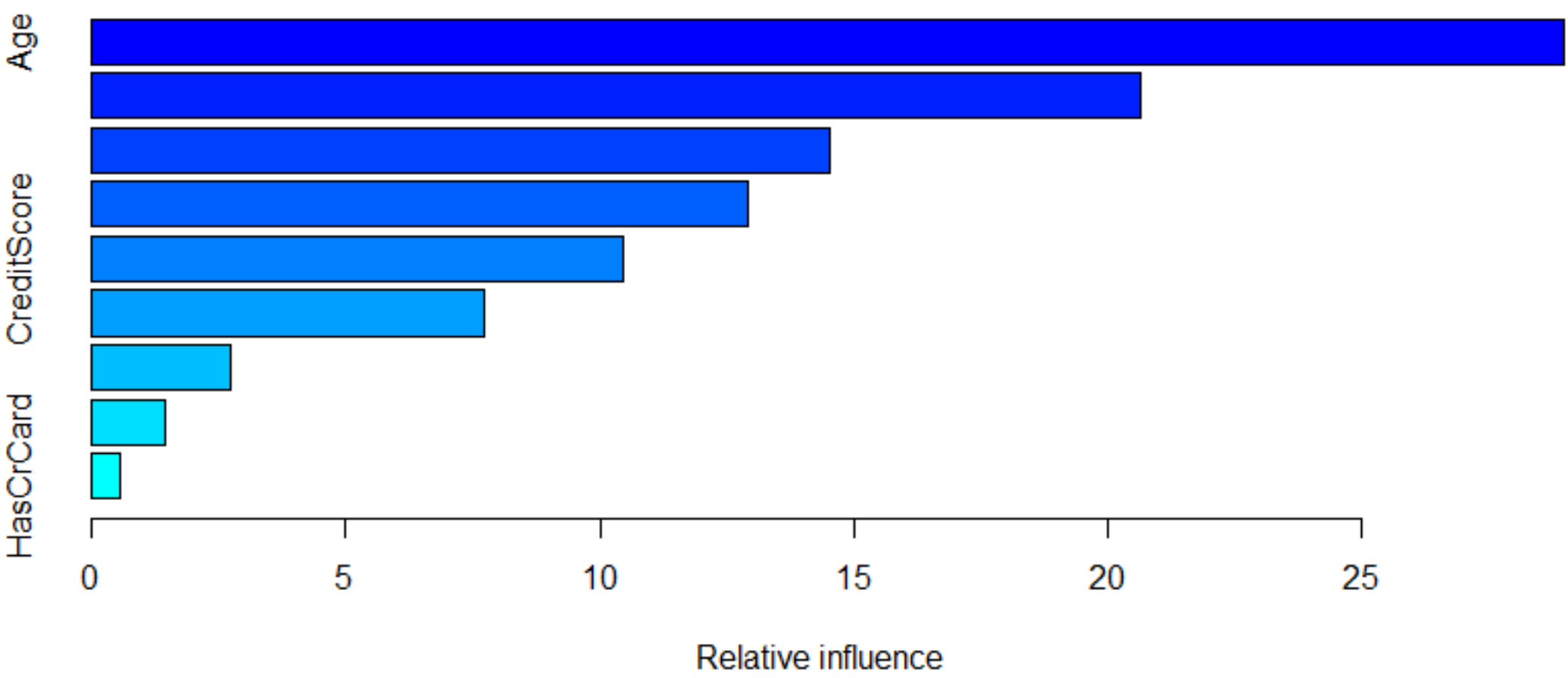
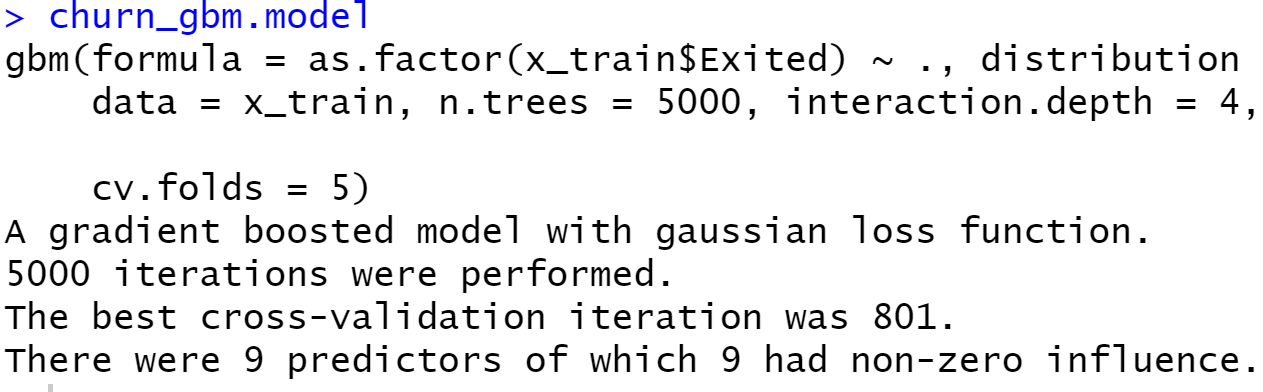
The accuracy for the target variable ‘hasCrCard’ is 70% using this method, where the predictions fitted the data pretty well.

A close up of a map

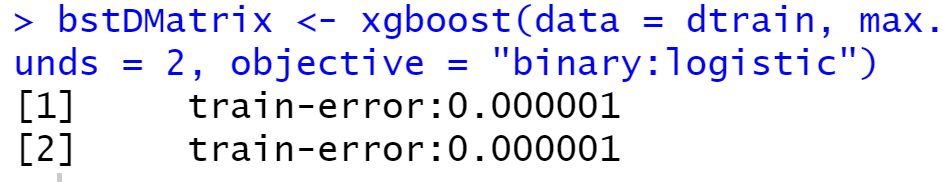
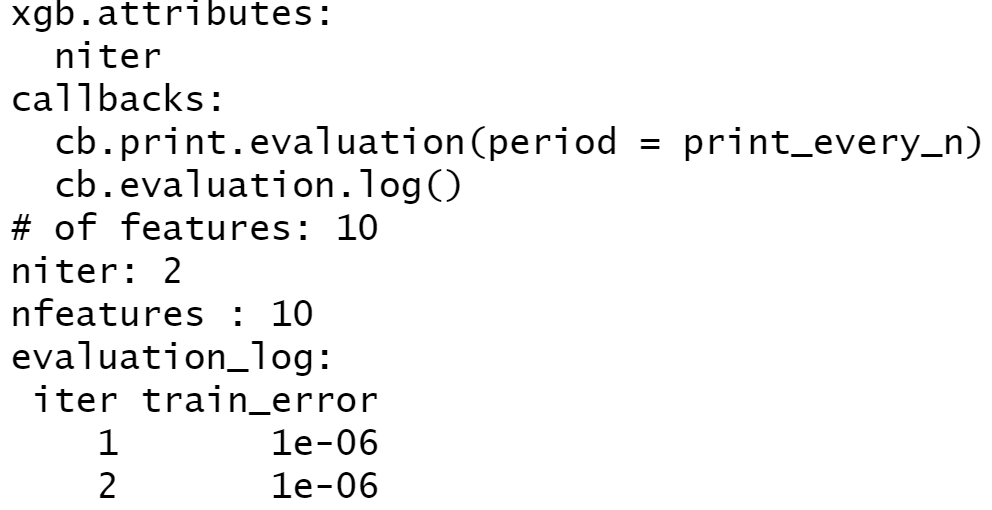
Description automatically generatedA screenshot of a cell phone

Description automatically generated

1. **Gradient Boosting Machine Algorithm:** The accuracy is almost 40%.



1. **Extreme Gradient Boosting (XGBOOST) Algorithm:** The accuracy is 78.93%



**INTERPRETATIONS**

1. From the customer’s Balance, Estimated Salary and Credit Score we determine that the customer’s possibility of exiting the bank service or not. Using the classification techniques, we were able to achieve approximately 85% of accuracy for the target variables.
2. To find the minimum credit score required for the customer to maintain his/her account is determined by the Hypothesis Testing and Descriptive Statistics. We found that 500 credit and more was the minimum requirement for the customer.
3. Using the customer’s Balance, Geographical Location and Estimated Salary we determine whether the customer holds a credit card or not, for this we’ve used classification as well as clustering techniques to determine its possibility. We achieved approximately 70% of accuracy to determine the target variables.

|  |  |
| --- | --- |
| **Techniques** | **Accuracy** |
| 1. Linear Modelling | 79.48% |
| 2. Logistic Modelling | 79.67% |
| 3. K-Means | 78.19% |
| 4. KNN | 80% |
| 5. SVM | 85.83% |
| 6. Naïve Bayes | 82.7% |
| 7. Decision Tree | 85.8% |
| 8. Random Forest | 85.93% |
| 9. Gradient Boost | 65.6% |
| 10. XGBoost | 79.67% |

**CONCLUSION**

We conclude that our churn modelling can be classified and clustered using various different algorithms and clustering techniques. However, since the dataset is huge and to avoid the overfitting of data, Random Forest would be the best option to predict the target variables and determine the customer’s outcome, also it fitted the data with minimum error.