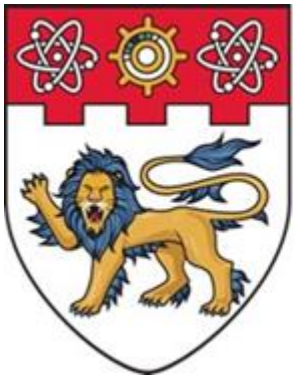


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 Analytics I: Visual and Predictive Techniques

Semester 1, AY 2020/21

TEAM PROJECT

Prepared by	Bairi Sahitya (U1922676B) Cai Xinrui (U1921389A) Ernest Ang Cheng Han (U1921310H) Malcolm Tan Yen Da (U1910206G)
Seminar Group	03
Team	07
Tutorial Instructor	Professor Hyeokkoo Eric Kwon
Date of Submission	1 November 2020

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Case Context

White Rock, a multinational Asset Management firm, is looking to utilise analytics to improve their business process to make faster and more informed decisions. This project aims to respond to their request for areas of focus and a Proof-Of-Concept (POC).

Business Problem & Opportunity

The area of focus for this project is that of customer retention, as it is one of the greater problems faced by companies today. This is especially so in the finance & asset management industry where there is strong competition which lead to intensive poaching efforts and eventual loss of customers. Moreover, it was discussed that traditional methods of data collection and analytics were simply too ineffective and unclear for White Rock to obtain substantial results.

As such, we defined the **business problem** as the low retention rate of customers of White Rock and lack of clarity in factors affecting it. With the trend and success of data analytics, the **business opportunity** was recognized as the potential for an automated analytics approach to speed up the process of identifying customers who are at risk of leaving and the factors which have the greatest impact on these customers.

Approach & Solution

Our proposed analytics solution seeks to develop a model capable of accurately predicting a customer's retention based on information about the customer as well as their interaction with White Rock. A dataset containing 9997 records was used and data cleaning was performed to provide high credibility and relevance for subsequent analysis. Data exploration was also conducted to gain a clearer understanding of the dataset.

The core design of our solution is built around using predicting models in R. We utilized various models namely Logistic Regression and Classification and Regression Tree (CART) and compared the models on the basis of accuracy and false negative rate. The CART model was evaluated to produce the best results, achieving an accuracy score and false negative rate of 97.0% and 9.97% respectively.

The model prediction complemented with expert opinion concluded that "AverageOfCustomerFeedbackOnService" and "PersonalAdvisor" were the more significant variables that can affect customer retention. The implementation plan is to provide White Rock with a simple dashboard displaying each customer's retention rate. Suggestions on how to employ Internal & External factors separately were given too. We further explored the possibility of creating a more holistic solution by suggesting White Rock to further utilise the identified factors to create personalised strategies for each customer.

1 BUSINESS PROBLEM

1.1 Customer Retention In The Finance Industry

Customer retention is one of the key success metrics of any organisation, including White Rock.

Customer retention measures not only how successful they are at acquiring new customers, but how successful they are at satisfying existing customers. These customers are contributing to a company's revenue and thus are of importance. According to Mixpanel's 2017 Product Benchmarks report, the finance industry faces a poor retention rate of 25% (Bernazzani, S., 2020), and this is likely due to the variety of supply of asset management firms available in the market, which leads to intensive poaching efforts from competitors.

Consequently, identifying the factors leading to low retention rate may be difficult.

Customer retention by its very nature is individualized, and will vary by many other factors such as the type of services or products provided. The traditional way of collecting data for analysis is through customer satisfaction surveys. (Qualtrics, n.d.), However, there is no objective way to verify the data collected. Also, the time lag between survey generating process, surveying process, and finally coming out with an analysis is significant. Moreover, too frequent satisfaction surveying may also lead to customer burnout (Survey Methods, 2018).

1.2 The benefits of Good Customer Retention

Our focus on customer retention stems from the potential benefits White Rock could reap from solving the problem identified above. Research has shown that high customer retention brings about greater profits. An increase in customer retention rate by 5% can result in an increase of 25 to 95% profit margins (Bain and Company, 2001). This is done through revenue and cost.

1. Higher Revenue

a. A retained customer is 60-70% more likely to purchase, as compared to a new customer which falls to around 5-20% (Saleh K., 2017). As such, it can be said that a retained customer spends more than a new customer.

b. Better brand image due to a strong hold of the consumer market which increases confidence of other customers to join in. After a Millennial finds a brand that they like, 80% of them will continue to give a company their business due to the familiarity. (Neff A., 2018)

2. Lower Cost

a. It is 5 times more expensive to acquire new customers than to retain them (Saleh K., 2017). This is due to the increased cost of marketing efforts. By spending more on new customers who may not generate as much revenue, the company may not even break even on those customers.

b. It provides free marketing as loyal customers often act as word-of-mouth marketers for the company. Moreover, word-of-mouth was cited as the best source for new business (Hubspot Research, 2018)

1.3 Rethinking the Problem: Identifying Factors of Retention

It is very hard for companies to accurately determine the causes of high customer churn rate. There are many variables surfaced in previous research that suggest certain factors are more important than others, especially those related to customer satisfaction and experiences (Yu, M. P., Grustani, H. G., & Intano, M. S., 2017). It is important for White Rock to be both able to predict accurately whether a customer is leaving and what White Rock can work on in their business processes to keep churn rate as low as possible.

In order to achieve this, for this project, we have divided our available information into two categories: internal and external. We define internal variables as those that White Rock has a certain extent of control over by altering business processes and external variables as those that White Rock has no control over. After running the final model, we hope to maximise the accuracy in predicting customer churn and what are the most influential factors behind this.

In the original dataset, we could only identify one internal factor named "IsActiveMember". In order to proceed with the analysis, we forged 8 more input variables which will be further discussed in Section 3.2 and Appendix A-1. These 8 forged variables are carefully selected because they are good measures in evaluating effectiveness of business processes.

Thus, if White Rock can identify the related factors accurately and even rank them according to their relative importance, they would be able to better plan their limited resources. Allowing them to target different groups of customers and improve retention rate using the most effective tools. Thus driving the most profits for the company.

2 ANALYTICS SOLUTION

2.1 The Opportunity: Automated Analytics Process To Support Retention

Data analytics is constantly evolving in this technologically driven world. Having shifted from descriptive to predictive and now prescriptive analytics, companies are now focusing more on actionable insights compared to simply monitoring data. For example, companies like Netflix and Amazon have witnessed success in using predictive data to draw meaningful insights to improve customer retention (Harris, A. B., 2018).

In order for White Rock to gain a competitive advantage, they have to speed up their process of identifying current relevant factors that contribute to changes in retention rate, and obtaining these relevant data quickly. Hence, we believe that there is a business opportunity for faster and more accurate retention predictions through an automated analytics approach.

For our project, we plan to address the business problem identified earlier and leverage upon the opportunity for automation by introducing a **Retention Analysis Model**. This will be a predictive model which **identifies customers of White Rock who are at high risk of leaving**. By unearthing possible churn before it happens, White Rock can then plan in advance the appropriate actions to take.

In addition, the model will also identify **factors that have the greatest impact on customer retention**. Our approach of looking into external and internal factors separately gives White Rock more information to work with and allows them to re-work their strategies accordingly.

2.2 Project Feasibility

While there is potential in adopting an analytical approach towards the business problem, it is critical to establish the feasibility of the project by ensuring that the problem has suitable characteristics for it to be tackled through the use of predictive models.

Firstly, the **predictive needs** of identifying important customers at risk of leaving and relevant factors that cause it will allow for White Rock to make timely interventions.

Subsequently, there exists the **imperfect knowledge** of White Rock not being fully aware of the intentions of their customers. There is a need for an analytics model that can recognise patterns from past customer data.

Lastly, the **training data is available** and will be addressed in the Section 3 of our report.

2.3 Desired Business Outcomes

Our focus on predicting customer retention and the factors affecting it bring about the following the business outcomes:

- **Identifying customers for Targeted Retention Efforts:** By developing a predictive model that accurately identifies customers who are at risk of leaving, targeted strategies can be executed towards them in an attempt to reduce actual loss.

- **Improving resource allocation and reducing cost**: By uncovering the more significant factors of customer retention, White Rock will be able to identify the important areas of focus and delegate more resources to build stronger strategies.

3 DATA CLEANING & EXPLORATION

3.1 Initial Dataset

The dataset we have chosen is the provided Bank Churn dataset available on Kaggle via the link: https://www.kaggle.com/shrutimechlearn/churn-modelling. The initial dataset consists of 10000 rows and 14 columns in a record data format. Each row represents a record and each column represents a variable. The description of these 14 initial variables is found in our Data Dictionary. The variable "CustomerID" is unique among all the 10000 customers and serves as the primary key attribute in this database.

3.2 Forged Variables

In order to demonstrate our models with the method described in our Analytics Solution section, we have decided to create 8 more variables for each of the observations in our dataset using R. Their description can be found in the Data Dictionary and the method used to generate them can be found in Appendix A-1.

3.3 Data Preparation and Data Cleaning

Data Cleaning is an essential step to prepare our data for further use in our models. Our data was cleaned in the following ways:

A. Outliers in the variable CreditScore

When we performed a box-plot on the continuous variable "CreditScore", we identified 2 outliers with the value 68 and 81, as shown in the diagram (Fig 1) below.

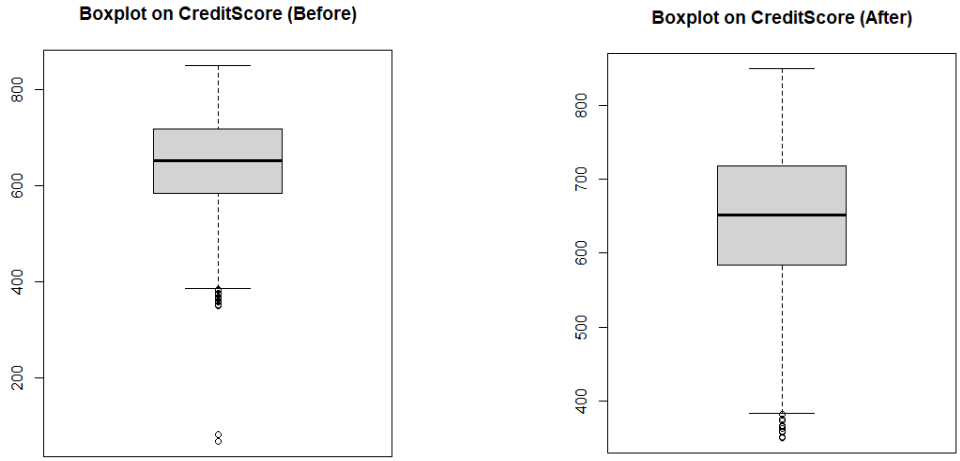


Fig 1: Effect of changing two outliers in the variable "CreditScore"

Considering that the two values are more than 4 IQR (Interquartile Range) away from the median value, we have considered two ways to handle these two outliers. The first way is to delete the two rows from our data and the second way is to treat it as a wrong value due to data entry errors and to correct them accordingly.

Since deleting rows should always be our last resort and our Logistic Regression model cannot handle NA values, we have decided the latter approach. We have made the assumption that these two outliers were caused by data entry errors, namly when the data was entered, the last digit of the number was omitted.

To correct these two wrong values, we have decided to multiply these two values by 10 and obtained 680 and 810 respectively, both falling inside the median +/- 1.5(IQR) range. After performing the multiplication, we have arrived at the new box plot as shown in the diagram above.

B. Redundant Data ("HasCrCard")

During the stage of modelling, we have identified the variable "HasCrCard" (whether a customer has a credit card with the bank) as a redundant variable. This is because it is irrelevant to the business problem we are trying to solve for WhiteRock. WhiteRock is not a bank so it will not be issuing credit cards to its customers. Therefore we have deleted the column entirely from our dataset.

C. NA in variables FinancialLiteracy and LastContactByABanker

There were 1 NA and 2 NAs in variable FinancialLiteracy and LastContactByABanker respectively, we removed these three rows from our dataset because the proportion is relatively small while there did not exist obvious solutions to appropriately replace those NA values. This has decreased our number of rows in the dataset to 9997 from 10000.

3.4 Data Exploration

We applied data visualisation techniques to derive insights from the data. The data visualisation packages used include the basic R package and ggplot2.

A. Distribution of Exited

The diagram (Fig 2) below has shown the frequency table for the variable "Exited". 20.4% (3 s.f.) of the customers have left, which is a relatively small proportion. In our Logistic Regression model, we have stratified the output variable during train-test split.

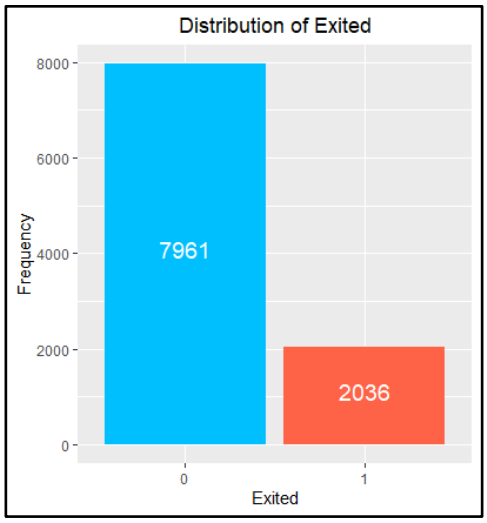


Fig 2: Frequency barplot of output variable

B. Distribution of Exited with respect all the variables

Data exploration should be purpose driven, our focus is to analyse the reasons and factors behind for why our customers had chosen to leave/stay. Therefore, pairwise distribution plots must be done first to understand more about the correlation between a variable and the output variable.

Appendix A-3 has shown the distribution for the output variable Exited with respect to all the variables in our dataset, except for "CustomerID" and "Surname", which are the key attributes and irrelevant to the analysis.

If the variable is a continuous variable, we have displayed the distribution in both jittered scatter plots and box-plots. If the variable is a categorical variable, we have displayed the distribution in the form of stacked barcharts (both frequency and proportion).

C. Findings from variables in the original dataset

Out of the 8 input variables from the original dataset, 7 of them ("CreditScore", "Geography", "Gender", "Age", "Balance", "NumOfProducts", "EstimatedSalary") are grouped under external factors while "IsActiveMember" is grouped under internal factors.

Of the 7 internal factors, several interesting observations surfaced. Firstly, the German branch of the bank has a higher proportion of customers leaving (Fig 3). Secondly, the variable "EstimatedSalary" is uniformly distributed in the dataset from the value of 0 to 200,000, showing no significant difference (Fig 4) between staying customers and leaving customers.

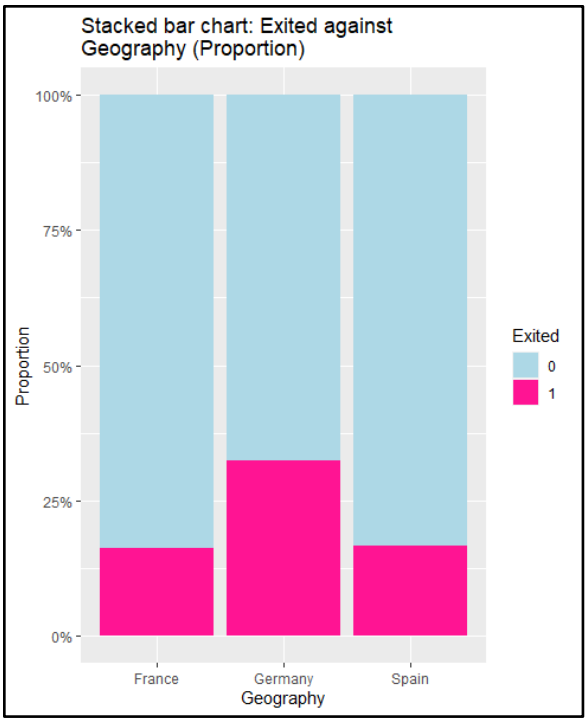


Fig 3: Stacked barplot (proportion) for variable "Geography"

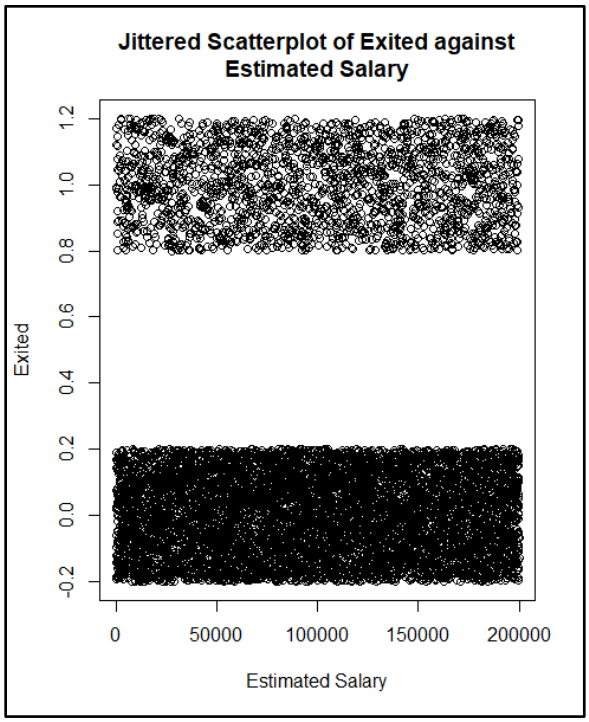


Fig 4: Jittered scatterplot for the variable "Estimated Salary"

For the only internal factor "IsActiveMember", we can observe clearly that active members have a lower proportion of customers leaving (Fig 5). There are 14.3% (3 s.f.) of customers who are active members and have left the company. Whereas there are 26.9% (3 s.f.) of customers who are inactive members and have left the company.

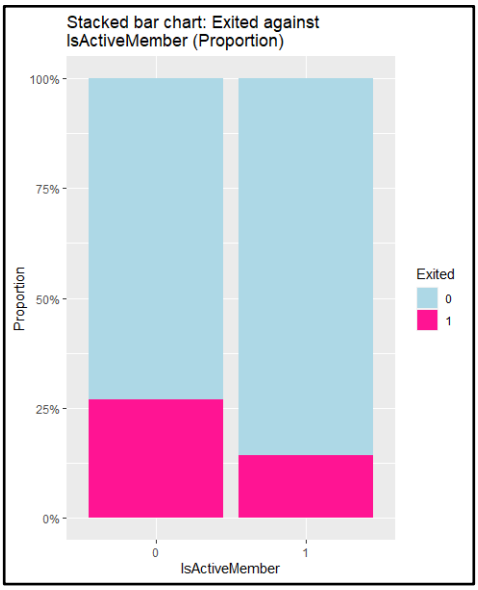


Fig 5: Proportion stacked barplot for variable "IsActiveMember"

4 LOGISTIC REGRESSION MODEL

4.1 Overview:

A logistic regression model was chosen as a contender for the predictive binary classification of customer retention. The final logistic model used consisted of only input variables that are statistically significant at 5 percent alpha and fitted using the glm function in R. The model fitting process, along with some other details on this model, is documented in Appendix D.

4.2 Considerations in choosing final model

When backward elimination was run on the model with all the variables, the variables chosen were the same as those manually selected based on their statistical significance. Furthermore, performing clustering on the dataset did not result in clear clustering among the dataset. While adding the clusters did improve the model accuracy, it was by a very small amount and was thus deemed insignificant. More details on the clustering process can be seen in appendix E.

The Adjusted GVIF values of the final logistic regression model are all below 2, suggesting that there are no multicollinearity issues between the explanatory variables.

4.3 Results:

The confusion matrix of the results using the test set is as follows:

	Retained (Negative)	Exit (Positive)
Retained (Negative)	2339 (78%)	49 (1.6%)
Exit (Positive)	84 (2.8%)	527 (17.6%)

Actual Results
Predicted Results
Accuracy: 95.6% (3 s.f.)
Type I Error (False Positive): 2.05% (3 s.f.)
Type II Error (False Negative): 13.8% (3 s.f.)

4.4 Insights:

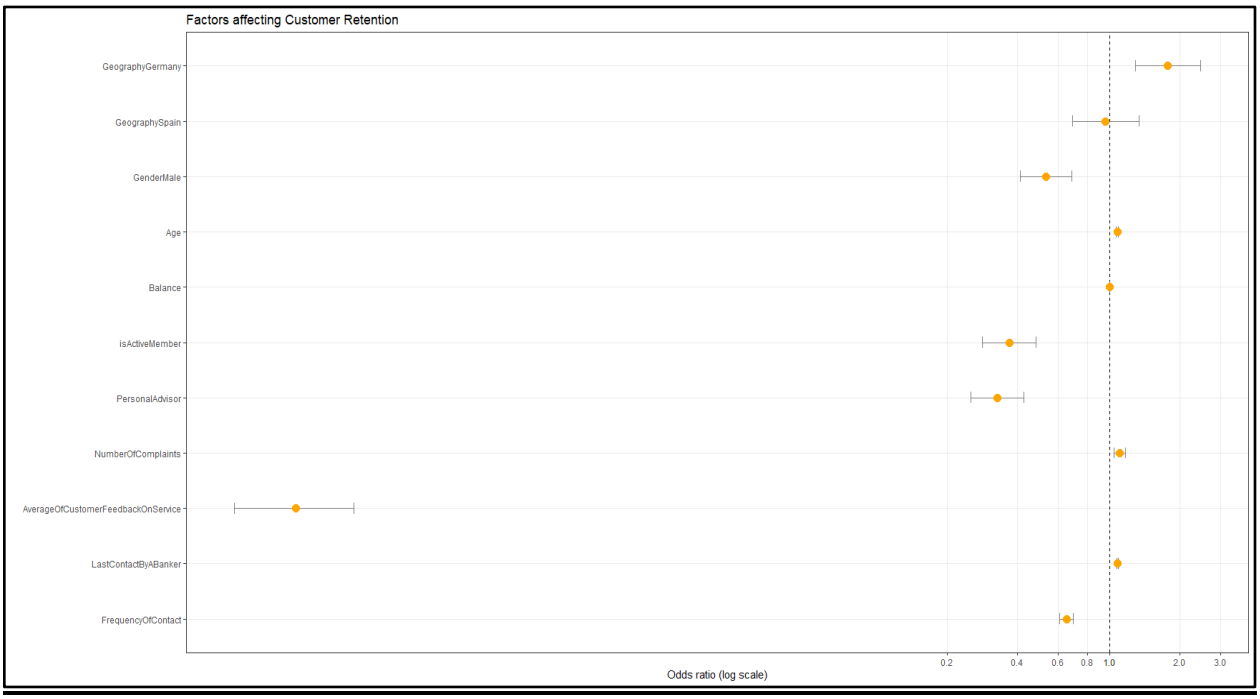


Fig 6 Odds Ratio of the predictive factors

The odds ratio, along with their confidence intervals are shown above. It is to be noted that an odds ratio of 0.5 is an equivalent departure from 1 as an odds ratio of 2. Thus a log scale is chosen for the x-axis to reflect this relative multiplicative effect.

The top 5 most significant factors based on the odds ratio seen above are:

1. AverageOfCustomerFeedbackOnService
2. PersonalAdvisor
3. IsActiveMember
4. GenderMale
5. GeographyGermany

While the first four factors have a negative effect on “Exited”, i.e when the variable increases or changes from baseline reference level, the odds of Exited = 1 decreases, “GeographyGermany” has a positive effect. This suggests the odds of a customer from Germany exiting is higher than those from France or Spain. As can be further seen from the confidence interval of “GeographySpain”, the customer being located in Spain seems to have no effect on their exit.

While the first three factors are rather intuitive, the effect of gender and geography may be further analysed to be understood better. For instance, there may be larger organisational issues in the Germany branch of the company.

5 CART

5.1 Overview:

CART (Classification and Regression Trees) utilizes both classification and regression in order to make accurate predictions on a given dataset. Implementation of this algorithm can be found in the RPART (Recursive Partitioning and Regression Trees) package. Generation of CART model, along with some other details is documented in Appendix C.

5.2 Results:

	Retained (Negative)	Exit (Positive)
Retained (Negative)	7865 (78.67%)	96 (0.96%)
Exit (Positive)	203 (2.03%)	1833 (18.34)

Actual Results
Predicted Results
Accuracy: 97.0% (3 s.f.)
Type I Error (False Positive): 1.21% (3 s.f.)
Type II Error (False Negative): 9.97% (3 s.f.)

5.3 Model Generation:

The first phase of CART involves growing the tree to its maximum by testing every possible value from all of the X variables and evaluating their respective gini indices to determine the best binary split factor and value(s) at every node of the decision tree (i.e. the smallest weighted average gini index). This process is repeated till the lenient stopping condition is met.

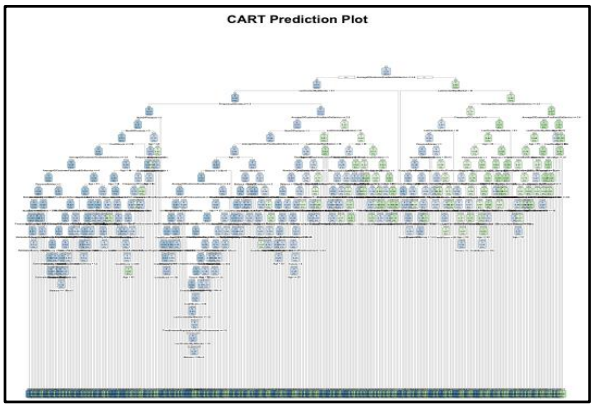


Fig 7 Full CART Diagram (Unpruned)

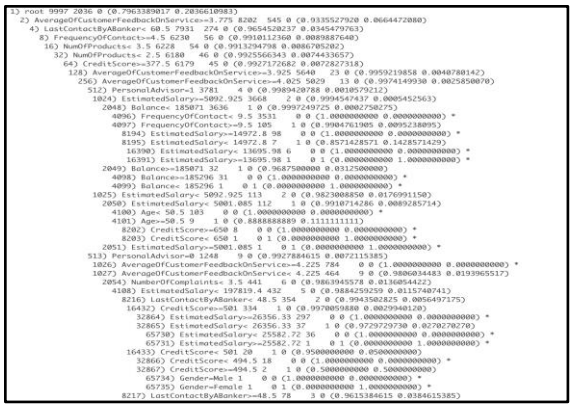


Fig 8 Segment of CART (Unpruned)

The second phase is to prune the CART to its minimum by pruning the CART at its weakest link using minimum cost complexity pruning (one standard deviation rule) in order to prevent overfitting the decision tree model.

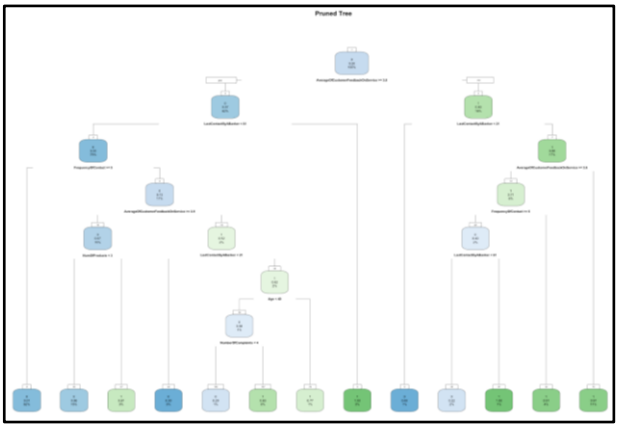
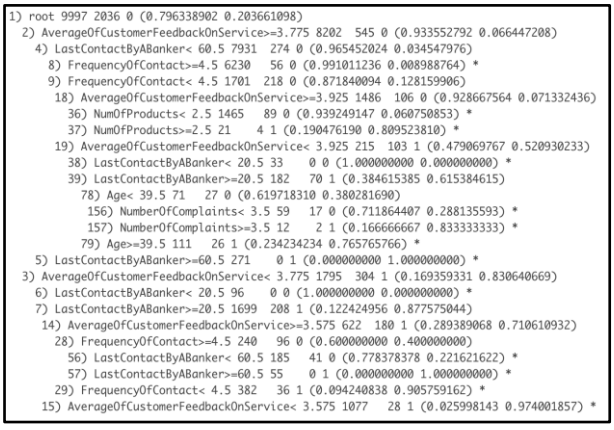


Fig 9 Full CART Diagram (Pruned)



```
1) root 9997 2036 0 (0.796338902 0.203661098)
  2) AverageOfCustomerFeedbackOnService>=3.775 8202  545 0 (0.933552792 0.066447208)
    4) LastContactByABanker< 60.5 7931  274 0 (0.965452024 0.034547976)
      8) FrequencyOfContact>=4.5 6230   56 0 (0.991011236 0.008988764) *
      9) FrequencyOfContact< 4.5 1701  218 0 (0.871840094 0.128159906)
       18) AverageOfCustomerFeedbackOnService>=3.925 1486  106 0 (0.928667564 0.071332436)
         36) NumOfProducts< 2.5 1465   89 0 (0.939249147 0.060750853) *
         37) NumOfProducts>=2.5 21    4 1 (0.190476190 0.809523810) *
       19) AverageOfCustomerFeedbackOnService< 3.925 215  103 1 (0.479069767 0.520930233)
         38) LastContactByABanker< 20.5 33    0 0 (1.000000000 0.000000000) *
         39) LastContactByABanker>=20.5 182   70 1 (0.384615385 0.615384615)
           78) Age< 39.5 71   27 0 (0.619718310 0.380281690)
            156) NumberOfComplaints< 3.5 59   17 0 (0.711864407 0.288135593) *
            157) NumberOfComplaints>=3.5 12    2 1 (0.166666667 0.833333333) *
           79) Age>=39.5 111   26 1 (0.234234234 0.765765766) *
    5) LastContactByABanker>=60.5 271    0 1 (0.000000000 1.000000000) *
  3) AverageOfCustomerFeedbackOnService< 3.775 1795  304 1 (0.169359331 0.830640669)
    6) LastContactByABanker< 20.5 96    0 0 (1.000000000 0.000000000) *
    7) LastContactByABanker>=20.5 1699  208 1 (0.122424956 0.877575044)
     14) AverageOfCustomerFeedbackOnService>=3.575 622  180 1 (0.289389068 0.710610932)
       28) FrequencyOfContact>=4.5 240   96 0 (0.600000000 0.400000000)
         56) LastContactByABanker< 60.5 185   41 0 (0.778378378 0.221621622) *
         57) LastContactByABanker>=60.5 55    0 1 (0.000000000 1.000000000) *
       29) FrequencyOfContact< 4.5 382   36 1 (0.094240838 0.905759162) *
     15) AverageOfCustomerFeedbackOnService< 3.575 1077   28 1 (0.025998143 0.974001857) *
```

Fig 10 Full results of CART (Pruned)

5.4 Insights:

Unlike the logistic regression model, CART uses a 10 fold cross validation algorithm and is hence able to filter the best variable to use at every of its nodes, along with the best binary split value for that particular variable in order to give the most accurate prediction in determining whether a customer would churn or continue using the organisation's services. Therefore, it is not necessary for us to categorize the variables in order to find the more accurate external and internal predictors for our solution as we did for the logistic regression model.

We will also examine the stopping rules used at the terminal nodes of the CART model since these classifications produce the purest child nodes. By doing so, we can narrow down the variables that could accurately determine if a particular customer would churn or retain in that financial year before performing further analysis on these variables to ensure its coherence and validity. Below are some of the stopping rules used in the terminal nodes:

For Customer Retained (Exited = 0)

1034) PersonalAdvisor=1 84 0 0 (1.0000000000 0.0000000000) *
968) Gender=Male 4 0 0 (1.0000000000 0.0000000000) *
4720) Geography=France,Germany 8 0 0 (1.0000000000 0.0000000000) *
268) IsActiveMember=1 3 0 0 (1.0000000000 0.0000000000) *
1026) AverageOfCustomerFeedbackOnService>=4.225 784 0 0 (1.0000000000 0.0000000000) *

For Customer Churned (Exited = 1)

65735) Gender=Female 1 0 1 (0.0000000000 1.0000000000) *
2551) Age>=41 15 0 1 (0.0000000000 1.0000000000) *
16390) EstimatedSalary< 13695.98 6 0 0 (1.0000000000 0.0000000000) *
1945) FinancialLiteracy=2 4 0 1 (0.0000000000 1.0000000000) *
931) CreditScore< 606 9 0 1 (0.0000000000 1.0000000000) *
2075) NumOfProducts< 1.5 4 0 1 (0.0000000000 1.0000000000) *

From examining the stopping rules which classify churned customers in the pure terminal nodes of the CART model, we can also infer that apart from the internal organisational factors which would obviously impact retention and/or churn rate, many external factors such as gender and age seem to play a significant role in determining cases of churned customers. This is expected given that such physiological factors do influence investment decisions as well as the likelihood of customers to stop using a company's products or service (Aren, S., & Aydemir, S. D., 2015).

It is important to mention that the respective variables and values can be utilized to advise and possibly necessitate follow up action from the internal management with the caveat that some of the results could be unintuitive, or even counterintuitive. The results generated should be further re-evaluated with industrial and professional knowledge from the relevant subject matter experts for a more accurate interpretation.

6 MODEL EVALUATION & RECOMMENDATIONS

6.1 Model Selection

As discussed in the Business Problem, the question that WhiteRock wants answers to is whether a customer is at risk of leaving (Exited == 1) and are more likely to be interested in a model that provides greater accuracy and prediction rates. Therefore, more attention should be given to the accuracy score and false negative rate of each model. The false negative rate is important as it tells us the percentage of customers that are predicted to be retained (negative) when in actuality, they

have exited (positive). In a technical aspect, this shows the overestimation of model predictive capabilities which may result in White Rock having lower retention rates, since they are not identifying the customers' intention accurately.

Comparing the two models, **CART** performed better in terms of accuracy and false negative rate. Thus, it is the final and better model chosen to predict retention rate of customers in the asset management industry.

Comparison of Logistic Regression Model and CART

Model	Test Set Accuracy	False Negative Rate
Logistic Regression	95.6%	13.8%
CART	97.0%	9.97%

While CART is superior in terms of accuracy, there are still some merits to the logistic regression model. The odd ratios that are derived from it are able to quantify the positive and negative effects of each input variable on the output variable, something that CART is unable to do. Thus, White Rock can still consider keeping the logistic regression model as a supplement.

The more noticeable and significant variables that were indicated in both model analysis include "AverageOfCustomerFeedbackOnService" and "PersonalAdvisor". This suggests that White Rock can expect higher retention from customers that are more satisfied and informed. Other significant variables identified also include "Age" and "Credit Score".

6.2 Complementing With Literature Review & Expert Opinion

Ideally, the selection of most significant variables should consider the opinions of relevant subject matter experts. To give a more holistic report, this section briefly discusses the major variables identified by professionals and relevant retention studies.

According to a study on customer retention, customers strongly agree that satisfaction of the products and services has influenced them in continuously patronizing. Similarly, customers agree that being comfortable and developing personal friendships with staff influenced them to retain (Yu, M. P., Grustani, H. G., & Intano, M. S., 2017). These are consistent with our result above that showcases "AverageOfCustomerFeedbackOnService" and "PersonalAdvisor" having high predictive power for customer retention.

However, we acknowledge the fact that our model was developed on one dataset and cannot entirely capture all factors that are significant in predicting retention. Thus, we researched into other factors that were not identified as strong variables in the result of our CART model.

Following a study by Gebze Institute of Technology, it was identified that financial literacy is one of the strongest factors of investment contributions (Aren, S., & Aydemir, S. D., 2015). With low financial literacy, they have lower risk appetite and are less likely to make transactions. Hence, it may be possible to suggest that low literacy can initiate the withdrawal process and cause customers to leave. However, the CART model suggests that high financial literacy is a stronger predictor of attrition. We believe that this is because customers with more financial knowledge tend to do more research on other asset management firms and are more likely to leave if they find another more attractive platform. Since this contradicts the results obtained above, more research is needed to ascertain a conclusive answer.

7 IMPLEMENTATION & RECOMMENDATION

7.1 Implementation of the Solution

An interactive dashboard can be created where managers of different departments within White Rock can provide information pertaining to a customer. The trained CART model can then be implemented on this database to give an accurate prediction of the retention status of a customer.

To give information of the variable importance on retention, a combination of the CART model and the logistic regression model should be used. Managers can use this information accordingly as explained in the next few sections.

It should be noted that the factors we have chosen are only a few of the variables we found to be significant based on our research. White Rock should constantly update the database and explore new factors that may contribute to changes in the probability of a customer leaving White Rock. The factors should continue to be holistic as discussed and should give a wide representation of White Rock's customer profile and their business processes. The model should be revised frequently and continually trained with new data to ensure it stays relevant and accurate.

7.2 Recommendations based on Significant Internal Factors

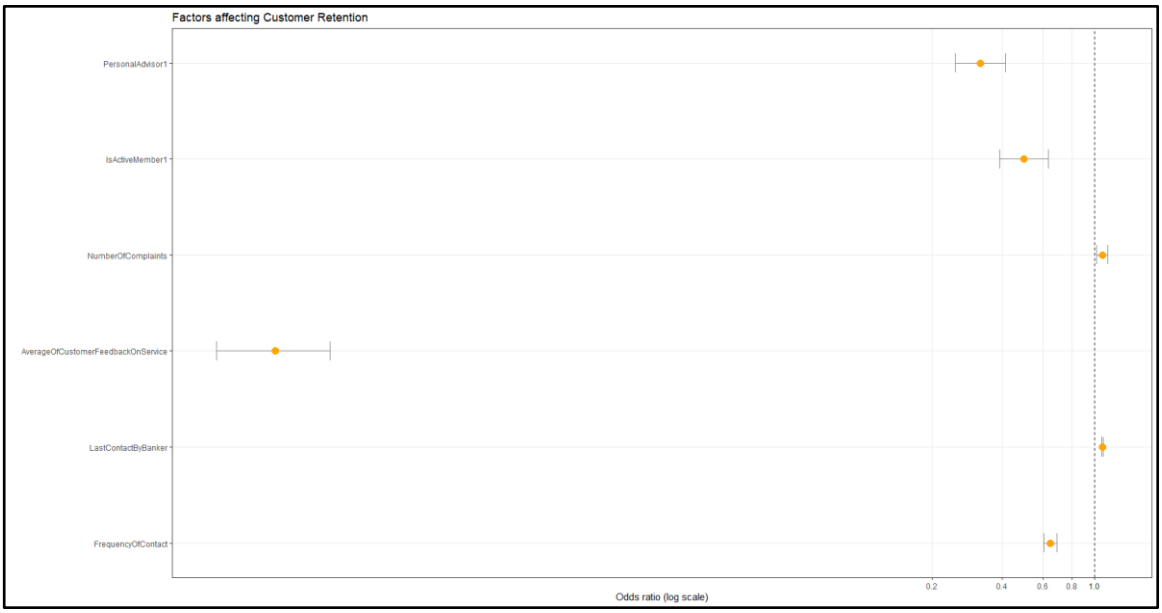


Fig 11 Odds Ratio of the predictive internal factors obtained when a logistic regression model is fitted with only internal variables

Analysing the significant internal factors can allow the company to identify the internal business processes they can focus on to improve customer retention.

Based on the results above, the internal variables with the highest impact are "AverageOfCustomerFeedbackOnService", "PersonalAdvisor" and "IsActiveMember", all having a negative effect. The large deviation seen in the odds ratio of "AverageOfCustomerFeedbackOnService" suggests that customer satisfaction plays a huge role in retaining the customers. Thus, the company can do further analysis to understand what drove customers to give high/ low feedback ratings and improve the respective aspects of customer service they are lacking in. Similarly, the company could benefit from investing in more personal advisors who can assist the customers in terms of investment and post-investment support. They could also market the benefits of the membership card more among their customers and target the benefits to their customers.

7.3 Recommendations based on Significant External Factors

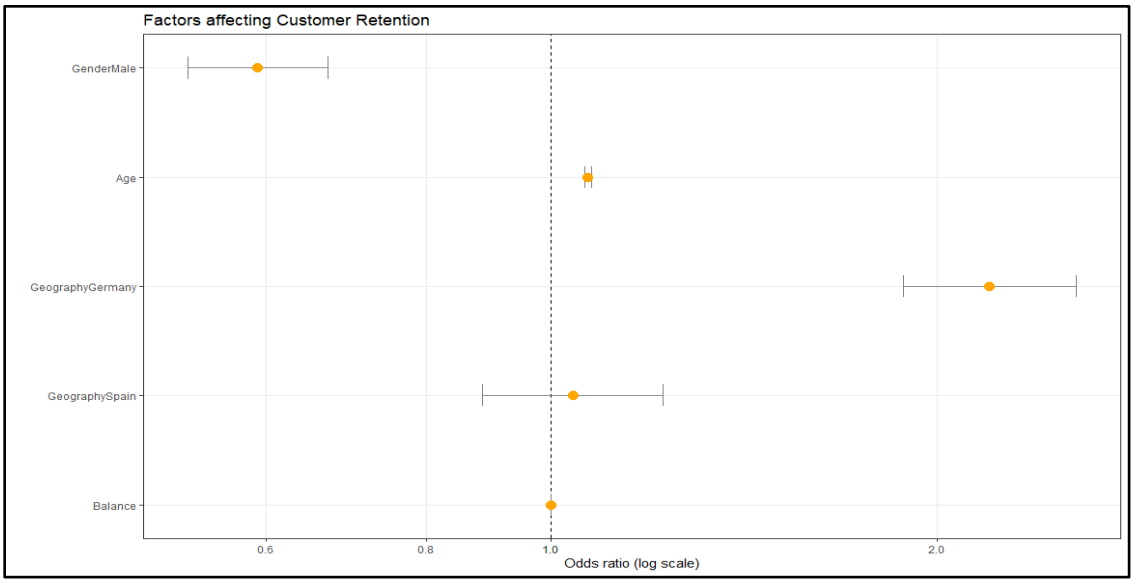


Fig 12 Odds Ratio of the predictive external factors obtained when a logistic regression model is fitted with only external variables

Analysing the significant external factors can be useful in two ways.

Firstly, it allows the company to understand their customer profile better. This information enables the company to do more targeted outreach to customers who are more likely to retain. As discussed in the previous section and as can be seen from figure 12, younger males may be a more favourable customer profile to target as they are more likely to retain.

Secondly, by doing further analysis on why certain customer proles are more likely to exit, the company can implement solutions that can address significant concerns of these profile groups. For instance, in a survey by WealthiHer Network on what women think about the meaning of wealth, 59% of the respondents replied with "Being able to provide for their family/children and about security and comfort" (Warwick-Ching, 2019). As such women may be more interested in topics such as caring for ageing parents. Managers within WhiteRock can adapt their programs to introduce such conversations with their female customers. WhiteRock should do further analysis within their own customer base to better understand their needs as well.

7.4 Comparing impact of Internal Factors against External

When the logistic regression models were fitted separately with only external or internal variables, the model with only external factors had a testset accuracy of 79.3% while the one with only internal factors had a testset accuracy of 95.0%. This suggests that internal factors are much better predictors of customer retention compared to external factors. Thus, should the company have organisational limitations, they will be better off focusing on improving internal business processes rather than increasing effort to market to a favourable customer profile.

8 PROJECT EXTENSION

8.1 Creating Personalised Variable Analysis

As discussed earlier, personalisation is the key to improving customer retention (Pierrot, 2019). One way this project can be expanded to create more personal solutions is to understand how the significant internal retention predictors vary with each of the external variables. For instance, are certain internal factors more important in predicting retention for females compared to males? Analysis like this can be done across all the external variables to come to a meaningful conclusion. Using this information for all the external factors, White Rock is able to identify more accurately the important business processes for a particular customer, based on their customer profile.

8.1 Variable Analysis of Younger vs Older Customers

To give an example, a separate logistic regression process is run on two groups, one with Age<39 and one with Age>=39. 39 is the mean age of the entire dataset. The implementation and results of these models can be seen in Appendix F. From the results, it was observed that "NumberOfComplaints" is a statistically significant positive factor for older customers but not for younger customers. Furthermore, "IsActiveMember" has a more significant odds ratio (more deviant from 1) for older customers at 0.3757178 for older customers and 0.6379621 for younger customers. As a result, White Rock may benefit from focusing more on membership card usage and number of complaints for older customers than younger customers.

9 LIMITATION & FUTURE DIRECTIONS

9.1 Limitations

During the stage of data exploration, we found out that the data obtained had inaccuracies which impeded its overall credibility and ultimately the predictions that our logistic regression and CART models made from it. One example which clearly illustrates this was the information collected on "EstimatedSalary". By visualising its behavior through a box plot and a scatter plot, we realise that the "EstimatedSalary" variable is uniformly distributed from the value of 0 to 200,000 which is extremely unlikely in reality.

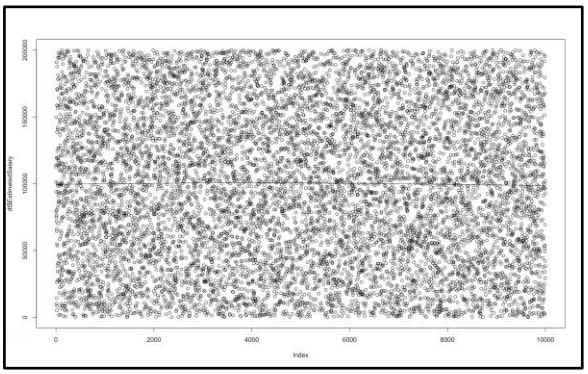


Fig 13 Scatter plot of "EstimatedSalary" variable

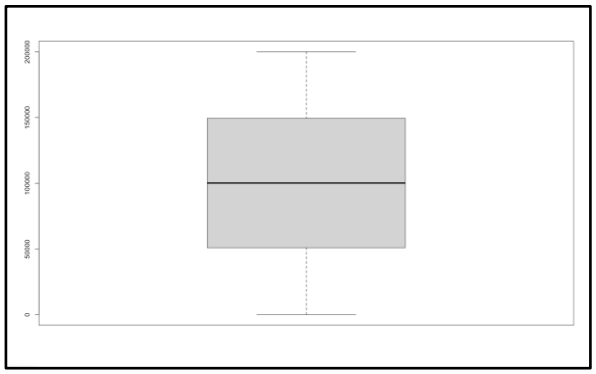


Fig 14 Box plot of "EstimatedSalary" variable

Additionally, information collected on customer churn was only from France, Germany and Spain with a significantly higher amount of data collected from France compared to Germany and Spain. Given the different demographics and geopolitical climate of individual nations which are factors that could impact retention and/or churn rate (Coval, 2001), limiting the geographical element of our data set to 3 countries with unequal representation diminishes the accuracy of the predictions made from our models.

Finally, it is essential that we recognise that our models are built upon the data provided by Kaggle. Generally, data collected on customer retention and/or churn for an organisation would vary from firm to firm even if they belonged to a common industry given that each company will collate different types of information and variables from its own pool of customers. Our models need to be adjusted and cleaned accordingly before implementation in order to ensure optimisation and accuracy in results produced and predictions made.

Even then, ultimately, our proposals made are built solely on statistical models and figures which could be misleading and provide false direction to organisations (Lebeid, 2018). Analysts and internal management should complement the predictions made with consultations from the relevant departments with the necessary domain knowledge as well as with literature and research written by experts from the same industrial field to further bring credibility to any final decisions made.

9.2 Future Research Directions

Upon identifying the previously mentioned pitfalls to our proposal, the majority of the flaws can be derived to the common issue of having a data set with a lack of depth in information collated. There are multiple areas for further research in this area of our project in order to further elevate the quality of our solution.

Effectively, the original data set only encompasses 11 different factors related to customer churn, out of which, few have been observed to be inaccurate as previously mentioned. Though both internal and external factors have been considered and collated, more emphasis can be placed in

collecting data directed towards more variables which could prove to be accurate predictors but were not sufficiently addressed in this dataset. These factors include capital appreciation, liquidity, financial security, expected returns, and risk minimization (Hemalatha, 2019).

Moreover, the data collected for categorical variables could be diversified to include a wider spectrum of categories. This is to minimize the impact of sampling bias, where data is collated from only specific groups of people (Riffenburgh, 2012), which would otherwise limit the generalizability of our model's predictions. In other words, findings from our solution can only be applicable to organisations whose collected data share similarities with our data set which is highly improbable as previously mentioned (Bhandari, 2020).

Additionally, the majority of the data collected are physical in nature. Though conventional and mainstream forms of data collected are useful to organisations and applicable for our model to make predictions, the data could also encompass audio and visual data as such novel forms of data collected forms the basis for emotion-based behavior analysis and research in communication (Ingram, 2017). This brings another dimension to our analysis on customer behavior and allows us to predict customer retention and/or churn rate based on the sentiments and what customers are feeling when they are using the organisation's services (Cunningham, 2020). This would otherwise be impossible with traditional forms of data collected.

Apart from revamping and revising our data set to make it more all encompassing, further research and development can be done on our predictive models as well. Apart from logistic regression and CART, White Rocks could invest in other prediction models and deep learning technologies such as Random Foresting and building Neural Networks which could potentially reduce prediction error as well as prediction speed (Lobato, 2017). Utilisation of such advanced artificial intelligence models will bring even more quality and credibility to the table and transform our tool into a multifaceted analytic tool with the potential to aid organisations in their decision making and allocation of resources to the right department so as to reduce their customer churn rate.

APPENDIX A - Data Cleaning and Preparation

Appendix A-1 (Data Preparation and the forging process)

1. Generating Personal Advisor:

```
randPA <- rep (0,nrow(data))

for (row in 1:nrow(data))

{

  if (data$Exited[row] == 0){

    randPA[row] <- rbinom(1, size =1, prob = 0.75)

  } else {
```

```
    randPA[row] <- rbinom(1, size =1, prob = 0.5)

  }

}

data$PersonalAdvisor <- randPA
```

2. Generating Financial Literacy:

```
randFL <- rep (0,nrow(data))

randFL <- sample(x = c(0,1,2), size = nrow(data),replace = TRUE, prob = c (0.2,0.4,0.4))

data$FinancialLiteracy <- randFL
```

3. Generating Number of Complains

```
randC <- rep (0,nrow(data))

for (row in 1:nrow(data))

{

  if (data$Exited[row] == 0){

    randC[row] <- sample(x=c(0,1,2,3,4,5,6,7,8,9,10), size = 1, replace = TRUE, prob =
c(0.61,0.10,0.10,0.09,0.03,0.02,0.01,0.01,0.01,0.01,0.01))

  } else {

    randC[row] <- sample(x=c(0,1,2,3,4,5,6,7,8,9,10), size = 1, replace = TRUE, prob =
c(0.48,0.12,0.12,0.11,0.05,0.04,0.03,0.02,0.01,0.01,0.01))

  }

}

data$NumberOfComplaints <- randC
```

4. Generating AverageOfCustomerFeedbackOnService

```
#The original code that resulted in Perfect Separation

for (row in 1:nrow(data))

{

  if (data$Exited[row] == 0){

    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 100,

    replace = TRUE, prob = c(0.01,0.06,0.08,0.20,0.65)))

  } else {

    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 100,

    replace = TRUE, prob = c(0.19,0.22,0.18,0.11,0.3)))
```

```
  }
}
#The code that eliminated the problem of Perfect Separation
for (row in 1:nrow(data))
{
  if (data$Exited[row] == 0){
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 20,
    replace = TRUE, prob = c(0.01,0.11,0.13,0.15,0.6)))
  } else {
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 20,
    replace = TRUE, prob = c(0.19,0.12,0.18,0.11,0.5)))
  }
}
data$AverageOfCustomerFeedbackOnService <- randCFS
```

5. Generating UnresolvedComplaint

```
for (row in 1:nrow(data)){
  if (data$NumberOfComplaints[row] == 0){
    randUC[row] <- 0
  }
  else {
    randUC[row] <- rbinom(1, size = 1, prob = 0.08)
  }
}
data$UnresolvedComplaint <- randUC
```

6. Generating LastContactByABanker

```
for (row in 1:10000)
{
  if (data$Exited[row] == 0){
    randLCB[row] <- sample(x= c(1:60), size = 1, prob = rep(1/60, 60))
  } else {
```

```
    randLCB[row] <- sample(x= c(21:100), size = 1, prob = rep(1/80, 80))

  }

}

data$LastContactByABanker <- randLCB
```

7. Generating TimeBetweenRegistrationAndFirstInvestment

```
randRTI <- sample(x=c(1:105), size = nrow(data), replace = TRUE, prob = rep(1/105, 105))

data$TimeBetweenRegistrationAndFirstInvestment <- randRTI
```

8. Generating FrequencyOfContact

```
randFOC <- rep(0, nrow(data))

for (row in 1:10000)

{

  if (data$Exited[row] == 0){

    randFOC[row] <- sample(x = c(1:10), size = 1, prob =
c(0.025,0.025,0.05,0.1,0.3,0.3,0.1,0.05,0.025,0.025))

  } else {

    randFOC[row] <- sample(x = c(1:10), size = 1, prob =
c(0.05,0.1,0.3,0.3,0.1,0.05,0.025,0.025,0.025,0.025))

  }

}

data$FrequencyOfContact <- randFOC
```

Appendix A-2 (Data Preparation and the problem of perfect separation)

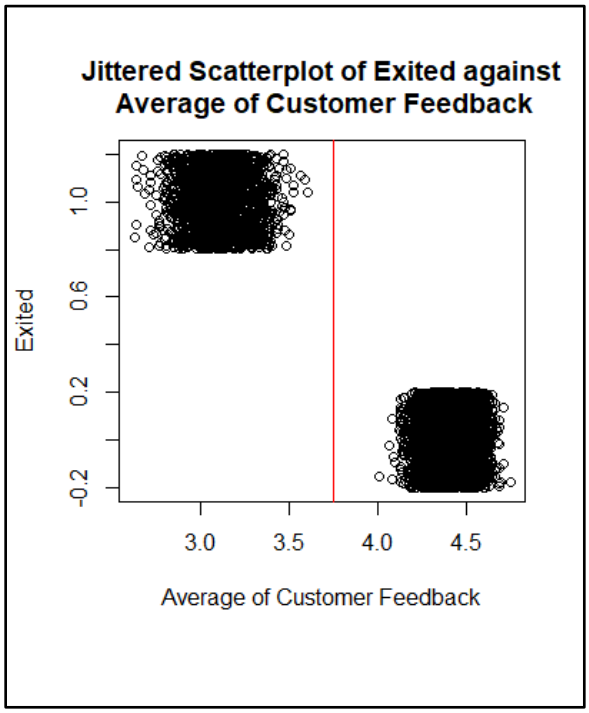
When generating variable "AverageOfCustomerFeedbackOnService", due to the inappropriate probabilities assigned to the two cases (Exited ==1 and Exited ==0), we have encountered the problem of perfect separation for this variable with respect to the output variable. Therefore, our CART model chose only AverageOfCustomerFeedbackOnService as the predicting variable while our Logistic Regression model could not converge. After realising this, we came back to the stage of data preparation and adjusted the probabilities accordingly, to ensure that we will eliminate the problem of perfect separation. We achieved this by decreasing the size of samples in sample() function from 100 to 20 while ensuring that the difference between the two sets of probabilities is not too wide. The diagram below demonstrates the problem of perfect separation and the effect after change.

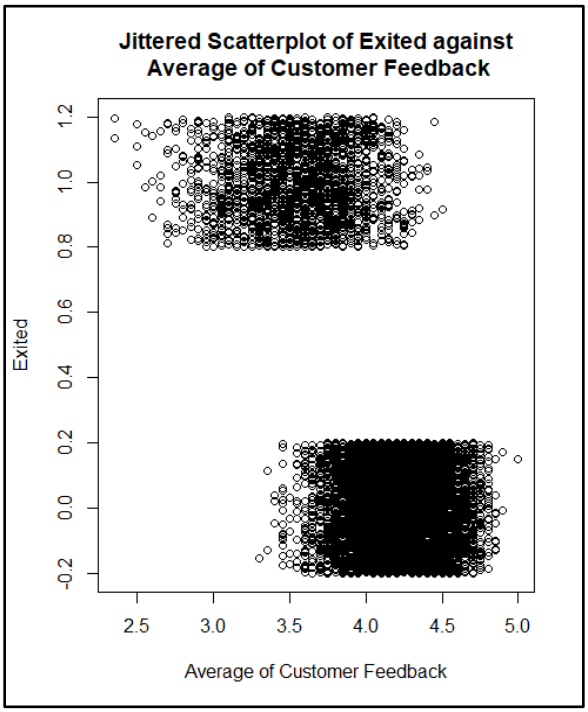
Before:

```r
#The original code that resulted in Perfect Separation
for (row in 1:nrow(data))
{
  if (data$Exited[row] == 0){
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 100,
    replace = TRUE, prob = c(0.01,0.06,0.08,0.20,0.65)))

  } else {
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 100,
    replace = TRUE, prob = c(0.19,0.22,0.18,0.11,0.3)))
  }
}
```

```r
#The code that eliminated the problem of Perfect Separation
for (row in 1:nrow(data))
{
  if (data$Exited[row] == 0){
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 20,
    replace = TRUE, prob = c(0.01,0.11,0.13,0.15,0.6)))

  } else {
    randCFS[row] <- mean(sample(x= c(1,2,3,4,5), size = 20,
    replace = TRUE, prob = c(0.19,0.12,0.18,0.11,0.5)))
  }
}
```

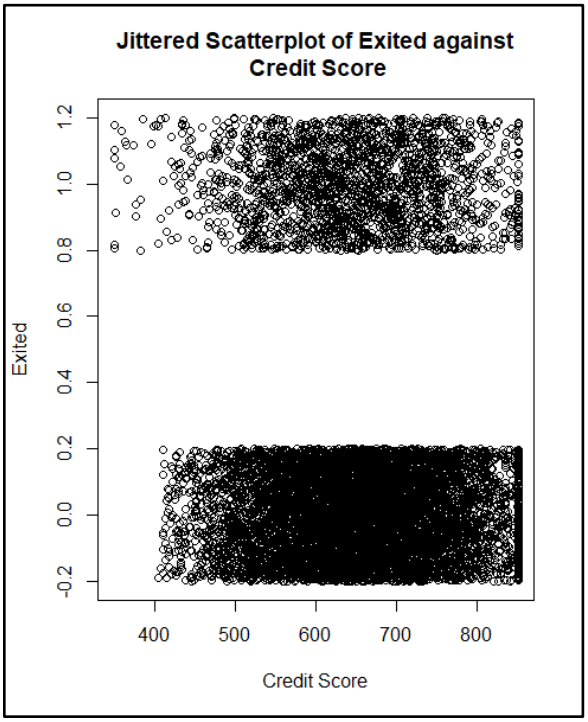
After:

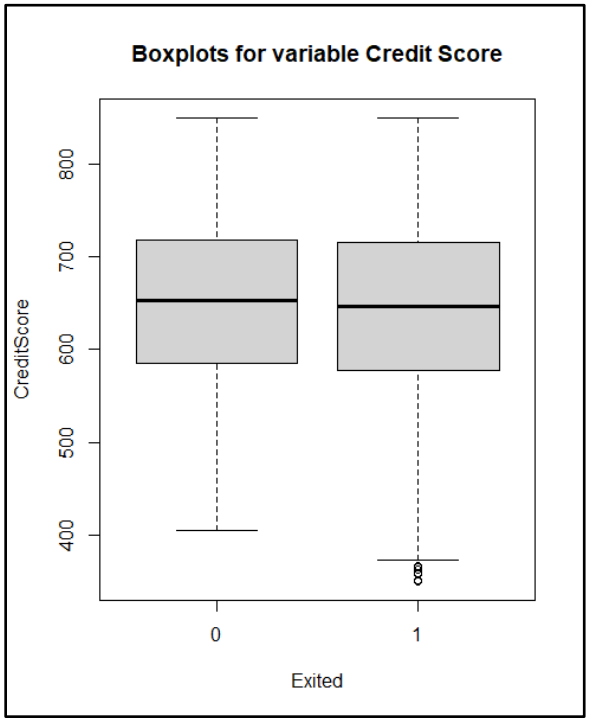




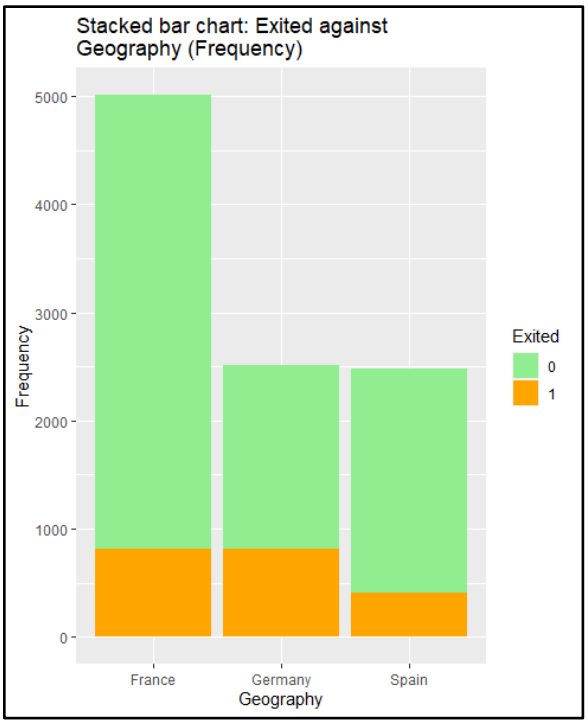
Appendix A-3

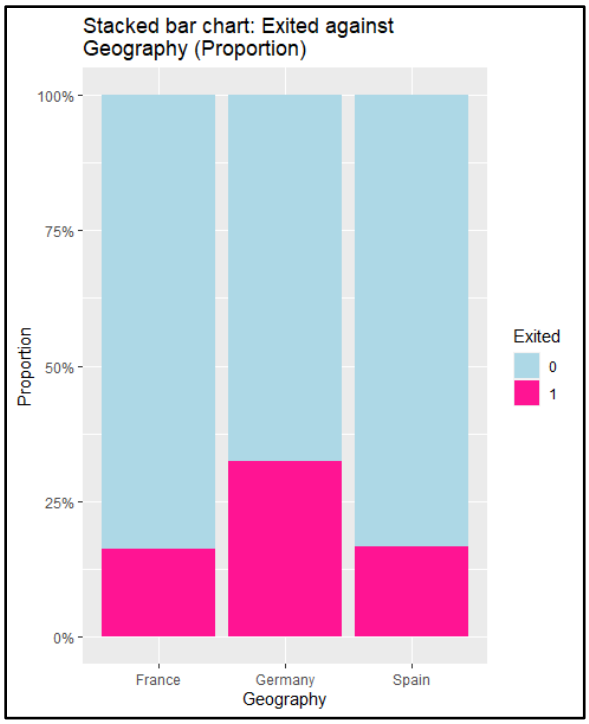
Variable: Credit Score (Continuous)



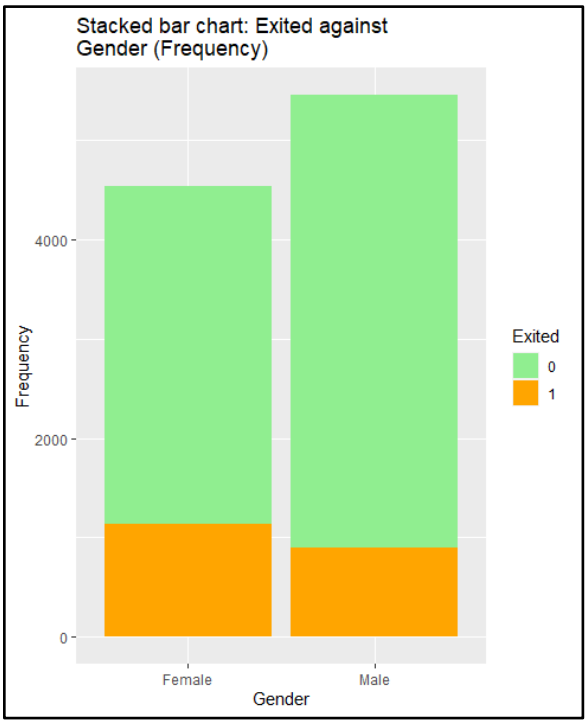


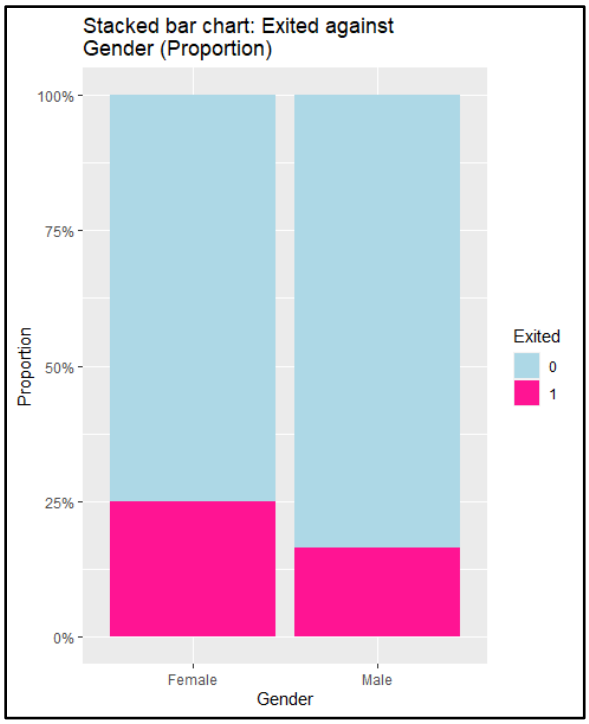
Variable: Geography (Categorical)



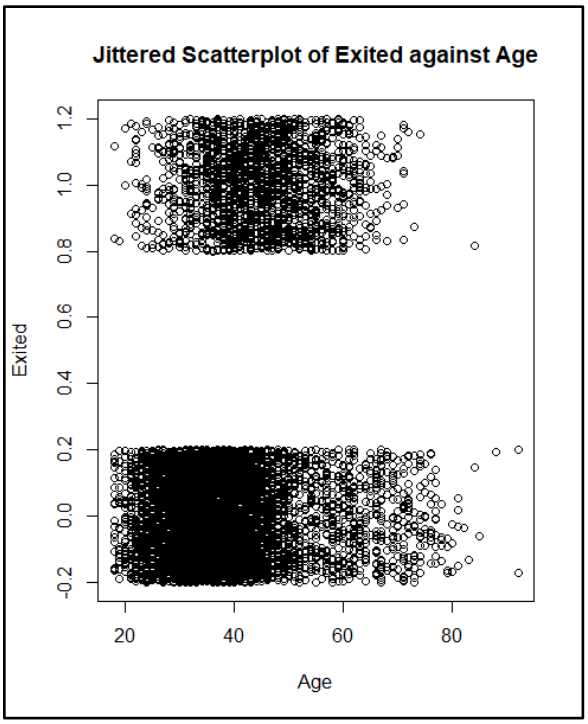


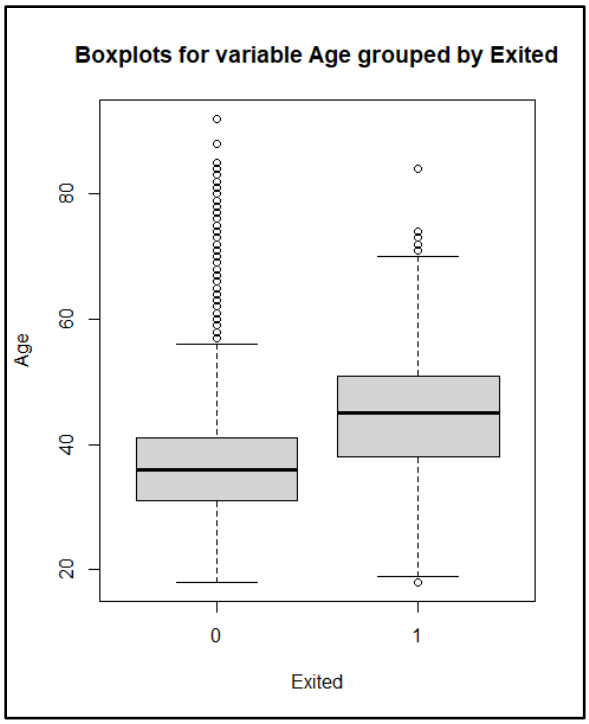
Variable: Gender (Categorical)



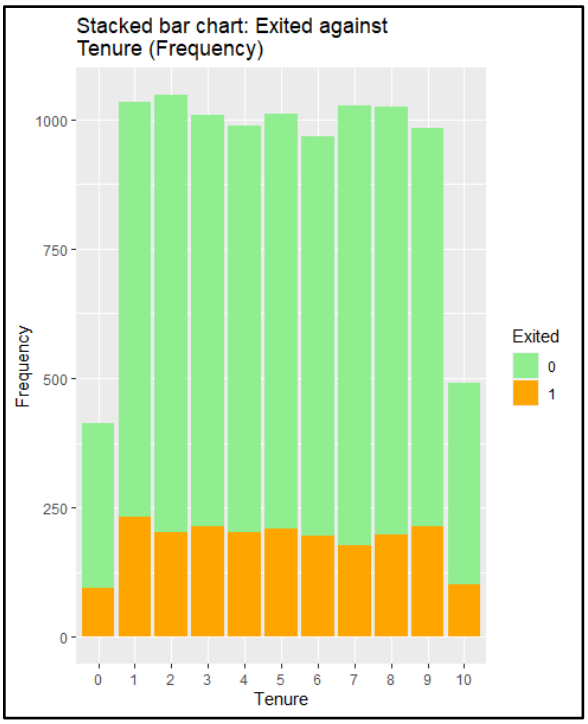


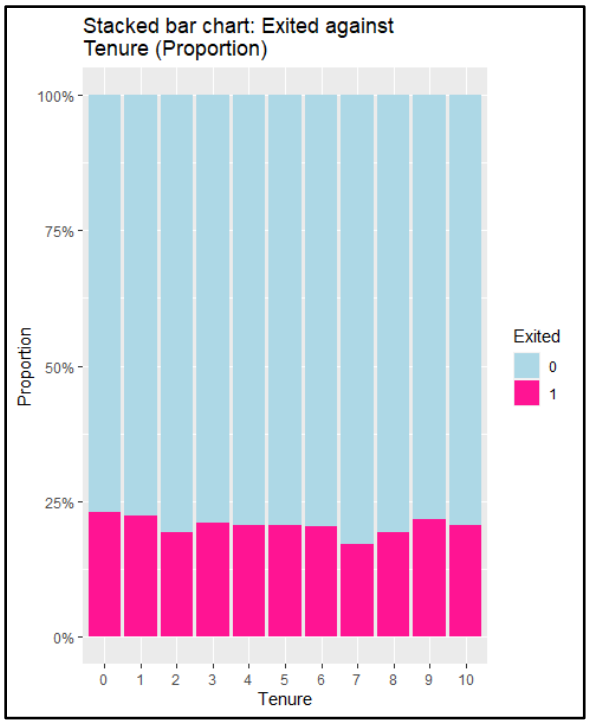
Variable: Age (Continuous)



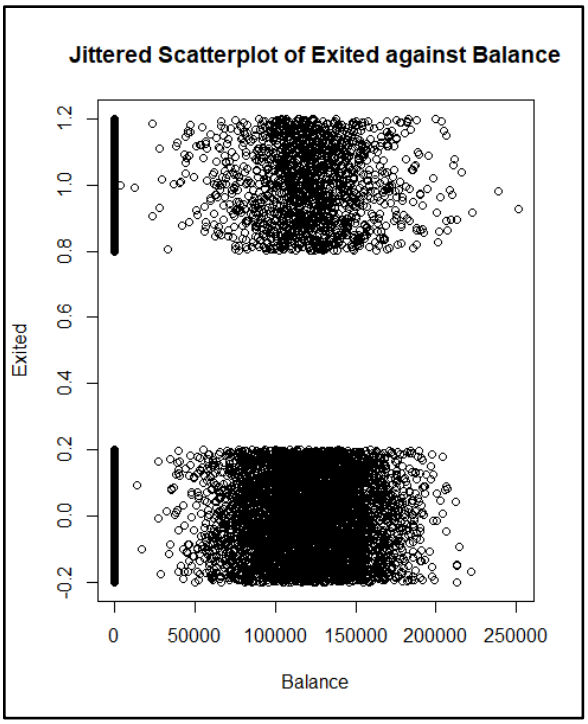


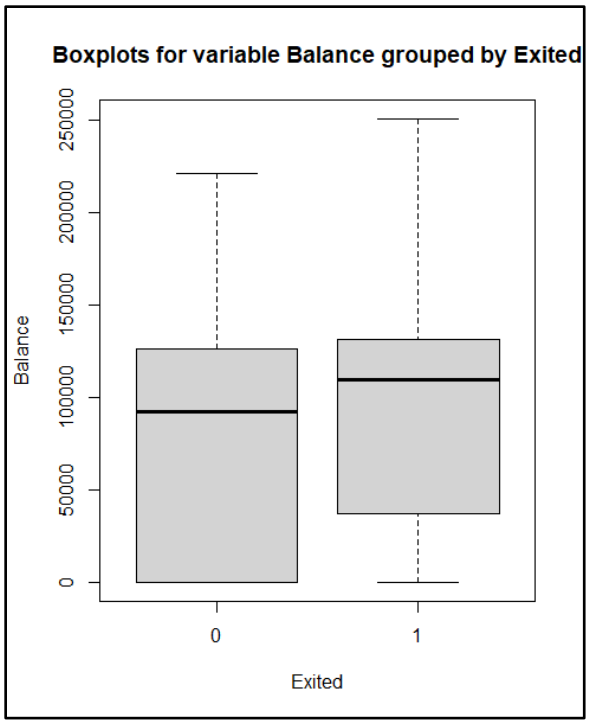
Variable: Tenure (Categorical)



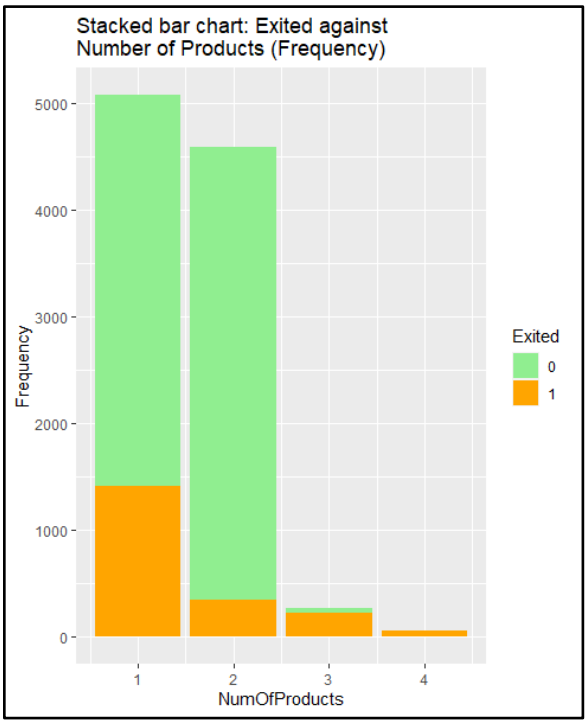


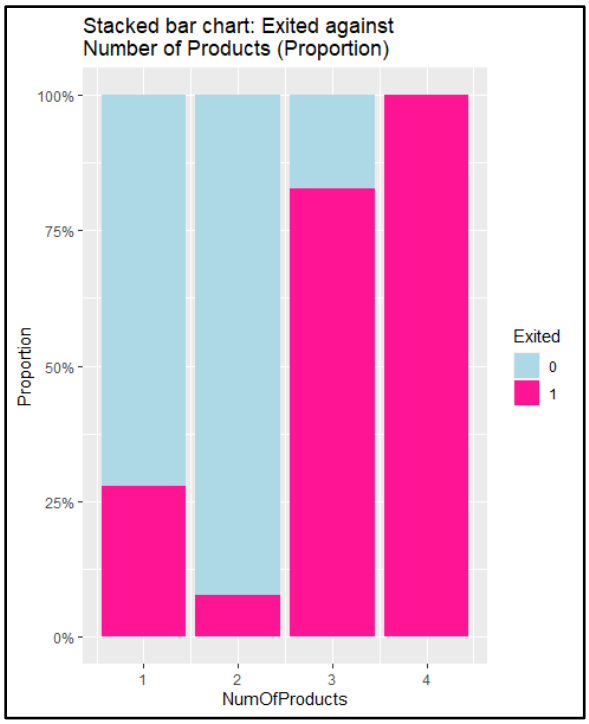
Variable: Balance (Continuous)



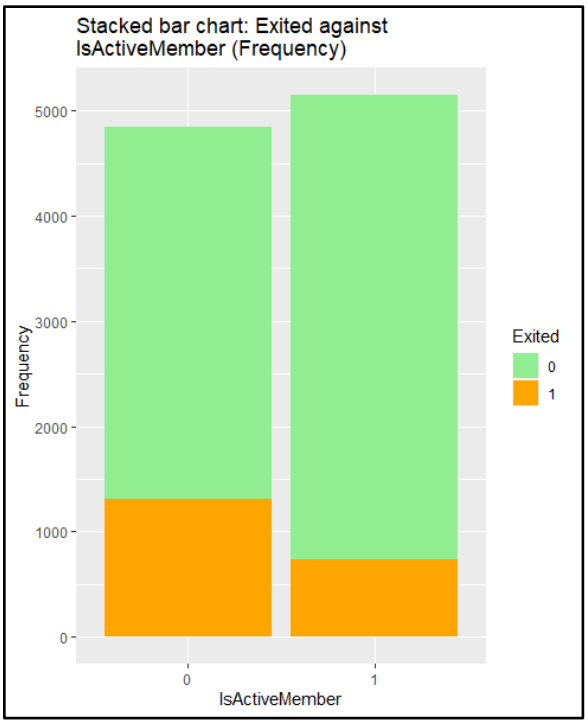


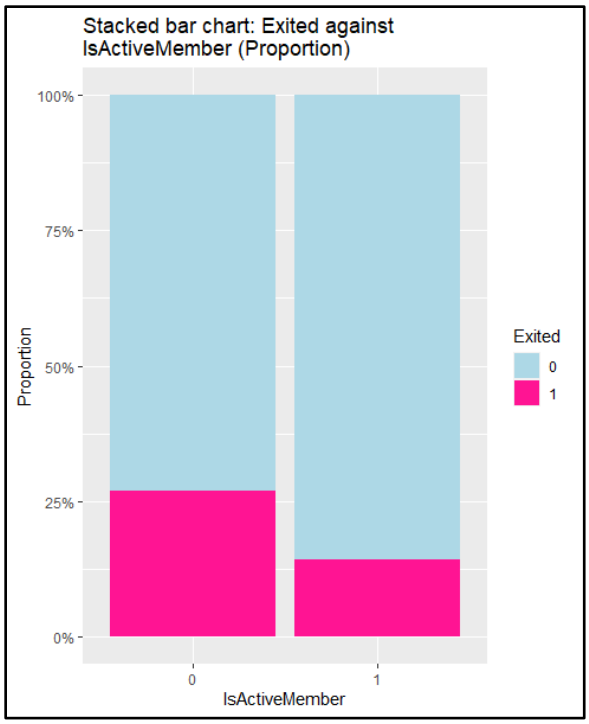
Variable: Number of Products (Categorical)



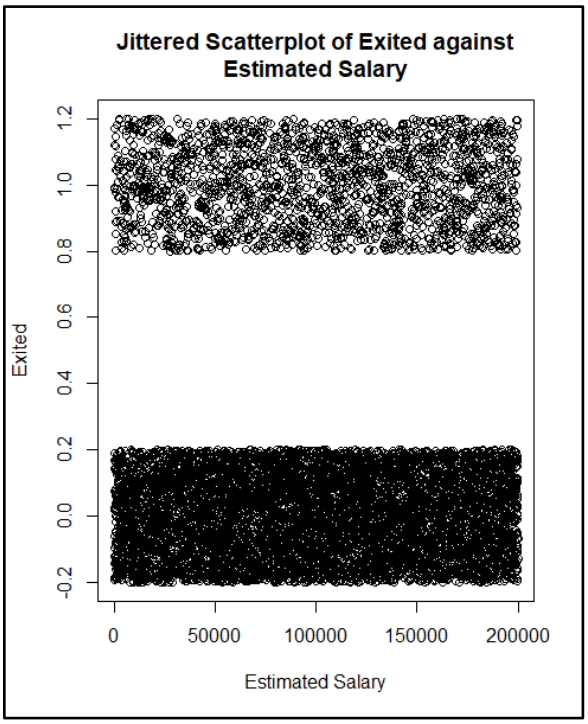


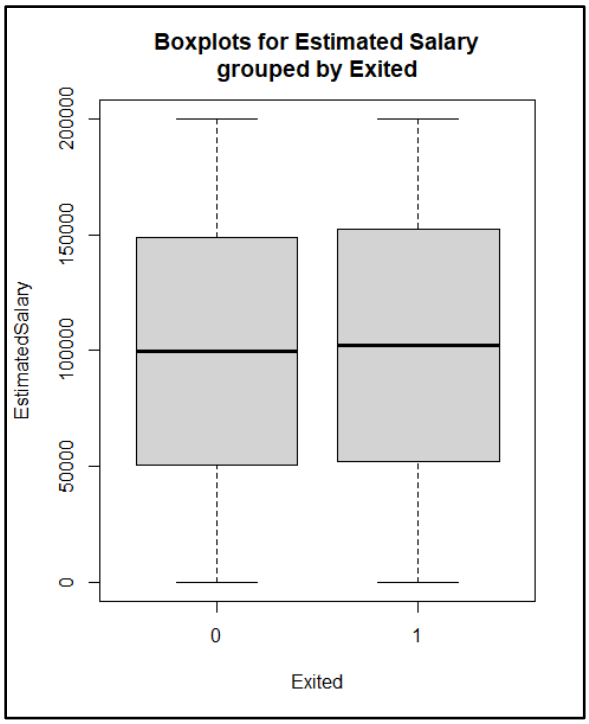
Variable: IsActiveMember (Categorical)



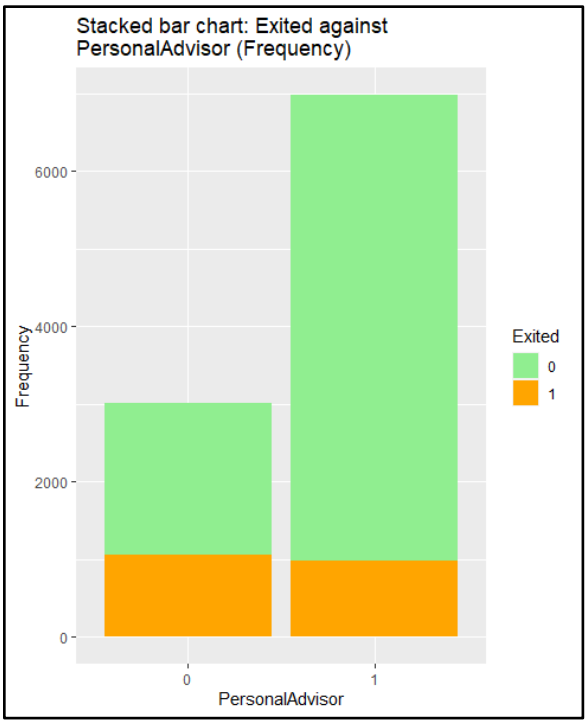


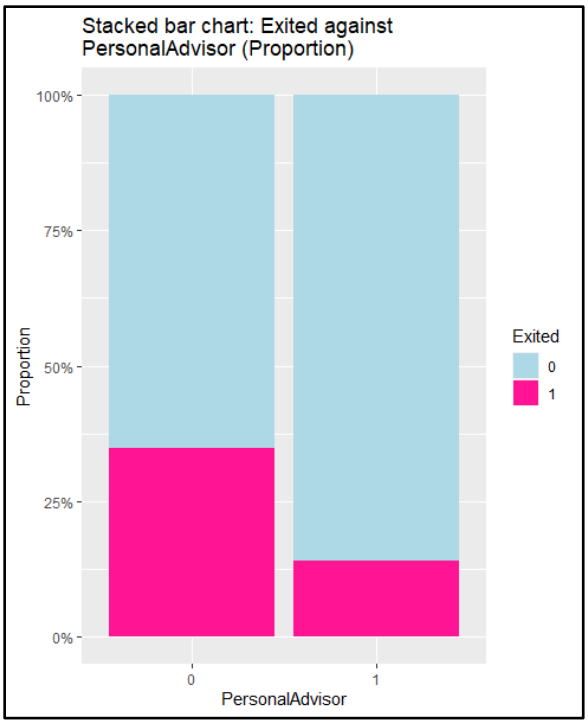
Variable: Estimated Salary (Continuous)



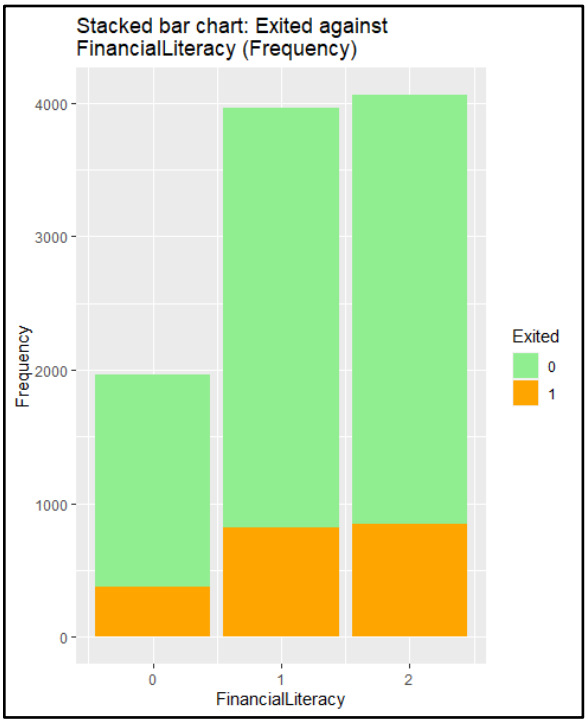


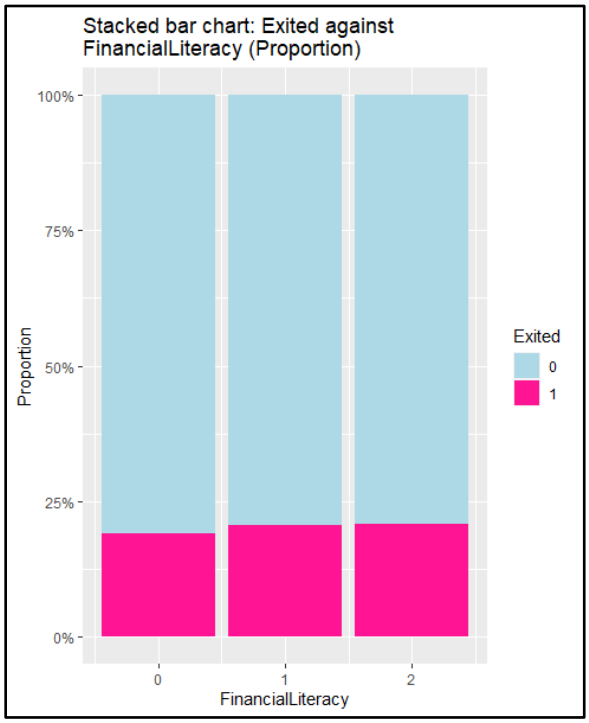
Variable: PersonalAdvisor (Categorical)



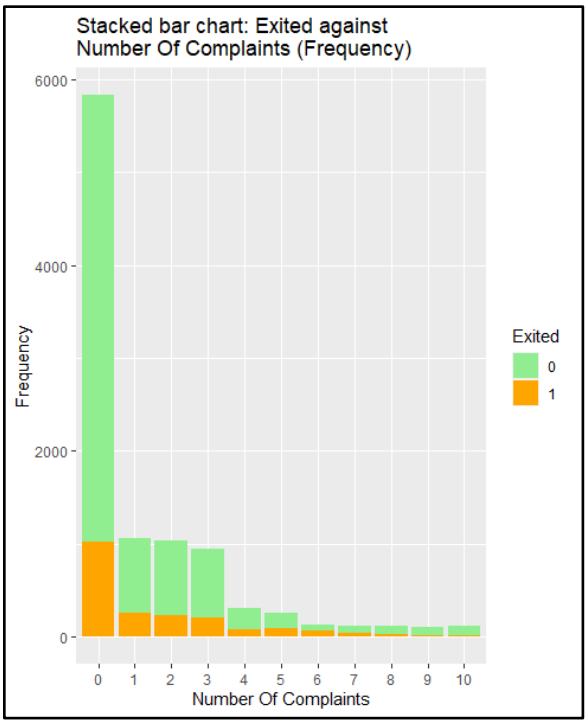


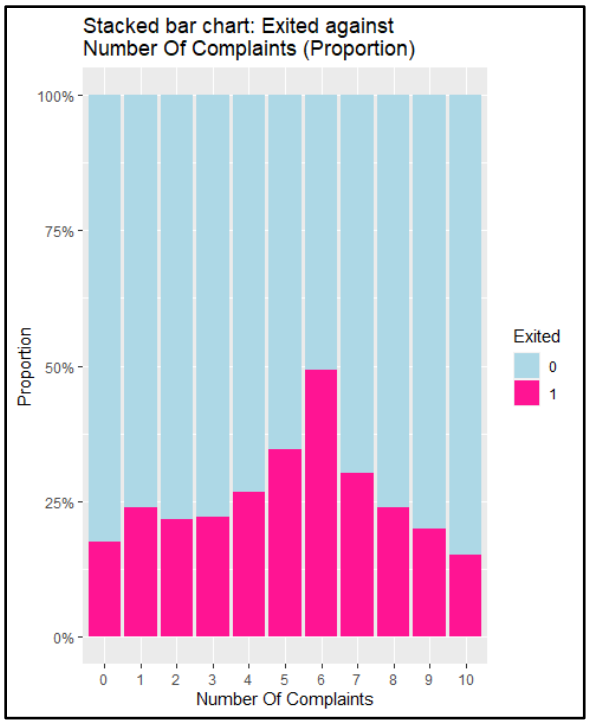
Variable: FinancialLiteracy (categorical)



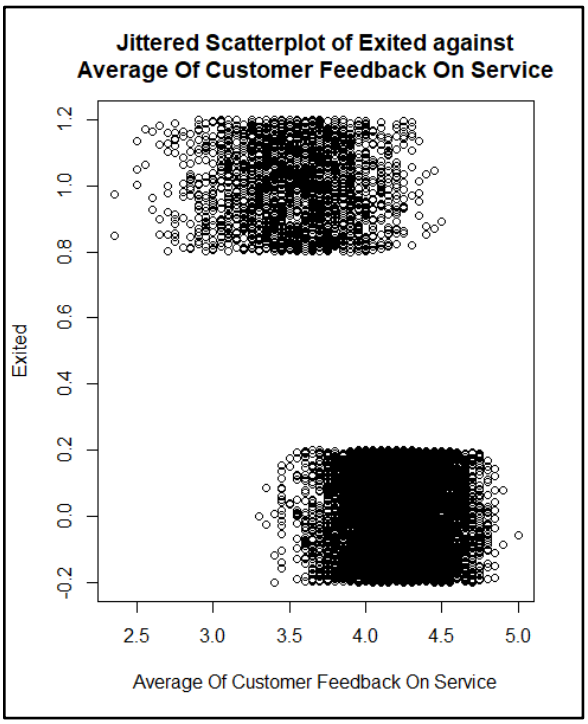


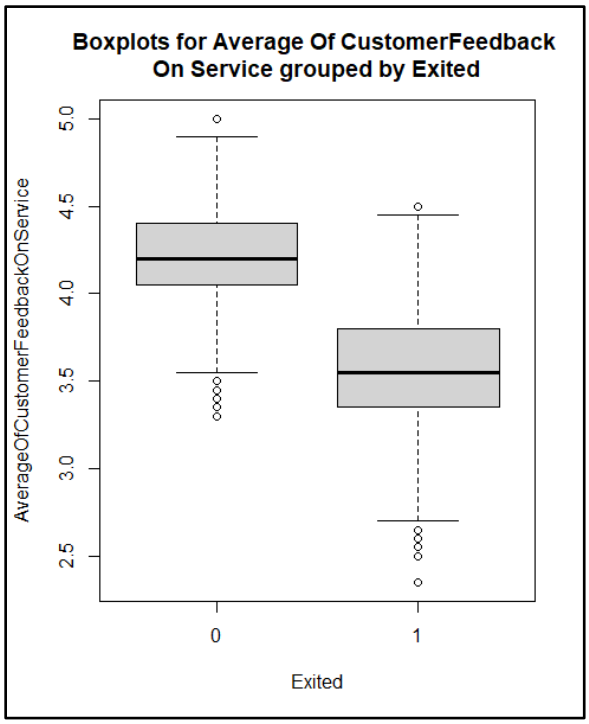
Variable: NumberOfComplaints (Categorical)



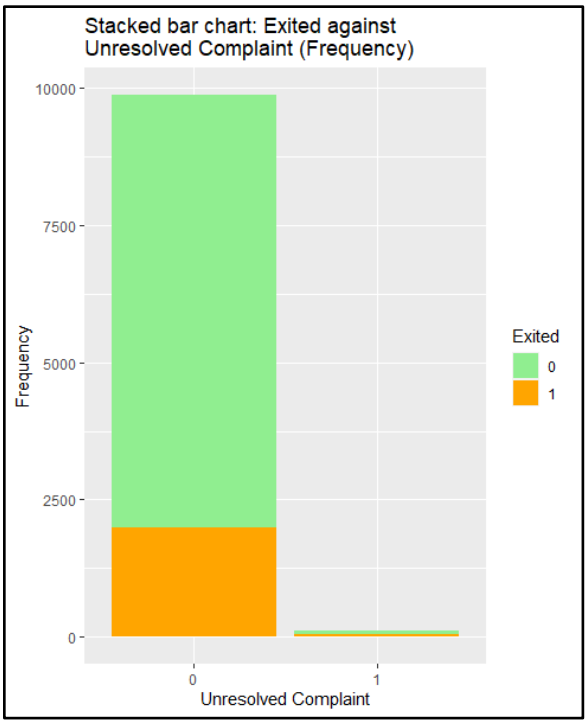


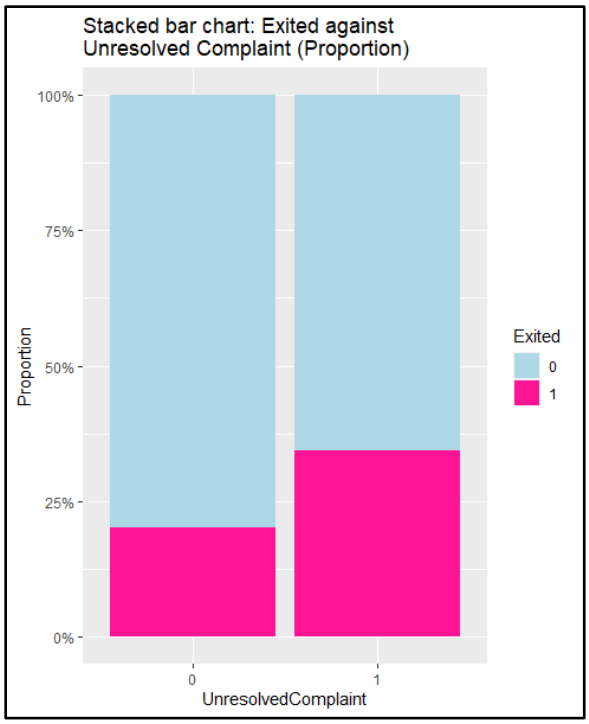
Variable: AverageOfCustomerFeedbackOnService (Continuous)



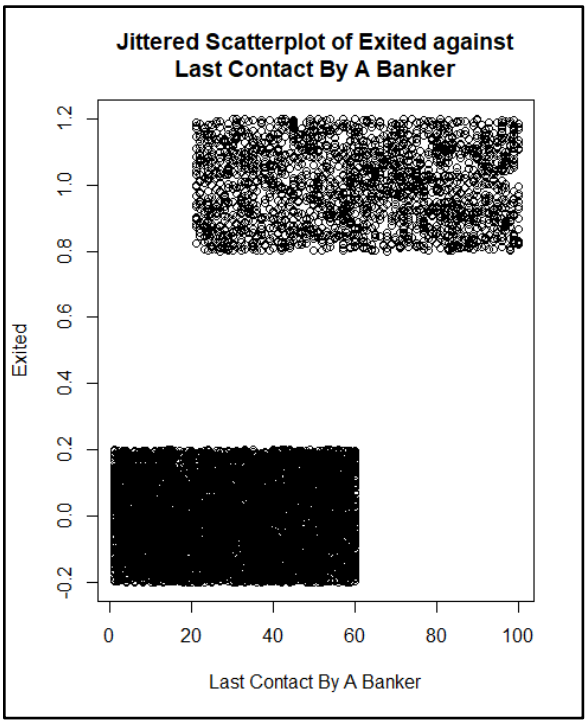


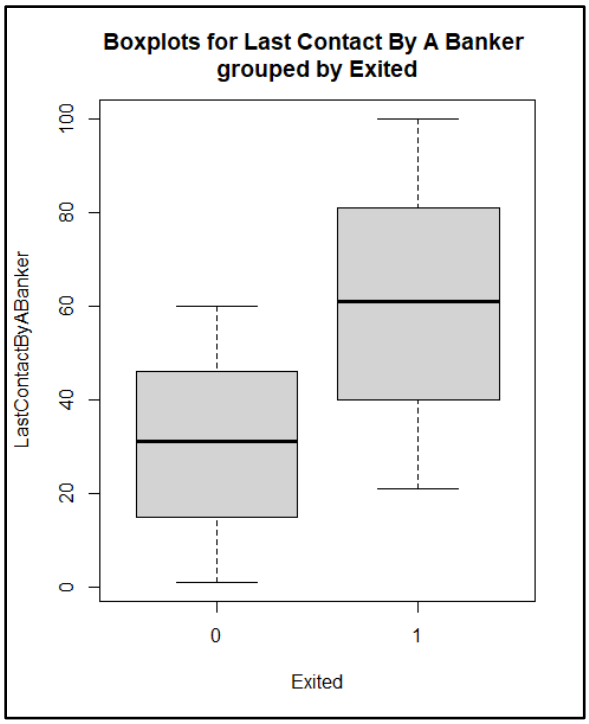
Variable: UnresolvedComplaint (Categorical)



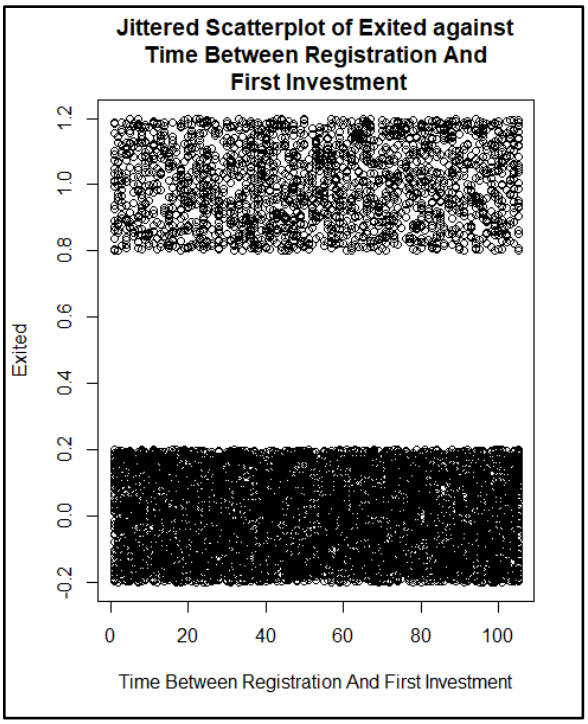


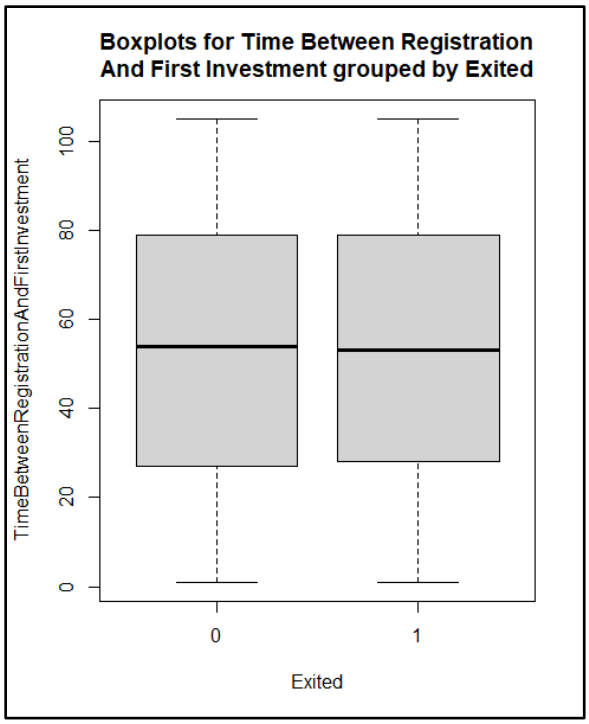
Variable: LastContactByABanker (Continuous)



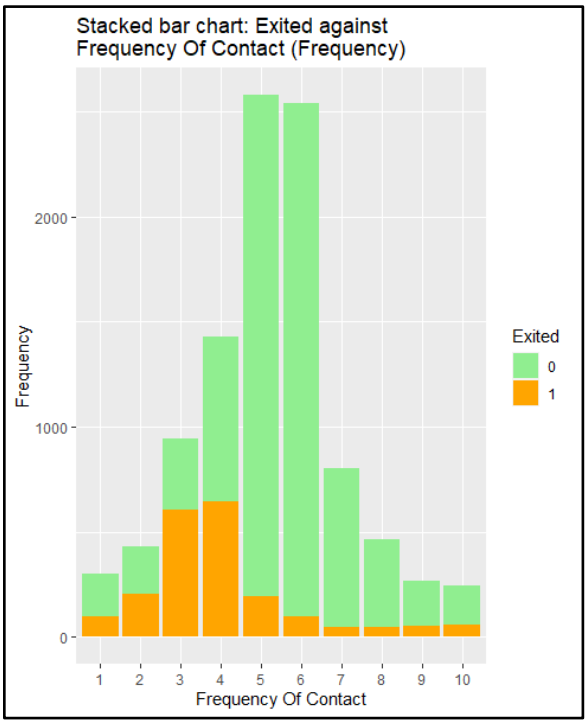


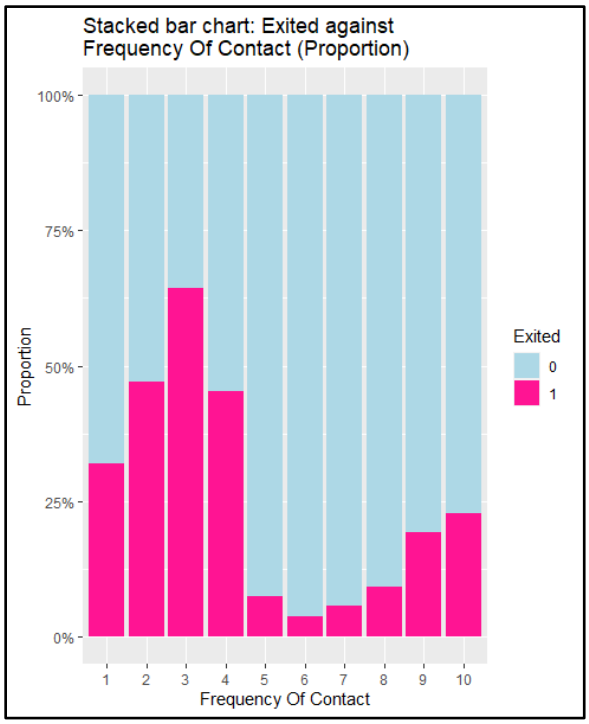
Variable: TimeBetweenRegistrationAndFirstInvestment (Continuous)





Variable: FrequencyOfContact (categorical)





APPENDIX B - LOGISTIC REGRESSION

For generating our CART model, we first import our data set and store it into a data table via the fread() function found in the data.table package. After which, we factorize all of the categorical variables via the factor() function

```
retention <- fread("data.csv", stringsAsFactors = T)

#Prepare data set
cols = ncol(retention)
retention <- retention[,5:cols]
retention$Exited <- factor(retention$Exited)
retention$FinancialLiteracy <- factor(retention$FinancialLiteracy)
retention$PersonalAdvisor <- factor(retention$PersonalAdvisor)
retention$UnresolvedComplaint <- factor(retention$UnresolvedComplaint)
retention$IsActiveMember <- factor(retention$IsActiveMember)
summary(retention)
```

We then split the data into train and test sets with a 7:3 ratio. The set.seed(2) ensures consistency in the split each time the code is run.

```
set.seed(2)
train <- sample.split(Y = retention$Exited, SplitRatio = 0.7)
trainset <- subset(retention, train == T)
testset <- subset(retention, train == F)
```

The logistic regression model is then fit on all the variables in the train set using the glm function.

```
m1 <- glm(Exited ~ . , family = binomial, data = trainset)
```

The results are as follows:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7691  -0.1554  -0.0525  -0.0095   4.3550

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            2.706e+01  1.289e+00  20.996  < 2e-16 ***
CreditScore                           -4.012e-04  6.643e-04  -0.604 0.545880
GeographyGermany                       5.878e-01  1.648e-01   3.567 0.000361 ***
GeographySpain                        -3.366e-02  1.689e-01  -0.199 0.842004
GenderMale                            -6.283e-01  1.322e-01  -4.754 2.00e-06 ***
Age                                    7.391e-02  6.300e-03  11.731  < 2e-16 ***
Tenure                                 6.086e-03  2.302e-02   0.264 0.791470
Balance                                3.343e-06  1.211e-06   2.760 0.005779 **
NumOfProducts                          2.076e-03  1.016e-01   0.020 0.983690
IsActiveMember1                       -1.003e+00  1.369e-01  -7.329 2.31e-13 ***
EstimatedSalary                        1.064e-06  1.148e-06   0.927 0.353819
PersonalAdvisor1                      -1.112e+00  1.354e-01  -8.215  < 2e-16 ***
FinancialLiteracy1                    -5.656e-02  1.855e-01  -0.305 0.760441
FinancialLiteracy2                     6.619e-02  1.821e-01   0.363 0.716248
NumberOfComplaints                     9.557e-02  2.889e-02   3.308 0.000939 ***
AverageOfCustomerFeedbackOnService    -8.093e+00  3.045e-01 -26.576  < 2e-16 ***
UnresolvedComplaint1                   7.120e-01  5.013e-01   1.420 0.155484
LastContactByABanker                   7.620e-02  4.022e-03  18.946  < 2e-16 ***
TimeBetweenRegistrationAndFirstInvestment  7.216e-05  2.185e-03   0.033 0.973650
FrequencyOfContact                    -4.307e-01  3.494e-02 -12.327  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7073.5  on 6997  degrees of freedom
Residual deviance: 1640.4  on 6978  degrees of freedom
AIC: 1680.4

Number of Fisher Scoring iterations: 8
```

At 5% significance level, only Geography, Gender, Age, Balance, isActiveMember, PersonalAdvisor, NumberOfComplaints, AverageOfCustomerFeedbackOnService, LastContactByABanker and FrequencyOfContact are statistically significant. A new model is fit with only these variables.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7833  -0.1579  -0.0527  -0.0095   4.3149

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       2.686e+01  1.184e+00  22.681  < 2e-16 ***
GeographyGermany                  5.752e-01  1.639e-01   3.510 0.000447 ***
GeographySpain                   -4.233e-02  1.680e-01  -0.252 0.801034
GenderMale                       -6.319e-01  1.315e-01  -4.806 1.54e-06 ***
Age                               7.387e-02  6.285e-03  11.753  < 2e-16 ***
Balance                           3.287e-06  1.185e-06   2.774 0.005531 **
IsActiveMember1                  -9.953e-01  1.362e-01  -7.309 2.69e-13 ***
PersonalAdvisor1                 -1.115e+00  1.352e-01  -8.245  < 2e-16 ***
NumberOfComplaints                9.656e-02  2.873e-02   3.361 0.000777 ***
AverageOfCustomerFeedbackOnService -8.067e+00  3.021e-01 -26.704  < 2e-16 ***
LastContactByABanker              7.649e-02  4.018e-03  19.034  < 2e-16 ***
FrequencyOfContact               -4.308e-01  3.482e-02 -12.370  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7073.5  on 6997  degrees of freedom
Residual deviance: 1644.3  on 6986  degrees of freedom
AIC: 1668.3

Number of Fisher Scoring iterations: 8
```

The odds ratio of all the variables and their confidence intervals are calculated using m2.

```
OR <- exp(coef(m2))
OR
OR.CI <- exp(confint(m2))
OR.CI
```

```
> OR
                  (Intercept)              GeographyGermany                GeographySpain                    GenderMale                           Age
                4.642077e+11                 1.777556e+00                 9.585526e-01                 5.316041e-01                  1.076666e+00
                      Balance                IsActiveMember1               PersonalAdvisor1            NumberOfComplaints AverageOfCustomerFeedbackOnService
                1.000003e+00                 3.696108e-01                 3.280519e-01                 1.101374e+00                  3.137351e-04
          LastContactByABanker             FrequencyOfContact
                1.079490e+00                 6.500027e-01
> OR.CI <- exp(confint(m2))
Waiting for profiling to be done...
> OR.CI
                                        2.5 %       97.5 %
(Intercept)                       4.832394e+10 5.031906e+12
GeographyGermany                  1.289908e+00 2.453027e+00
GeographySpain                    6.883521e-01 1.330367e+00
GenderMale                        4.102582e-01 6.871099e-01
Age                               1.063569e+00 1.090120e+00
Balance                           1.000001e+00 1.000006e+00
IsActiveMember1                   2.823603e-01 4.817167e-01
PersonalAdvisor1                  2.513199e-01 4.270950e-01
NumberOfComplaints                1.040428e+00 1.164592e+00
AverageOfCustomerFeedbackOnService 1.705385e-04 5.577450e-04
LastContactByABanker              1.071238e+00 1.088258e+00
FrequencyOfContact                6.066593e-01 6.954404e-01
```

This information is visualised using the following code. The plotting is done using ggplot and a log scale is used for the x-axis. The values of odds ratios and their confidence intervals are manually coded. The resultant graph clearly shows the deviation of odds ratio from 1.

```
boxOdds <- c(OR[c(seq(2,length(OR),1))])
n_ci <-  length(OR.CI)
boxCILow<-OR.CI[c(seq(2,n_ci/2,1))]
boxCIHigh<- OR.CI[c(seq(n_ci/2+2,n_ci,1))]
#----visualisation of the odds ratios-----
boxLabels <- names(boxOdds)
df <- data.frame(yAxis = length(boxLabels):1, boxOdds, boxCILow, boxCIHigh)
p <- ggplot(df, aes(x = boxOdds, y = yAxis))
p + geom_vline(aes(xintercept = 1), size = .25, linetype = "dashed") +
  geom_errorbarh(aes(xmax = boxCIHigh, xmin = boxCILow), size = .5, height = .2, color = "gray50") +
  geom_point(size = 3.5, color = "orange") +
  theme_bw() +
  theme(panel.grid.minor = element_blank()) +
  scale_y_continuous(breaks = length(boxLabels):1, labels = boxLabels) +
  scale_x_continuous(breaks = c(seq(0,1,0.2), seq(1,5,1) )) +
  coord_trans(x = "log10") +
  ylab("") +
  xlab("Odds ratio (log scale)") +
  ggtitle("Factors affecting Customer Retention")
```

The model is used to predict the Exited values for both the train set and test set. The confusion matrix of the results, as well as the accuracy of the prediction are shown below:

```
> # Confusion Matrix on Trainset
> threshold <- 0.5
> prob.train <- predict(m2, type = 'response')
> m2.predict.train <- ifelse(prob.train > threshold, 1, 0)
> table1 <- table(Trainset.Actual = trainset$Exited, m2.predict.train, deparse.level = 2)
> table1
               m2.predict.train
Trainset.Actual    0    1
              0 5456  117
              1  200 1225
> round(prop.table(table1),3)
               m2.predict.train
Trainset.Actual     0     1
              0 0.780 0.017
              1 0.029 0.175
> # Overall Accuracy
> mean(m2.predict.train == trainset$Exited)
[1] 0.9547013
> # Confusion Matrix on Testset
> prob.test <- predict(m2, newdata = testset, type = 'response')
> m2.predict.test <- ifelse(prob.test> threshold, 1, 0)
> table2 <- table(Testset.Actual = testset$Exited, m2.predict.test, deparse.level = 2)
> table2
               m2.predict.test
Testset.Actual    0    1
             0 2339   49
             1   84  527
> round(prop.table(table2), 3)
               m2.predict.test
Testset.Actual     0     1
             0 0.780 0.016
             1 0.028 0.176
> # Overall Accuracy
> mean(m2.predict.test == testset$Exited)
[1] 0.9556519
```

Backward elimination was done on the full model using the step function, which eliminates variables based on Akaike information criterion. The variables chosen using this method are the same as those chosen based on significance at 5% significance level.

```
m3<- step(m1)
```

```
Step:  AIC=1668.3
Exited ~ Geography + Gender + Age + Balance + IsActiveMember +
    PersonalAdvisor + NumberOfComplaints + AverageOfCustomerFeedbackOnService +
    LastContactByABanker + FrequencyOfContact

                                  Df Deviance    AIC
<none>                                 1644.3 1668.3
- Balance                           1   1652.0 1674.0
- NumberOfComplaints                1   1655.2 1677.2
- Geography                         2   1658.9 1678.9
- Gender                            1   1667.8 1689.8
- IsActiveMember                    1   1700.1 1722.1
- PersonalAdvisor                   1   1713.5 1735.5
- Age                               1   1787.4 1809.4
- FrequencyOfContact                1   1813.0 1835.0
- LastContactByABanker              1   2249.7 2271.7
- AverageOfCustomerFeedbackOnService  1   3562.5 3584.5
> |
```

The GVIF values are checked using the vif() function. The adjusted GVIF values are all below 2.

```
> vif(m2)
                                     GVIF Df GVIF^(1/(2*Df))
Geography                        1.234499  2        1.054078
Gender                           1.017886  1        1.008903
Age                              1.109149  1        1.053162
Balance                          1.234699  1        1.111170
IsActiveMember                   1.084838  1        1.041556
PersonalAdvisor                  1.029585  1        1.014685
NumberOfComplaints               1.017776  1        1.008849
AverageOfCustomerFeedbackOnService 1.182418  1        1.087391
LastContactByABanker             1.118933  1        1.057796
FrequencyOfContact               1.038108  1        1.018876
> |
```

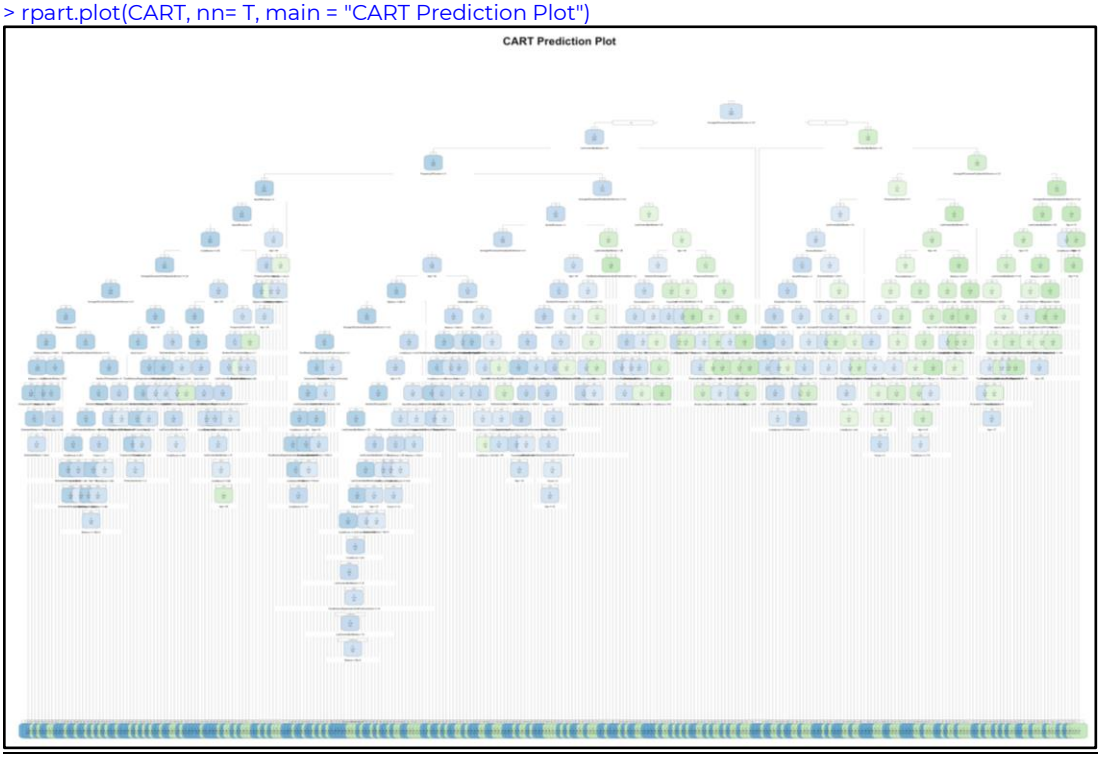
APPENDIX C - CART MODEL

For generating our CART model, we first import our data set and store it into a data table via the fread() function found in the data.table package. After which, we factorize all of the categorical variables via the factor() function

```
## Import data set
> library(data.table)
> dt <- fread('https://raw.githubusercontent.com/ernestang98/Customer-Churn/main/data.csv', stringsAsFactors=T)

## Prepare data set
> dt$HasCrCard<- factor(dt$HasCrCard)
> dt$IsActiveMember<- factor(dt$IsActiveMember)
> dt$Exited<- factor(dt$Exited)
> dt$PersonalAdvisor<- factor(dt$PersonalAdvisor)
> dt$FinancialLiteracy<- factor(dt$FinancialLiteracy)
> dt$UnresolvedComplaint<- factor(dt$UnresolvedComplaint)
> cols = ncol(dt)
> pData <- dt[,11:cols]
```

Next we will start building our CART model using the rpart package. To allow the CART to grow to its maximum length, we set the complexity parameter (cp) to 0. To eventually visualize our classification tree, we will utilize the rpart.plot package as well

```
## Building CART Model
> library(rpart)
> library(rpart.plot)
> set.seed(2004)
> CART <- rpart(Exited ~ ., data = pData, method = 'class', control = rpart.control(minsplit = 2, cp = 0))
> rpart.plot(CART, nn= T, main = "CART Prediction Plot")
```



```
> print(CART)
n= 9997

node), split, n, loss, yval, (yprob)
      * denotes terminal node

    1) root 9997 2036 0 (0.7963389017 0.2036610983)
      2) AverageOfCustomerFeedbackOnService>=3.775 8202  545 0 (0.9335527920 0.0664472080)
        4) LastContactByABanker< 60.5 7931  274 0 (0.9654520237 0.0345479763)
          8) FrequencyOfContact>=4.5 6230   56 0 (0.9910112360 0.0089887640)
           16) NumOfProducts< 3.5 6228   54 0 (0.9913294798 0.0086705202)
             32) NumOfProducts< 2.5 6180   46 0 (0.9925566343 0.0074433657)
               64) CreditScore>=377.5 6179   45 0 (0.9927172682 0.0072827318)
                128) AverageOfCustomerFeedbackOnService>=3.925 5640   23 0 (0.9959219858 0.0040780142)
                  256) AverageOfCustomerFeedbackOnService>=4.025 5029   13 0 (0.9974149330 0.0025850070)
                    512) PersonalAdvisor=1 3781    4 0 (0.9989420788 0.0010579212)
                     1024) EstimatedSalary>=5092.925 3668    2 0 (0.9994547437 0.0005452563)
                       2048) Balance< 185071 3636    1 0 (0.9997249725 0.0002750275)
                         4096) FrequencyOfContact< 9.5 3531    0 0 (1.0000000000 0.0000000000) *
                         4097) FrequencyOfContact>=9.5 105    1 0 (0.9904761905 0.0095238095)
                           8194) EstimatedSalary>=14972.8 98    0 0 (1.0000000000 0.0000000000) *
                           8195) EstimatedSalary< 14972.8 7    1 0 (0.8571428571 0.1428571429)
                            16390) EstimatedSalary< 13695.98 6    0 0 (1.0000000000 0.0000000000) *
                            16391) EstimatedSalary>=13695.98 1    0 1 (0.0000000000 1.0000000000) *
                       2049) Balance>=185071 32    1 0 (0.9687500000 0.0312500000)
                         4098) Balance>=185296 31    0 0 (1.0000000000 0.0000000000) *
                         4099) Balance< 185296 1    0 1 (0.0000000000 1.0000000000) *
                     1025) EstimatedSalary< 5092.925 113    2 0 (0.9823008850 0.0176991150)
                       2050) EstimatedSalary< 5001.085 112    1 0 (0.9910714286 0.0089285714)
                         4100) Age< 50.5 103    0 0 (1.0000000000 0.0000000000) *
                         4101) Age>=50.5 9    1 0 (0.8888888889 0.1111111111)
                           8202) CreditScore>=650 8    0 0 (1.0000000000 0.0000000000) *
                           8203) CreditScore< 650 1    0 1 (0.0000000000 1.0000000000) *
                       2051) EstimatedSalary>=5001.085 1    0 1 (0.0000000000 1.0000000000) *
                    513) PersonalAdvisor=0 1248    9 0 (0.9927884615 0.0072115385)
                     1026) AverageOfCustomerFeedbackOnService>=4.225 784    0 0 (1.0000000000 0.0000000000) *
                     1027) AverageOfCustomerFeedbackOnService< 4.225 464    9 0 (0.9806034483 0.0193965517)
                       2054) NumberOfComplaints< 3.5 441    6 0 (0.9863945578 0.0136054422)
                         4108) EstimatedSalary< 197819.4 432    5 0 (0.9884259259 0.0115740741)
                           8216) LastContactByABanker< 48.5 354    2 0 (0.9943502825 0.0056497175)
                            16432) CreditScore>=501 334    1 0 (0.9970059880 0.0029940120)
                              32864) EstimatedSalary>=26356.33 297    0 0 (1.0000000000 0.0000000000) *
                              32865) EstimatedSalary< 26356.33 37    1 0 (0.9729729730 0.0270270270)
                                65730) EstimatedSalary< 25582.72 36    0 0 (1.0000000000 0.0000000000) *
                                65731) EstimatedSalary>=25582.72 1    0 1 (0.0000000000 1.0000000000) *
                            16433) CreditScore< 501 20    1 0 (0.9500000000 0.0500000000)
                              32866) CreditScore< 494.5 18    0 0 (1.0000000000 0.0000000000) *
                              32867) CreditScore>=494.5 2    1 0 (0.5000000000 0.5000000000)
                                65734) Gender=Male 1    0 0 (1.0000000000 0.0000000000) *
                                65735) Gender=Female 1    0 1 (0.0000000000 1.0000000000) *
                           8217) LastContactByABanker>=48.5 78    3 0 (0.9615384615 0.0384615385)
                            16434) Tenure>=0.5 76    2 0 (0.9736842105 0.0263157895)
                              32868) Age< 49.5 72    1 0 (0.9861111111 0.0138888889)
                                65736) Balance< 152149.5 67    0 0 (1.0000000000 0.0000000000) *
                                65737) Balance>=152149.5 5    1 0 (0.8000000000 0.2000000000)
                                 131474) Balance>=156135.1 4    0 0 (1.0000000000 0.0000000000) *
                                 131475) Balance< 156135.1 1    0 1 (0.0000000000 1.0000000000) *
                              32869) Age>=49.5 4    1 0 (0.7500000000 0.2500000000)
                                65738) CreditScore>=695.5 3    0 0 (1.0000000000 0.0000000000) *
                                65739) CreditScore< 695.5 1    0 1 (0.0000000000 1.0000000000) *
                            16435) Tenure< 0.5 2    1 0 (0.5000000000 0.5000000000)
                              32870) CreditScore< 699 1    0 0 (1.0000000000 0.0000000000) *
                              32871) CreditScore>=699 1    0 1 (0.0000000000 1.0000000000) *
                         4109) EstimatedSalary>=197819.4 9    1 0 (0.8888888889 0.1111111111)
                           8218) EstimatedSalary>=198184.8 8    0 0 (1.0000000000 0.0000000000) *
                           8219) EstimatedSalary< 198184.8 1    0 1 (0.0000000000 1.0000000000) *
                       2055) NumberOfComplaints>=3.5 23    3 0 (0.8695652174 0.1304347826)
                         4110) AverageOfCustomerFeedbackOnService< 4.175 17    0 0 (1.0000000000 0.0000000000) *
                         4111) AverageOfCustomerFeedbackOnService>=4.175 6    3 0 (0.5000000000 0.5000000000)
                           8222) FrequencyOfContact>=5.5 3    0 0 (1.0000000000 0.0000000000) *
                           8223) FrequencyOfContact< 5.5 3    0 1 (0.0000000000 1.0000000000) *
                  257) AverageOfCustomerFeedbackOnService< 4.025 611   10 0 (0.9836333879 0.0163666121)
                    514) Age< 50.5 547    4 0 (0.9926873857 0.0073126143)
                     1028) HasCrCard=1 389    0 0 (1.0000000000 0.0000000000) *
                     1029) HasCrCard=0 158    4 0 (0.9746835443 0.0253164557)
```

```
        2058) TimeBetweenRegistrationAndFirstInvestment< 103.5 155   3 0 (0.9806451613 0.0193548387)
          4116) EstimatedSalary< 190464.2 151   2 0 (0.9867549669 0.0132450331)
            8232) NumberOfComplaints< 6.5 146   1 0 (0.9931506849 0.0068493151)
              16464) FrequencyOfContact< 9.5 139   0 0 (1.0000000000 0.0000000000) *
              16465) FrequencyOfContact>=9.5 7   1 0 (0.8571428571 0.1428571429)
                32930) FinancialLiteracy=1,2 6   0 0 (1.0000000000 0.0000000000) *
                32931) FinancialLiteracy=0 1   0 1 (0.0000000000 1.0000000000) *
            8233) NumberOfComplaints>=6.5 5   1 0 (0.8000000000 0.2000000000)
              16466) CreditScore< 699 4   0 0 (1.0000000000 0.0000000000) *
              16467) CreditScore>=699 1   0 1 (0.0000000000 1.0000000000) *
          4117) EstimatedSalary>=190464.2 4   1 0 (0.7500000000 0.2500000000)
            8234) Tenure< 4 3   0 0 (1.0000000000 0.0000000000) *
            8235) Tenure>=4 1   0 1 (0.0000000000 1.0000000000) *
        2059) TimeBetweenRegistrationAndFirstInvestment>=103.5 3   1 0 (0.6666666667 0.3333333333)
          4118) Geography=France,Spain 2   0 0 (1.0000000000 0.0000000000) *
          4119) Geography=Germany 1   0 1 (0.0000000000 1.0000000000) *
      515) Age>=50.5 64   6 0 (0.9062500000 0.0937500000)
        1030) EstimatedSalary< 135952.9 46   1 0 (0.9782608696 0.0217391304)
          2060) FrequencyOfContact< 8.5 43   0 0 (1.0000000000 0.0000000000) *
          2061) FrequencyOfContact>=8.5 3   1 0 (0.6666666667 0.3333333333)
            4122) Age>=59 2   0 0 (1.0000000000 0.0000000000) *
            4123) Age< 59 1   0 1 (0.0000000000 1.0000000000) *
        1031) EstimatedSalary>=135952.9 18   5 0 (0.7222222222 0.2777777778)
          2062) CreditScore< 702.5 14   2 0 (0.8571428571 0.1428571429)
            4124) EstimatedSalary>=141965.7 13   1 0 (0.9230769231 0.0769230769)
              8248) LastContactByABanker>=24 11   0 0 (1.0000000000 0.0000000000) *
              8249) LastContactByABanker< 24 2   1 0 (0.5000000000 0.5000000000)
                16498) CreditScore>=602.5 1   0 0 (1.0000000000 0.0000000000) *
                16499) CreditScore< 602.5 1   0 1 (0.0000000000 1.0000000000) *
            4125) EstimatedSalary< 141965.7 1   0 1 (0.0000000000 1.0000000000) *
          2063) CreditScore>=702.5 4   1 1 (0.2500000000 0.7500000000)
            4126) Gender=Female 1   0 0 (1.0000000000 0.0000000000) *
            4127) Gender=Male 3   0 1 (0.0000000000 1.0000000000) *
    129) AverageOfCustomerFeedbackOnService< 3.925 539   22 0 (0.9591836735 0.0408163265)
      258) Age< 43.5 427   2 0 (0.9953161593 0.0046838407)
        516) Age< 37.5 311   0 0 (1.0000000000 0.0000000000) *
        517) Age>=37.5 116   2 0 (0.9827586207 0.0172413793)
          1034) PersonalAdvisor=1 84   0 0 (1.0000000000 0.0000000000) *
          1035) PersonalAdvisor=0 32   2 0 (0.9375000000 0.0625000000)
            2070) HasCrCard=1 17   0 0 (1.0000000000 0.0000000000) *
            2071) HasCrCard=0 15   2 0 (0.8666666667 0.1333333333)
              4142) EstimatedSalary>=77795.64 9   0 0 (1.0000000000 0.0000000000) *
              4143) EstimatedSalary< 77795.64 6   2 0 (0.6666666667 0.3333333333)
                8286) CreditScore< 644 4   0 0 (1.0000000000 0.0000000000) *
                8287) CreditScore>=644 2   0 1 (0.0000000000 1.0000000000) *
      259) Age>=43.5 112   20 0 (0.8214285714 0.1785714286)
        518) FrequencyOfContact< 8.5 101   11 0 (0.8910891089 0.1089108911)
          1036) NumberOfComplaints< 5.5 95   7 0 (0.9263157895 0.0736842105)
            2072) LastContactByABanker< 59 94   6 0 (0.9361702128 0.0638297872)
              4144) TimeBetweenRegistrationAndFirstInvestment>=6.5 90   4 0 (0.9555555556 0.0444444444)
                8288) FinancialLiteracy=1,2 68   0 0 (1.0000000000 0.0000000000) *
                8289) FinancialLiteracy=0 22   4 0 (0.8181818182 0.1818181818)
                  16578) LastContactByABanker< 36.5 13   0 0 (1.0000000000 0.0000000000) *
                  16579) LastContactByABanker>=36.5 9   4 0 (0.5555555556 0.4444444444)
                    33158) CreditScore>=695 4   0 0 (1.0000000000 0.0000000000) *
                    33159) CreditScore< 695 5   1 1 (0.2000000000 0.8000000000)
                      66318) Age< 45.5 1   0 0 (1.0000000000 0.0000000000) *
                      66319) Age>=45.5 4   0 1 (0.0000000000 1.0000000000) *
              4145) TimeBetweenRegistrationAndFirstInvestment< 6.5 4   2 0 (0.5000000000 0.5000000000)
                8290) CreditScore>=633.5 2   0 0 (1.0000000000 0.0000000000) *
                8291) CreditScore< 633.5 2   0 1 (0.0000000000 1.0000000000) *
            2073) LastContactByABanker>=59 1   0 1 (0.0000000000 1.0000000000) *
          1037) NumberOfComplaints>=5.5 6   2 1 (0.3333333333 0.6666666667)
            2074) NumOfProducts>=1.5 2   0 0 (1.0000000000 0.0000000000) *
            2075) NumOfProducts< 1.5 4   0 1 (0.0000000000 1.0000000000) *
        519) FrequencyOfContact>=8.5 11   2 1 (0.1818181818 0.8181818182)
          1038) Tenure>=6.5 3   1 0 (0.6666666667 0.3333333333)
            2076) CreditScore< 665 2   0 0 (1.0000000000 0.0000000000) *
            2077) CreditScore>=665 1   0 1 (0.0000000000 1.0000000000) *
          1039) Tenure< 6.5 8   0 1 (0.0000000000 1.0000000000) *
  65) CreditScore< 377.5 1   0 1 (0.0000000000 1.0000000000) *
33) NumOfProducts>=2.5 48   8 0 (0.8333333333 0.1666666667)
  66) Age< 43.5 40   3 0 (0.9250000000 0.0750000000)
    132) FrequencyOfContact< 7.5 37   1 0 (0.9729729730 0.0270270270)
      264) Balance< 131553.9 34   0 0 (1.0000000000 0.0000000000) *
```

```
       265) Balance>=131553.9 3    1 0 (0.6666666667 0.3333333333)
         530) Age< 36 2    0 0 (1.0000000000 0.0000000000) *
         531) Age>=36 1    0 1 (0.0000000000 1.0000000000) *
       133) FrequencyOfContact>=7.5 3    1 1 (0.3333333333 0.6666666667)
         266) CreditScore>=635 1    0 0 (1.0000000000 0.0000000000) *
         267) CreditScore< 635 2    0 1 (0.0000000000 1.0000000000) *
      67) Age>=43.5 8    3 1 (0.3750000000 0.6250000000)
       134) Balance< 37137.43 4    1 0 (0.7500000000 0.2500000000)
         268) IsActiveMember=1 3    0 0 (1.0000000000 0.0000000000) *
         269) IsActiveMember=0 1    0 1 (0.0000000000 1.0000000000) *
       135) Balance>=37137.43 4    0 1 (0.0000000000 1.0000000000) *
  17) NumOfProducts>=3.5 2    0 1 (0.0000000000 1.0000000000) *
 9) FrequencyOfContact< 4.5 1701  218 0 (0.8718400941 0.1281599059)
  18) AverageOfCustomerFeedbackOnService>=3.925 1486  106 0 (0.9286675639 0.0713324361)
    36) NumOfProducts< 2.5 1465   89 0 (0.9392491468 0.0607508532)
      72) AverageOfCustomerFeedbackOnService>=4.025 1255   49 0 (0.9609561753 0.0390438247)
       144) Age< 42.5 993   22 0 (0.9778449144 0.0221550856)
         288) Balance< 218307.3 992   21 0 (0.9788306452 0.0211693548)
           576) AverageOfCustomerFeedbackOnService>=4.225 626    5 0 (0.9920127796 0.0079872204)
            1152) TimeBetweenRegistrationAndFirstInvestment>=1.5 621    4 0 (0.9935587762 0.0064412238)
              2304) UnresolvedComplaint=0 615    3 0 (0.9951219512 0.0048780488)
                4608) LastContactByABanker< 57.5 583    2 0 (0.9965694683 0.0034305317)
                  9216) CreditScore>=519.5 532    1 0 (0.9981203008 0.0018796992)
                    18432) TimeBetweenRegistrationAndFirstInvestment< 94.5 459    0 0 (1.0000000000 0.0000000000) *
                    18433) TimeBetweenRegistrationAndFirstInvestment>=94.5 73    1 0 (0.9863013699 0.0136986301)
                      36866) CreditScore< 759 63    0 0 (1.0000000000 0.0000000000) *
                      36867) CreditScore>=759 10    1 0 (0.9000000000 0.1000000000)
                        73734) CreditScore>=761 9    0 0 (1.0000000000 0.0000000000) *
                        73735) CreditScore< 761 1    0 1 (0.0000000000 1.0000000000) *
                  9217) CreditScore< 519.5 51    1 0 (0.9803921569 0.0196078431)
                    18434) CreditScore< 517.5 49    0 0 (1.0000000000 0.0000000000) *
                    18435) CreditScore>=517.5 2    1 0 (0.5000000000 0.5000000000)
                      36870) Geography=France 1    0 0 (1.0000000000 0.0000000000) *
                      36871) Geography=Germany 1    0 1 (0.0000000000 1.0000000000) *
                4609) LastContactByABanker>=57.5 32    1 0 (0.9687500000 0.0312500000)
                  9218) Age< 40.5 28    0 0 (1.0000000000 0.0000000000) *
                  9219) Age>=40.5 4    1 0 (0.7500000000 0.2500000000)
                    18438) EstimatedSalary< 152324.6 3    0 0 (1.0000000000 0.0000000000) *
                    18439) EstimatedSalary>=152324.6 1    0 1 (0.0000000000 1.0000000000) *
              2305) UnresolvedComplaint=1 6    1 0 (0.8333333333 0.1666666667)
                4610) LastContactByABanker< 56 5    0 0 (1.0000000000 0.0000000000) *
                4611) LastContactByABanker>=56 1    0 1 (0.0000000000 1.0000000000) *
            1153) TimeBetweenRegistrationAndFirstInvestment< 1.5 5    1 0 (0.8000000000 0.2000000000)
              2306) Geography=France,Germany 4    0 0 (1.0000000000 0.0000000000) *
              2307) Geography=Spain 1    0 1 (0.0000000000 1.0000000000) *
           577) AverageOfCustomerFeedbackOnService< 4.225 366   16 0 (0.9562841530 0.0437158470)
            1154) CreditScore>=423 365   15 0 (0.9589041096 0.0410958904)
              2308) Age>=22.5 350   12 0 (0.9657142857 0.0342857143)
                4616) NumberOfComplaints< 5.5 330    9 0 (0.9727272727 0.0272727273)
                  9232) LastContactByABanker< 21.5 110    0 0 (1.0000000000 0.0000000000) *
                  9233) LastContactByABanker>=21.5 220    9 0 (0.9590909091 0.0409090909)
                    18466) LastContactByABanker>=22.5 216    7 0 (0.9675925926 0.0324074074)
                      36932) LastContactByABanker>=24.5 205    5 0 (0.9756097561 0.0243902439)
                        73864) Tenure>=0.5 198    4 0 (0.9797979798 0.0202020202)
                          147728) CreditScore>=611.5 115    0 0 (1.0000000000 0.0000000000) *
                          147729) CreditScore< 611.5 83    4 0 (0.9518072289 0.0481927711)
                            295458) CreditScore< 610 82    3 0 (0.9634146341 0.0365853659)
                              590916) LastContactByABanker>=34.5 61    0 0 (1.0000000000 0.0000000000) *
                              590917) LastContactByABanker< 34.5 21    3 0 (0.8571428571 0.1428571429)
                                1181834) TimeBetweenRegistrationAndFirstInvestment>=17.5 20    2 0 (0.9000000000 0.1000000000)
                                  2363668) LastContactByABanker< 31.5 14    0 0 (1.0000000000 0.0000000000) *
                                  2363669) LastContactByABanker>=31.5 6    2 0 (0.6666666667 0.3333333333)
                                    4727338) Balance< 85474.29 4    0 0 (1.0000000000 0.0000000000) *
                                    4727339) Balance>=85474.29 2    0 1 (0.0000000000 1.0000000000) *
                                1181835) TimeBetweenRegistrationAndFirstInvestment< 17.5 1    0 1 (0.0000000000 1.0000000000) *
                            295459) CreditScore>=610 1    0 1 (0.0000000000 1.0000000000) *
                        73865) Tenure< 0.5 7    1 0 (0.8571428571 0.1428571429)
                          147730) CreditScore< 751.5 6    0 0 (1.0000000000 0.0000000000) *
                          147731) CreditScore>=751.5 1    0 1 (0.0000000000 1.0000000000) *
                      36933) LastContactByABanker< 24.5 11    2 0 (0.8181818182 0.1818181818)
                        73866) Age< 36.5 7    0 0 (1.0000000000 0.0000000000) *
                        73867) Age>=36.5 4    2 0 (0.5000000000 0.5000000000)
                          147734) EstimatedSalary< 64867.6 2    0 0 (1.0000000000 0.0000000000) *
                          147735) EstimatedSalary>=64867.6 2    0 1 (0.0000000000 1.0000000000) *
                    18467) LastContactByABanker< 22.5 4    2 0 (0.5000000000 0.5000000000)
```

```
          36934) CreditScore< 622.2  0 0 (1.0000000000 0.0000000000) *
          36935) CreditScore>=622.2  0 1 (0.0000000000 1.0000000000) *
        4617) NumberOfComplaints>=5.5 20  3 0 (0.8500000000 0.1500000000)
          9234) TimeBetweenRegistrationAndFirstInvestment>=9.5 19  2 0 (0.8947368421 0.1052631579)
            18468) CreditScore< 787.17  1 0 (0.9411764706 0.0588235294)
              36936) Tenure< 8.5 13  0 0 (1.0000000000 0.0000000000) *
              36937) Tenure>=8.5 4  1 0 (0.7500000000 0.2500000000)
                73874) Tenure>=9.5 3  0 0 (1.0000000000 0.0000000000) *
                73875) Tenure< 9.5 1  0 1 (0.0000000000 1.0000000000) *
            18469) CreditScore>=787.2  1 0 (0.5000000000 0.5000000000)
              36938) CreditScore>=817.5 1  0 0 (1.0000000000 0.0000000000) *
              36939) CreditScore< 817.5 1  0 1 (0.0000000000 1.0000000000) *
          9235) TimeBetweenRegistrationAndFirstInvestment< 9.5 1  0 1 (0.0000000000 1.0000000000) *
      2309) Age< 22.5 15  3 0 (0.8000000000 0.2000000000)
        4618) NumOfProducts>=1.5 9  0 0 (1.0000000000 0.0000000000) *
        4619) NumOfProducts< 1.5 6  3 0 (0.5000000000 0.5000000000)
          9238) Tenure< 7.5 4  1 0 (0.7500000000 0.2500000000)
            18476) Balance< 135420.9 3  0 0 (1.0000000000 0.0000000000) *
            18477) Balance>=135420.9 1  0 1 (0.0000000000 1.0000000000) *
          9239) Tenure>=7.5 2  0 1 (0.0000000000 1.0000000000) *
    1155) CreditScore< 423 1  0 1 (0.0000000000 1.0000000000) *
  289) Balance>=218307.3 1  0 1 (0.0000000000 1.0000000000) *
145) Age>=42.5 262  27 0 (0.8969465649 0.1030534351)
  290) IsActiveMember=1 169  6 0 (0.9644970414 0.0355029586)
    580) Balance< 144157.8 151  3 0 (0.9801324503 0.0198675497)
      1160) TimeBetweenRegistrationAndFirstInvestment>=3.5 145  2 0 (0.9862068966 0.0137931034)
        2320) CreditScore>=480 136  1 0 (0.9926470588 0.0073529412)
          4640) EstimatedSalary>=25205.6 120  0 0 (1.0000000000 0.0000000000) *
          4641) EstimatedSalary< 25205.6 16  1 0 (0.9375000000 0.0625000000)
            9282) EstimatedSalary< 20528.69 15  0 0 (1.0000000000 0.0000000000) *
            9283) EstimatedSalary>=20528.69 1  0 1 (0.0000000000 1.0000000000) *
        2321) CreditScore< 480 9  1 0 (0.8888888889 0.1111111111)
          4642) CreditScore< 472 7  0 0 (1.0000000000 0.0000000000) *
          4643) CreditScore>=472 2  1 0 (0.5000000000 0.5000000000)
            9286) Geography=Germany 1  0 0 (1.0000000000 0.0000000000) *
            9287) Geography=France 1  0 1 (0.0000000000 1.0000000000) *
      1161) TimeBetweenRegistrationAndFirstInvestment< 3.5 6  1 0 (0.8333333333 0.1666666667)
        2322) Tenure>=3.5 5  0 0 (1.0000000000 0.0000000000) *
        2323) Tenure< 3.5 1  0 1 (0.0000000000 1.0000000000) *
    581) Balance>=144157.8 18  3 0 (0.8333333333 0.1666666667)
      1162) AverageOfCustomerFeedbackOnService>=4.075 16  1 0 (0.9375000000 0.0625000000)
        2324) Tenure>=1.5 13  0 0 (1.0000000000 0.0000000000) *
        2325) Tenure< 1.5 3  1 0 (0.6666666667 0.3333333333)
          4650) CreditScore>=740 2  0 0 (1.0000000000 0.0000000000) *
          4651) CreditScore< 740 1  0 1 (0.0000000000 1.0000000000) *
      1163) AverageOfCustomerFeedbackOnService< 4.075 2  0 1 (0.0000000000 1.0000000000) *
  291) IsActiveMember=0 93  21 0 (0.7741935484 0.2258064516)
    582) NumOfProducts>=1.5 45  1 0 (0.9777777778 0.0222222222)
      1164) Balance< 185167.44  0 0 (1.0000000000 0.0000000000) *
      1165) Balance>=185167 1  0 1 (0.0000000000 1.0000000000) *
    583) NumOfProducts< 1.5 48  20 0 (0.5833333333 0.4166666667)
      1166) AverageOfCustomerFeedbackOnService>=4.275 20  3 0 (0.8500000000 0.1500000000)
        2332) Age< 58.19  2 0 (0.8947368421 0.1052631579)
          4664) Tenure< 7.14  0 0 (1.0000000000 0.0000000000) *
          4665) Tenure>=7.5  2 0 (0.6000000000 0.4000000000)
            9330) CreditScore>=716.2  0 0 (1.0000000000 0.0000000000) *
            9331) CreditScore< 716.3  1 1 (0.3333333333 0.6666666667)
              18662) CreditScore< 622.5 1  0 0 (1.0000000000 0.0000000000) *
              18663) CreditScore>=622.5 2  0 1 (0.0000000000 1.0000000000) *
        2333) Age>=58 1  0 1 (0.0000000000 1.0000000000) *
      1167) AverageOfCustomerFeedbackOnService< 4.275 28  11 1 (0.3928571429 0.6071428571)
        2334) LastContactByABanker< 19.5 5  0 0 (1.0000000000 0.0000000000) *
        2335) LastContactByABanker>=19.5 23  6 1 (0.2608695652 0.7391304348)
          4670) EstimatedSalary>=121513.1 12  6 0 (0.5000000000 0.5000000000)
            9340) Balance>=72088.51 7  1 0 (0.8571428571 0.1428571429)
              18680) Age< 54.5 6  0 0 (1.0000000000 0.0000000000) *
              18681) Age>=54.5 1  0 1 (0.0000000000 1.0000000000) *
            9341) Balance< 72088.51 5  0 1 (0.0000000000 1.0000000000) *
          4671) EstimatedSalary< 121513.1 11  0 1 (0.0000000000 1.0000000000) *
73) AverageOfCustomerFeedbackOnService< 4.025 210  40 0 (0.8095238095 0.1904761905)
 146) Age< 44.5 166  18 0 (0.8915662651 0.1084337349)
  292) NumberOfComplaints< 5.5 157  14 0 (0.9108280255 0.0891719745)
    584) Balance< 122771.7 107  5 0 (0.9532710280 0.0467289720)
      1168) CreditScore< 765 97  3 0 (0.9690721649 0.0309278351)
        2336) PersonalAdvisor=1 67  0 0 (1.0000000000 0.0000000000) *
```

```
          2337) PersonalAdvisor=0 30    3 0 (0.9000000000 0.1000000000)
            4674) EstimatedSalary< 192509.8 29    2 0 (0.9310344828 0.0689655172)
              9348) TimeBetweenRegistrationAndFirstInvestment>=13 27    1 0 (0.9629629630 0.0370370370)
               18696) Tenure< 9.5 23    0 0 (1.0000000000 0.0000000000) *
               18697) Tenure>=9.5 4    1 0 (0.7500000000 0.2500000000)
                 37394) Age< 38 3    0 0 (1.0000000000 0.0000000000) *
                 37395) Age>=38 1    0 1 (0.0000000000 1.0000000000) *
              9349) TimeBetweenRegistrationAndFirstInvestment< 13 2    1 0 (0.5000000000 0.5000000000)
               18698) CreditScore< 661 1    0 0 (1.0000000000 0.0000000000) *
               18699) CreditScore>=661 1    0 1 (0.0000000000 1.0000000000) *
            4675) EstimatedSalary>=192509.8 1    0 1 (0.0000000000 1.0000000000) *
          1169) CreditScore>=765 10    2 0 (0.8000000000 0.2000000000)
            2338) CreditScore>=772 8    0 0 (1.0000000000 0.0000000000) *
            2339) CreditScore< 772 2    0 1 (0.0000000000 1.0000000000) *
         585) Balance>=122771.7 50    9 0 (0.8200000000 0.1800000000)
          1170) Balance>=129020.4 43    5 0 (0.8837209302 0.1162790698)
            2340) TimeBetweenRegistrationAndFirstInvestment< 100.5 42    4 0 (0.9047619048 0.0952380952)
              4680) Tenure< 5.5 23    0 0 (1.0000000000 0.0000000000) *
              4681) Tenure>=5.5 19    4 0 (0.7894736842 0.2105263158)
                9362) EstimatedSalary< 187863.1 18    3 0 (0.8333333333 0.1666666667)
                 18724) TimeBetweenRegistrationAndFirstInvestment>=43 11    0 0 (1.0000000000 0.0000000000) *
                 18725) TimeBetweenRegistrationAndFirstInvestment< 43 7    3 0 (0.5714285714 0.4285714286)
                   37450) Tenure< 8.5 5    1 0 (0.8000000000 0.2000000000)
                     74900) Age>=29 4    0 0 (1.0000000000 0.0000000000) *
                     74901) Age< 29 1    0 1 (0.0000000000 1.0000000000) *
                   37451) Tenure>=8.5 2    0 1 (0.0000000000 1.0000000000) *
                9363) EstimatedSalary>=187863.1 1    0 1 (0.0000000000 1.0000000000) *
            2341) TimeBetweenRegistrationAndFirstInvestment>=100.5 1    0 1 (0.0000000000 1.0000000000) *
          1171) Balance< 129020.4 7    3 1 (0.4285714286 0.5714285714)
            2342) EstimatedSalary>=87601.62 3    0 0 (1.0000000000 0.0000000000) *
            2343) EstimatedSalary< 87601.62 4    0 1 (0.0000000000 1.0000000000) *
       293) NumberOfComplaints>=5.5 9    4 0 (0.5555555556 0.4444444444)
         586) CreditScore>=687 5    0 0 (1.0000000000 0.0000000000) *
         587) CreditScore< 687 4    0 1 (0.0000000000 1.0000000000) *
      147) Age>=44.5 44   22 0 (0.5000000000 0.5000000000)
        294) LastContactByABanker< 21 13    0 0 (1.0000000000 0.0000000000) *
        295) LastContactByABanker>=21 31    9 1 (0.2903225806 0.7096774194)
         590) PersonalAdvisor=1 19    9 1 (0.4736842105 0.5263157895)
          1180) AverageOfCustomerFeedbackOnService>=3.975 11    3 0 (0.7272727273 0.2727272727)
            2360) CreditScore< 734.5 9    1 0 (0.8888888889 0.1111111111)
              4720) Geography=France,Germany 8    0 0 (1.0000000000 0.0000000000) *
              4721) Geography=Spain 1    0 1 (0.0000000000 1.0000000000) *
            2361) CreditScore>=734.5 2    0 1 (0.0000000000 1.0000000000) *
          1181) AverageOfCustomerFeedbackOnService< 3.975 8    1 1 (0.1250000000 0.8750000000)
            2362) CreditScore< 563 2    1 0 (0.5000000000 0.5000000000)
              4724) CreditScore>=539.5 1    0 0 (1.0000000000 0.0000000000) *
              4725) CreditScore< 539.5 1    0 1 (0.0000000000 1.0000000000) *
            2363) CreditScore>=563 6    0 1 (0.0000000000 1.0000000000) *
         591) PersonalAdvisor=0 12    0 1 (0.0000000000 1.0000000000) *
   37) NumOfProducts>=2.5 21    4 1 (0.1904761905 0.8095238095)
     74) LastContactByABanker< 28 3    0 0 (1.0000000000 0.0000000000) *
     75) LastContactByABanker>=28 18    1 1 (0.0555555556 0.9444444444)
      150) TimeBetweenRegistrationAndFirstInvestment< 12 1    0 0 (1.0000000000 0.0000000000) *
      151) TimeBetweenRegistrationAndFirstInvestment>=12 17    0 1 (0.0000000000 1.0000000000) *
  19) AverageOfCustomerFeedbackOnService< 3.925 215  103 1 (0.4790697674 0.5209302326)
   38) LastContactByABanker< 20.5 33    0 0 (1.0000000000 0.0000000000) *
   39) LastContactByABanker>=20.5 182   70 1 (0.3846153846 0.6153846154)
     78) Age< 39.5 71   27 0 (0.6197183099 0.3802816901)
      156) NumberOfComplaints< 3.5 59   17 0 (0.7118644068 0.2881355932)
        312) PersonalAdvisor=1 38    7 0 (0.8157894737 0.1842105263)
          624) TimeBetweenRegistrationAndFirstInvestment>=28.5 28    2 0 (0.9285714286 0.0714285714)
          1248) EstimatedSalary< 188093 26    1 0 (0.9615384615 0.0384615385)
            2496) LastContactByABanker< 44.5 19    0 0 (1.0000000000 0.0000000000) *
            2497) LastContactByABanker>=44.5 7    1 0 (0.8571428571 0.1428571429)
              4994) LastContactByABanker>=45.5 6    0 0 (1.0000000000 0.0000000000) *
              4995) LastContactByABanker< 45.5 1    0 1 (0.0000000000 1.0000000000) *
          1249) EstimatedSalary>=188093 2    1 0 (0.5000000000 0.5000000000)
            2498) CreditScore< 573.5 1    0 0 (1.0000000000 0.0000000000) *
            2499) CreditScore>=573.5 1    0 1 (0.0000000000 1.0000000000) *
          625) TimeBetweenRegistrationAndFirstInvestment< 28.5 10    5 0 (0.5000000000 0.5000000000)
          1250) Age< 35.5 7    2 0 (0.7142857143 0.2857142857)
            2500) CreditScore< 644.5 4    0 0 (1.0000000000 0.0000000000) *
            2501) CreditScore>=644.5 3    1 1 (0.3333333333 0.6666666667)
              5002) CreditScore>=710 1    0 0 (1.0000000000 0.0000000000) *
              5003) CreditScore< 710 2    0 1 (0.0000000000 1.0000000000) *
```

```
          1251) Age>=35.5 3    0 1 (0.0000000000 1.0000000000) *
        313) PersonalAdvisor=0 21   10 0 (0.5238095238 0.4761904762)
          626) HasCrCard=0 5    0 0 (1.0000000000 0.0000000000) *
          627) HasCrCard=1 16    6 1 (0.3750000000 0.6250000000)
            1254) EstimatedSalary< 36579.29 3    0 0 (1.0000000000 0.0000000000) *
            1255) EstimatedSalary>=36579.29 13    3 1 (0.2307692308 0.7692307692)
              2510) EstimatedSalary>=147539 2    0 0 (1.0000000000 0.0000000000) *
              2511) EstimatedSalary< 147539 11    1 1 (0.0909090909 0.9090909091)
                5022) CreditScore< 514.5 1    0 0 (1.0000000000 0.0000000000) *
                5023) CreditScore>=514.5 10    0 1 (0.0000000000 1.0000000000) *
      157) NumberOfComplaints>=3.5 12    2 1 (0.1666666667 0.8333333333)
        314) Tenure< 1.5 3    1 0 (0.6666666667 0.3333333333)
          628) CreditScore< 691.5 2    0 0 (1.0000000000 0.0000000000) *
          629) CreditScore>=691.5 1    0 1 (0.0000000000 1.0000000000) *
        315) Tenure>=1.5 9    0 1 (0.0000000000 1.0000000000) *
    79) Age>=39.5 111   26 1 (0.2342342342 0.7657657658)
      158) FrequencyOfContact< 2.5 23   11 1 (0.4782608696 0.5217391304)
        316) LastContactByABanker>=29.5 14    4 0 (0.7142857143 0.2857142857)
          632) Tenure>=3.5 10    1 0 (0.9000000000 0.1000000000)
            1264) CreditScore< 737.5 9    0 0 (1.0000000000 0.0000000000) *
            1265) CreditScore>=737.5 1    0 1 (0.0000000000 1.0000000000) *
          633) Tenure< 3.5 4    1 1 (0.2500000000 0.7500000000)
            1266) Geography=France 1    0 0 (1.0000000000 0.0000000000) *
            1267) Geography=Germany,Spain 3    0 1 (0.0000000000 1.0000000000) *
        317) LastContactByABanker< 29.5 9    1 1 (0.1111111111 0.8888888889)
          634) NumOfProducts>=1.5 1    0 0 (1.0000000000 0.0000000000) *
          635) NumOfProducts< 1.5 8    0 1 (0.0000000000 1.0000000000) *
      159) FrequencyOfContact>=2.5 88   15 1 (0.1704545455 0.8295454545)
        318) IsActiveMember=1 36   11 1 (0.3055555556 0.6944444444)
          636) FrequencyOfContact>=3.5 14    6 0 (0.5714285714 0.4285714286)
            1272) Geography=France,Spain 11    3 0 (0.7272727273 0.2727272727)
              2544) FinancialLiteracy=0,2 7    0 0 (1.0000000000 0.0000000000) *
              2545) FinancialLiteracy=1 4    1 1 (0.2500000000 0.7500000000)
                5090) Gender=Female 1    0 0 (1.0000000000 0.0000000000) *
                5091) Gender=Male 3    0 1 (0.0000000000 1.0000000000) *
            1273) Geography=Germany 3    0 1 (0.0000000000 1.0000000000) *
          637) FrequencyOfContact< 3.5 22    3 1 (0.1363636364 0.8636363636)
            1274) NumOfProducts>=1.5 5    2 1 (0.4000000000 0.6000000000)
              2548) CreditScore< 624.5 2    0 0 (1.0000000000 0.0000000000) *
              2549) CreditScore>=624.5 3    0 1 (0.0000000000 1.0000000000) *
            1275) NumOfProducts< 1.5 17    1 1 (0.0588235294 0.9411764706)
              2550) Age< 41 2    1 0 (0.5000000000 0.5000000000)
                5100) CreditScore>=692.5 1    0 0 (1.0000000000 0.0000000000) *
                5101) CreditScore< 692.5 1    0 1 (0.0000000000 1.0000000000) *
              2551) Age>=41 15    0 1 (0.0000000000 1.0000000000) *
        319) IsActiveMember=0 52    4 1 (0.0769230769 0.9230769231)
          638) Age< 40.5 2    1 0 (0.5000000000 0.5000000000)
            1276) CreditScore>=605.5 1    0 0 (1.0000000000 0.0000000000) *
            1277) CreditScore< 605.5 1    0 1 (0.0000000000 1.0000000000) *
          639) Age>=40.5 50    3 1 (0.0600000000 0.9400000000)
            1278) TimeBetweenRegistrationAndFirstInvestment>=92 9    2 1 (0.2222222222 0.7777777778)
              2556) Age< 48.5 3    1 0 (0.6666666667 0.3333333333)
                5112) CreditScore>=580 2    0 0 (1.0000000000 0.0000000000) *
                5113) CreditScore< 580 1    0 1 (0.0000000000 1.0000000000) *
              2557) Age>=48.5 6    0 1 (0.0000000000 1.0000000000) *
            1279) TimeBetweenRegistrationAndFirstInvestment< 92 41    1 1 (0.0243902439 0.9756097561)
              2558) TimeBetweenRegistrationAndFirstInvestment< 10 4    1 1 (0.2500000000 0.7500000000)
                5116) CreditScore< 580.5 1    0 0 (1.0000000000 0.0000000000) *
                5117) CreditScore>=580.5 3    0 1 (0.0000000000 1.0000000000) *
              2559) TimeBetweenRegistrationAndFirstInvestment>=10 37    0 1 (0.0000000000 1.0000000000) *
  5) LastContactByABanker>=60.5 271    0 1 (0.0000000000 1.0000000000) *
3) AverageOfCustomerFeedbackOnService< 3.775 1795  304 1 (0.1693593315 0.8306406685)
  6) LastContactByABanker< 20.5 96    0 0 (1.0000000000 0.0000000000) *
  7) LastContactByABanker>=20.5 1699  208 1 (0.1224249559 0.8775750441)
    14) AverageOfCustomerFeedbackOnService>=3.575 622  180 1 (0.2893890675 0.7106109325)
      28) FrequencyOfContact>=4.5 240   96 0 (0.6000000000 0.4000000000)
        56) LastContactByABanker< 60.5 185   41 0 (0.7783783784 0.2216216216)
        112) PersonalAdvisor=1 131   15 0 (0.8854961832 0.1145038168)
          224) NumOfProducts< 2.5 129   13 0 (0.8992248062 0.1007751938)
            448) Geography=France,Spain 97    5 0 (0.9484536082 0.0515463918)
              896) EstimatedSalary< 147582.5 70    1 0 (0.9857142857 0.0142857143)
              1792) NumberOfComplaints< 5 65    0 0 (1.0000000000 0.0000000000) *
              1793) NumberOfComplaints>=5 5    1 0 (0.8000000000 0.2000000000)
                3586) CreditScore>=535 4    0 0 (1.0000000000 0.0000000000) *
                3587) CreditScore< 535 1    0 1 (0.0000000000 1.0000000000) *
```

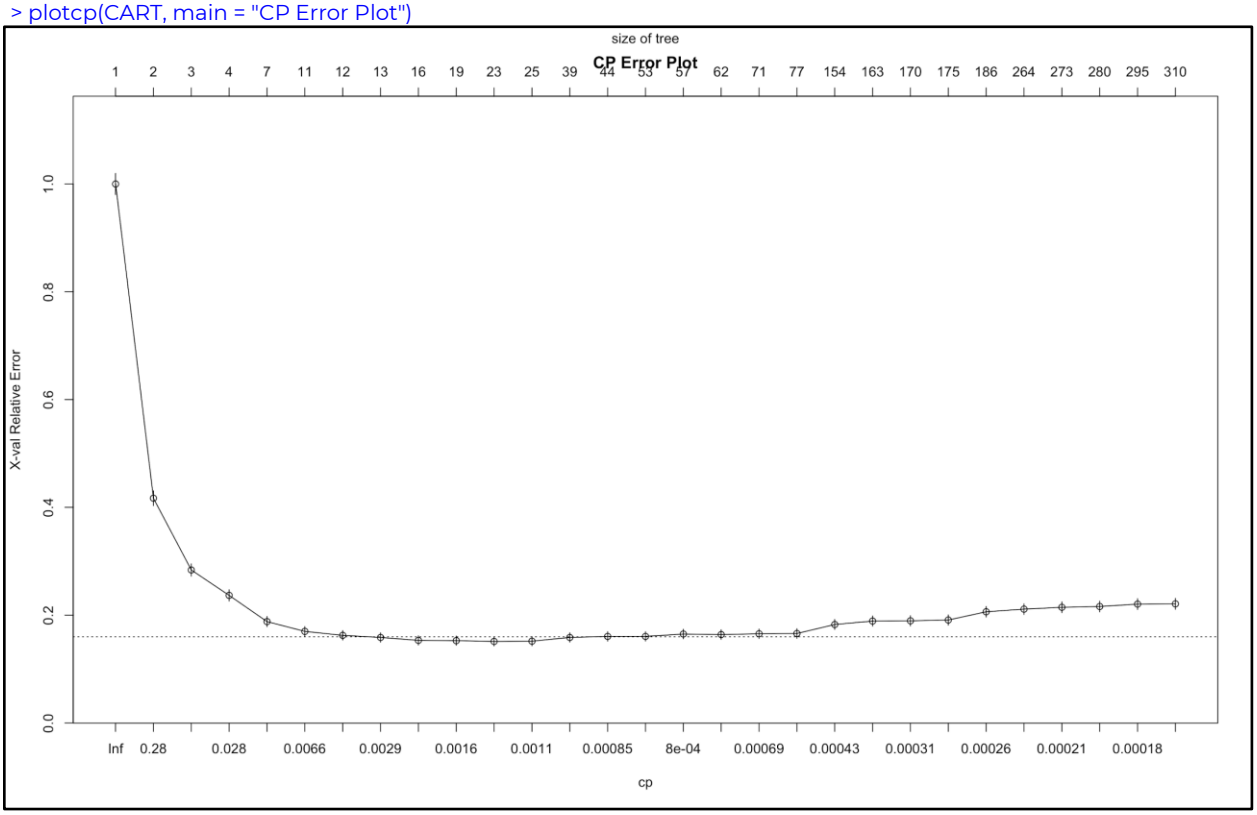
```
        897) EstimatedSalary>=147582.5 27    4 0 (0.8518518519 0.1481481481)
        1794) LastContactByABanker< 57.5 24    2 0 (0.9166666667 0.0833333333)
          3588) EstimatedSalary>=149057.8 23    1 0 (0.9565217391 0.0434782609)
            7176) LastContactByABanker>=24.5 21    0 0 (1.0000000000 0.0000000000) *
            7177) LastContactByABanker< 24.5 2    1 0 (0.5000000000 0.5000000000)
              14354) CreditScore< 674 1    0 0 (1.0000000000 0.0000000000) *
              14355) CreditScore>=674 1    0 1 (0.0000000000 1.0000000000) *
          3589) EstimatedSalary< 149057.8 1    0 1 (0.0000000000 1.0000000000) *
        1795) LastContactByABanker>=57.5 3    1 1 (0.3333333333 0.6666666667)
          3590) Geography=France 1    0 0 (1.0000000000 0.0000000000) *
          3591) Geography=Spain 2    0 1 (0.0000000000 1.0000000000) *
      449) Geography=Germany 32    8 0 (0.7500000000 0.2500000000)
        898) Age< 35.5 13    0 0 (1.0000000000 0.0000000000) *
        899) Age>=35.5 19    8 0 (0.5789473684 0.4210526316)
        1798) LastContactByABanker< 47.5 16    5 0 (0.6875000000 0.3125000000)
          3596) FrequencyOfContact< 8.5 12    2 0 (0.8333333333 0.1666666667)
            7192) EstimatedSalary< 188521.2 11    1 0 (0.9090909091 0.0909090909)
              14384) FinancialLiteracy=1,2 10    0 0 (1.0000000000 0.0000000000) *
              14385) FinancialLiteracy=0 1    0 1 (0.0000000000 1.0000000000) *
            7193) EstimatedSalary>=188521.2 1    0 1 (0.0000000000 1.0000000000) *
          3597) FrequencyOfContact>=8.5 4    1 1 (0.2500000000 0.7500000000)
            7194) CreditScore< 548 1    0 0 (1.0000000000 0.0000000000) *
            7195) CreditScore>=548 3    0 1 (0.0000000000 1.0000000000) *
        1799) LastContactByABanker>=47.5 3    0 1 (0.0000000000 1.0000000000) *
    225) NumOfProducts>=2.5 2    0 1 (0.0000000000 1.0000000000) *
  113) PersonalAdvisor=0 54   26 0 (0.5185185185 0.4814814815)
    226) EstimatedSalary< 35231.68 6    0 0 (1.0000000000 0.0000000000) *
    227) EstimatedSalary>=35231.68 48   22 1 (0.4583333333 0.5416666667)
      454) TimeBetweenRegistrationAndFirstInvestment>=65.5 18    6 0 (0.6666666667 0.3333333333)
        908) AverageOfCustomerFeedbackOnService>=3.675 13    2 0 (0.8461538462 0.1538461538)
        1816) Age< 45.5 9    0 0 (1.0000000000 0.0000000000) *
        1817) Age>=45.5 4    2 0 (0.5000000000 0.5000000000)
          3634) CreditScore>=671 2    0 0 (1.0000000000 0.0000000000) *
          3635) CreditScore< 671 2    0 1 (0.0000000000 1.0000000000) *
        909) AverageOfCustomerFeedbackOnService< 3.675 5    1 1 (0.2000000000 0.8000000000)
        1818) Tenure>=5 1    0 0 (1.0000000000 0.0000000000) *
        1819) Tenure< 5 4    0 1 (0.0000000000 1.0000000000) *
      455) TimeBetweenRegistrationAndFirstInvestment< 65.5 30   10 1 (0.3333333333 0.6666666667)
        910) Age< 28 2    0 0 (1.0000000000 0.0000000000) *
        911) Age>=28 28    8 1 (0.2857142857 0.7142857143)
        1822) IsActiveMember=1 12    6 0 (0.5000000000 0.5000000000)
          3644) CreditScore>=620 8    2 0 (0.7500000000 0.2500000000)
            7288) Tenure< 4.5 5    0 0 (1.0000000000 0.0000000000) *
            7289) Tenure>=4.5 3    1 1 (0.3333333333 0.6666666667)
              14578) CreditScore< 691 1    0 0 (1.0000000000 0.0000000000) *
              14579) CreditScore>=691 2    0 1 (0.0000000000 1.0000000000) *
          3645) CreditScore< 620 4    0 1 (0.0000000000 1.0000000000) *
        1823) IsActiveMember=0 16    2 1 (0.1250000000 0.8750000000)
          3646) TimeBetweenRegistrationAndFirstInvestment< 8.5 2    0 0 (1.0000000000 0.0000000000) *
          3647) TimeBetweenRegistrationAndFirstInvestment>=8.5 14    0 1 (0.0000000000 1.0000000000) *
 57) LastContactByABanker>=60.5 55    0 1 (0.0000000000 1.0000000000) *
29) FrequencyOfContact< 4.5 382   36 1 (0.0942408377 0.9057591623)
  58) LastContactByABanker< 60.5 202   36 1 (0.1782178218 0.8217821782)
  116) Age< 43.5 90   29 1 (0.3222222222 0.6777777778)
    232) PersonalAdvisor=1 48   22 1 (0.4583333333 0.5416666667)
      464) Tenure>=2.5 27   10 0 (0.6296296296 0.3703703704)
        928) TimeBetweenRegistrationAndFirstInvestment>=11 25    8 0 (0.6800000000 0.3200000000)
        1856) Tenure< 3.5 7    0 0 (1.0000000000 0.0000000000) *
        1857) Tenure>=3.5 18    8 0 (0.5555555556 0.4444444444)
          3714) CreditScore>=759 4    0 0 (1.0000000000 0.0000000000) *
          3715) CreditScore< 759 14    6 1 (0.4285714286 0.5714285714)
            7430) LastContactByABanker>=52.5 2    0 0 (1.0000000000 0.0000000000) *
            7431) LastContactByABanker< 52.5 12    4 1 (0.3333333333 0.6666666667)
              14862) Age< 32 6    2 0 (0.6666666667 0.3333333333)
                29724) Tenure< 4.5 4    0 0 (1.0000000000 0.0000000000) *
                29725) Tenure>=4.5 2    0 1 (0.0000000000 1.0000000000) *
              14863) Age>=32 6    0 1 (0.0000000000 1.0000000000) *
        929) TimeBetweenRegistrationAndFirstInvestment< 11 2    0 1 (0.0000000000 1.0000000000) *
      465) Tenure< 2.5 21    5 1 (0.2380952381 0.7619047619)
        930) CreditScore>=606 12    5 1 (0.4166666667 0.5833333333)
        1860) NumOfProducts< 2.5 8    3 0 (0.6250000000 0.3750000000)
          3720) FrequencyOfContact>=3.5 4    0 0 (1.0000000000 0.0000000000) *
          3721) FrequencyOfContact< 3.5 4    1 1 (0.2500000000 0.7500000000)
            7442) EstimatedSalary< 59991.96 1    0 0 (1.0000000000 0.0000000000) *
            7443) EstimatedSalary>=59991.96 3    0 1 (0.0000000000 1.0000000000) *
```

```
          1861) NumOfProducts>=2.5 4    0 1 (0.0000000000 1.0000000000) *
         931) CreditScore< 606 9    0 1 (0.0000000000 1.0000000000) *
       233) PersonalAdvisor=0 42    7 1 (0.1666666667 0.8333333333)
         466) CreditScore< 475 1    0 0 (1.0000000000 0.0000000000) *
         467) CreditScore>=475 41    6 1 (0.1463414634 0.8536585366)
         934) Age>=37.5 25    6 1 (0.2400000000 0.7600000000)
          1868) TimeBetweenRegistrationAndFirstInvestment< 45.5 11    5 1 (0.4545454545 0.5454545455)
           3736) LastContactByABanker>=43 3    0 0 (1.0000000000 0.0000000000) *
           3737) LastContactByABanker< 43 8    2 1 (0.2500000000 0.7500000000)
            7474) CreditScore>=848.5 1    0 0 (1.0000000000 0.0000000000) *
            7475) CreditScore< 848.5 7    1 1 (0.1428571429 0.8571428571)
             14950) Age>=42.5 2    1 0 (0.5000000000 0.5000000000)
               29900) CreditScore>=718.5 1    0 0 (1.0000000000 0.0000000000) *
               29901) CreditScore< 718.5 1    0 1 (0.0000000000 1.0000000000) *
             14951) Age< 42.5 5    0 1 (0.0000000000 1.0000000000) *
          1869) TimeBetweenRegistrationAndFirstInvestment>=45.5 14    1 1 (0.0714285714 0.9285714286)
           3738) Tenure< 1.5 2    1 0 (0.5000000000 0.5000000000)
            7476) CreditScore< 738.5 1    0 0 (1.0000000000 0.0000000000) *
            7477) CreditScore>=738.5 1    0 1 (0.0000000000 1.0000000000) *
           3739) Tenure>=1.5 12    0 1 (0.0000000000 1.0000000000) *
         935) Age< 37.5 16    0 1 (0.0000000000 1.0000000000) *
     117) Age>=43.5 112    7 1 (0.0625000000 0.9375000000)
       234) Balance< 86650.97 37    6 1 (0.1621621622 0.8378378378)
         468) CreditScore< 463 2    0 0 (1.0000000000 0.0000000000) *
         469) CreditScore>=463 35    4 1 (0.1142857143 0.8857142857)
         938) LastContactByABanker< 22.5 4    2 0 (0.5000000000 0.5000000000)
          1876) CreditScore>=544.5 2    0 0 (1.0000000000 0.0000000000) *
          1877) CreditScore< 544.5 2    0 1 (0.0000000000 1.0000000000) *
         939) LastContactByABanker>=22.5 31    2 1 (0.0645161290 0.9354838710)
          1878) LastContactByABanker>=57.5 5    2 1 (0.4000000000 0.6000000000)
           3756) EstimatedSalary>=148766.5 2    0 0 (1.0000000000 0.0000000000) *
           3757) EstimatedSalary< 148766.5 3    0 1 (0.0000000000 1.0000000000) *
          1879) LastContactByABanker< 57.5 26    0 1 (0.0000000000 1.0000000000) *
       235) Balance>=86650.97 75    1 1 (0.0133333333 0.9866666667)
         470) Geography=Spain 15    1 1 (0.0666666667 0.9333333333)
         940) Balance< 111186 4    1 1 (0.2500000000 0.7500000000)
          1880) Balance>=110280 1    0 0 (1.0000000000 0.0000000000) *
          1881) Balance< 110280 3    0 1 (0.0000000000 1.0000000000) *
         941) Balance>=111186 11    0 1 (0.0000000000 1.0000000000) *
         471) Geography=France,Germany 60    0 1 (0.0000000000 1.0000000000) *
   59) LastContactByABanker>=60.5 180    0 1 (0.0000000000 1.0000000000) *
 15) AverageOfCustomerFeedbackOnService< 3.575 1077   28 1 (0.0259981430 0.9740018570)
   30) AverageOfCustomerFeedbackOnService>=3.425 370   26 1 (0.0702702703 0.9297297297)
     60) LastContactByABanker< 59.5 212   26 1 (0.1226415094 0.8773584906)
      120) Age< 38.5 66   16 1 (0.2424242424 0.7575757576)
        240) LastContactByABanker>=29.5 49   16 1 (0.3265306122 0.6734693878)
          480) EstimatedSalary< 6924.565 3    0 0 (1.0000000000 0.0000000000) *
          481) EstimatedSalary>=6924.565 46   13 1 (0.2826086957 0.7173913043)
           962) IsActiveMember=1 21    9 1 (0.4285714286 0.5714285714)
            1924) CreditScore>=619.5 13    5 0 (0.6153846154 0.3846153846)
              3848) CreditScore< 702 7    1 0 (0.8571428571 0.1428571429)
                7696) Geography=France,Germany 6    0 0 (1.0000000000 0.0000000000) *
                7697) Geography=Spain 1    0 1 (0.0000000000 1.0000000000) *
              3849) CreditScore>=702 6    2 1 (0.3333333333 0.6666666667)
                7698) CreditScore>=752.5 3    1 0 (0.6666666667 0.3333333333)
                 15396) Age< 36.5 2    0 0 (1.0000000000 0.0000000000) *
                 15397) Age>=36.5 1    0 1 (0.0000000000 1.0000000000) *
                7699) CreditScore< 752.5 3    0 1 (0.0000000000 1.0000000000) *
            1925) CreditScore< 619.5 8    1 1 (0.1250000000 0.8750000000)
              3850) TimeBetweenRegistrationAndFirstInvestment>=91.5 1    0 0 (1.0000000000 0.0000000000) *
              3851) TimeBetweenRegistrationAndFirstInvestment< 91.5 7    0 1 (0.0000000000 1.0000000000) *
           963) IsActiveMember=0 25    4 1 (0.1600000000 0.8400000000)
            1926) TimeBetweenRegistrationAndFirstInvestment< 23 6    3 0 (0.5000000000 0.5000000000)
              3852) LastContactByABanker>=39.5 3    0 0 (1.0000000000 0.0000000000) *
              3853) LastContactByABanker< 39.5 3    0 1 (0.0000000000 1.0000000000) *
            1927) TimeBetweenRegistrationAndFirstInvestment>=23 19    1 1 (0.0526315789 0.9473684211)
              3854) Age< 22.5 1    0 0 (1.0000000000 0.0000000000) *
              3855) Age>=22.5 18    0 1 (0.0000000000 1.0000000000) *
        241) LastContactByABanker< 29.5 17    0 1 (0.0000000000 1.0000000000) *
      121) Age>=38.5 146   10 1 (0.0684931507 0.9315068493)
        242) Balance>=144717.6 18    6 1 (0.3333333333 0.6666666667)
          484) FrequencyOfContact>=4.5 10    4 0 (0.6000000000 0.4000000000)
            968) Gender=Male 4    0 0 (1.0000000000 0.0000000000) *
            969) Gender=Female 6    2 1 (0.3333333333 0.6666666667)
             1938) TimeBetweenRegistrationAndFirstInvestment>=68 2    0 0 (1.0000000000 0.0000000000) *
```

```
      1939) TimeBetweenRegistrationAndFirstInvestment< 68 4    0 1 (0.0000000000 1.0000000000) *
    485) FrequencyOfContact< 4.5 8    0 1 (0.0000000000 1.0000000000) *
  243) Balance< 144717.6 128    4 1 (0.0312500000 0.9687500000)
    486) Geography=Spain 26    3 1 (0.1153846154 0.8846153846)
      972) FrequencyOfContact>=4.5 8    3 1 (0.3750000000 0.6250000000)
      1944) FinancialLiteracy=1 4    1 0 (0.7500000000 0.2500000000)
        3888) Age< 59 3    0 0 (1.0000000000 0.0000000000) *
        3889) Age>=59 1    0 1 (0.0000000000 1.0000000000) *
      1945) FinancialLiteracy=2 4    0 1 (0.0000000000 1.0000000000) *
      973) FrequencyOfContact< 4.5 18    0 1 (0.0000000000 1.0000000000) *
    487) Geography=France,Germany 102    1 1 (0.0098039216 0.9901960784)
      974) Tenure< 0.5 5    1 1 (0.2000000000 0.8000000000)
      1948) CreditScore>=705 1    0 0 (1.0000000000 0.0000000000) *
      1949) CreditScore< 705 4    0 1 (0.0000000000 1.0000000000) *
      975) Tenure>=0.5 97    0 1 (0.0000000000 1.0000000000) *
  61) LastContactByABanker>=59.5 158    0 1 (0.0000000000 1.0000000000) *
31) AverageOfCustomerFeedbackOnService< 3.425 707    2 1 (0.0028288543 0.9971711457)
  62) Age>=73.5 2    1 0 (0.5000000000 0.5000000000)
  124) CreditScore>=583.5 1    0 0 (1.0000000000 0.0000000000) *
  125) CreditScore< 583.5 1    0 1 (0.0000000000 1.0000000000) *
  63) Age< 73.5 705    1 1 (0.0014184397 0.9985815603)
  126) Age< 23.5 12    1 1 (0.0833333333 0.9166666667)
    252) Age>=22.5 1    0 0 (1.0000000000 0.0000000000) *
    253) Age< 22.5 11    0 1 (0.0000000000 1.0000000000) *
  127) Age>=23.5 693    0 1 (0.0000000000 1.0000000000) *
```

Now we will need to prune the CART to its minimum to prevent overfitting. In order to evaluate the CP value in which the CART model will be pruned at, we need to find the cross validation (CV) error cap and the simplest node that is within the CV error cap (i.e. the nearest node under the CV error cap). To do this, we can first plot and print out the various CV errors and CP values at each node of our CART model.

```
> plotcp(CART, main = "CP Error Plot")
```




```
> printcp(CART)
```

```
Classification tree:
rpart(formula = Exited ~ ., data = pData, method = "class", control = rpart.control(minsplit = 2,
    cp = 0))

Variables actually used in tree construction:
 [1] Age                                  AverageOfCustomerFeedbackOnService
 [3] Balance                              CreditScore
 [5] EstimatedSalary                      FinancialLiteracy
 [7] FrequencyOfContact                   Gender
 [9] Geography                            HasCrCard
[11] IsActiveMember                       LastContactByABanker
[13] NumberOfComplaints                   NumOfProducts
[15] PersonalAdvisor                      Tenure
[17] TimeBetweenRegistrationAndFirstInvestment UnresolvedComplaint

Root node error: 2036/9997 = 0.20366

n= 9997

          CP nsplit rel error  xerror      xstd
1  0.58300589      0 1.0000000 1.00000 0.0197770
2  0.13310413      1 0.4169941 0.41699 0.0136900
3  0.04715128      2 0.2838900 0.28389 0.0114618
4  0.01686313      3 0.2367387 0.23674 0.0105200
5  0.00687623      6 0.1861493 0.18811 0.0094262
6  0.00638507     10 0.1571709 0.16994 0.0089766
7  0.00392927     11 0.1507859 0.16257 0.0087867
8  0.00212836     12 0.1468566 0.15864 0.0086834
9  0.00163720     15 0.1404715 0.15324 0.0085391
10 0.00147348     18 0.1355599 0.15275 0.0085259
11 0.00122790     22 0.1296660 0.15128 0.0084860
12 0.00098232     24 0.1272102 0.15177 0.0084993
13 0.00085953     38 0.1129666 0.15864 0.0086834
14 0.00084199     43 0.1085462 0.16061 0.0087352
15 0.00081860     52 0.1006876 0.16061 0.0087352
16 0.00078585     56 0.0972495 0.16503 0.0088505
17 0.00073674     61 0.0933202 0.16405 0.0088250
18 0.00065488     70 0.0839882 0.16552 0.0088632
19 0.00049116     76 0.0800589 0.16601 0.0088759
20 0.00036837    153 0.0402750 0.18271 0.0092952
21 0.00032744    162 0.0368369 0.18910 0.0094498
22 0.00029470    169 0.0343811 0.18959 0.0094616
23 0.00027287    174 0.0329077 0.19106 0.0094968
24 0.00024558    185 0.0294695 0.20629 0.0098521
25 0.00021829    263 0.0088409 0.21120 0.0099634
26 0.00021050    272 0.0068762 0.21464 0.0100405
27 0.00019646    279 0.0054028 0.21611 0.0100734
28 0.00016372    294 0.0024558 0.22053 0.0101711
29 0.00000000    309 0.0000000 0.22102 0.0101819
```

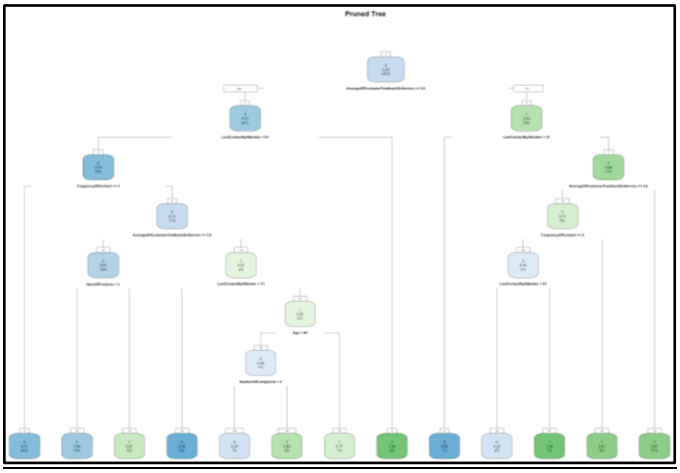
Next we can identify the node which has the lowest CV error and calculate the CV error cap by adding its CV error along with its error standard deviation. After doing so, we can find the CP value of the first node whose CV error's value is under the CV error cap.

```
## Deduce CP value - Manual Method
> CVerror.cap.M <- 0.15177 + 0.0084993
> CVerror.cap.M
[1] 0.1602693
> cp.M <- sqrt(0.00212836  * 0.00392927)
> cp.M
[1] 0.002891868
```

```
## Deduce CP value - Automated Method
> CVerror.cap <- CART$cptable[which.min(CART$cptable[,"xerror"]), "xerror"] +
CART$cptable[which.min(CART$cptable[,"xerror"]), "xstd"]
> i <- 1; j<- 4
> while (CART$cptable[i,j] > CVerror.cap) {i <- i + 1}
> cp = ifelse(i > 1, sqrt(CART$cptable[i,1] * CART$cptable[i-1,1]), 1)
> cp
[1] 0.002891867
```

After identifying the CP value, we can prune the CART model to its minimum:

```
## Pruning the Tree
> pCART <- prune(CART, cp = cp)
> rpart.plot(pCART, nn= T, main = "Pruned Tree")
```




```
> print(pCART)
n= 9997

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 9997 2036 0 (0.796338902 0.203661098)
   2) AverageOfCustomerFeedbackOnService>=3.775 8202  545 0 (0.933552792 0.066447208)
     4) LastContactByABanker< 60.5 7931  274 0 (0.965452024 0.034547976)
       8) FrequencyOfContact>=4.5 6230   56 0 (0.991011236 0.008988764) *
       9) FrequencyOfContact< 4.5 1701  218 0 (0.871840094 0.128159906)
        18) AverageOfCustomerFeedbackOnService>=3.925 1486  106 0 (0.928667564 0.071332436)
          36) NumOfProducts< 2.5 1465   89 0 (0.939249147 0.060750853) *
          37) NumOfProducts>=2.5 21    4 1 (0.190476190 0.809523810) *
        19) AverageOfCustomerFeedbackOnService< 3.925 215  103 1 (0.479069767 0.520930233)
          38) LastContactByABanker< 20.5 33    0 0 (1.000000000 0.000000000) *
          39) LastContactByABanker>=20.5 182   70 1 (0.384615385 0.615384615)
            78) Age< 39.5 71   27 0 (0.619718310 0.380281690)
             156) NumberOfComplaints< 3.5 59   17 0 (0.711864407 0.288135593) *
             157) NumberOfComplaints>=3.5 12    2 1 (0.166666667 0.833333333) *
            79) Age>=39.5 111   26 1 (0.234234234 0.765765766) *
     5) LastContactByABanker>=60.5 271    0 1 (0.000000000 1.000000000) *
   3) AverageOfCustomerFeedbackOnService< 3.775 1795  304 1 (0.169359331 0.830640669)
     6) LastContactByABanker< 20.5 96    0 0 (1.000000000 0.000000000) *
     7) LastContactByABanker>=20.5 1699  208 1 (0.122424956 0.877575044)
      14) AverageOfCustomerFeedbackOnService>=3.575 622  180 1 (0.289389068 0.710610932)
        28) FrequencyOfContact>=4.5 240   96 0 (0.600000000 0.400000000)
          56) LastContactByABanker< 60.5 185   41 0 (0.778378378 0.221621622) *
          57) LastContactByABanker>=60.5 55    0 1 (0.000000000 1.000000000) *
```

```
      29) FrequencyOfContact< 4.5 382   36 1 (0.094240838 0.905759162) *
     15) AverageOfCustomerFeedbackOnService< 3.575 1077   28 1 (0.025998143 0.974001857) *

> printcp(pCART)
Classification tree:
rpart(formula = Exited ~ ., data = pData, method = "class", control = rpart.control(minsplit = 2,
    cp = 0))

Variables actually used in tree construction:
[1] Age                          AverageOfCustomerFeedbackOnService FrequencyOfContact
[4] LastContactByABanker         NumberOfComplaints                 NumOfProducts

Root node error: 2036/9997 = 0.20366

n= 9997

        CP nsplit rel error  xerror     xstd
1 0.5830059      0   1.00000 1.00000 0.0197770
2 0.1331041      1   0.41699 0.41699 0.0136900
3 0.0471513      2   0.28389 0.28389 0.0114618
4 0.0168631      3   0.23674 0.23674 0.0105200
5 0.0068762      6   0.18615 0.18811 0.0094262
6 0.0063851     10   0.15717 0.16994 0.0089766
7 0.0039293     11   0.15079 0.16257 0.0087867
8 0.0028919     12   0.14686 0.15864 0.0086834
```

After pruning our CART model, we can finally use it to predict whether a particular customer would exit and stop using White Rock's services at the end of the financial year and calculate its prediction accuracy to help evaluate its credibility as a model.

```
## Predicting with Pruned CART Model
> cart.predict <- predict(pCART, newdata = pData, type = "class")
> results <- data.frame(pData, cart.predict)

## Confusion Matrix & Accuracy of Model
> table(results$cart.predict, pData$Exited, deparse.level = 2)
              results$cart.predict
pData$Exited    0    1
         0   7865   96
         1    203 1833
> mean(results$cart.predict == dt$Exited)
[1] 0.970091
```

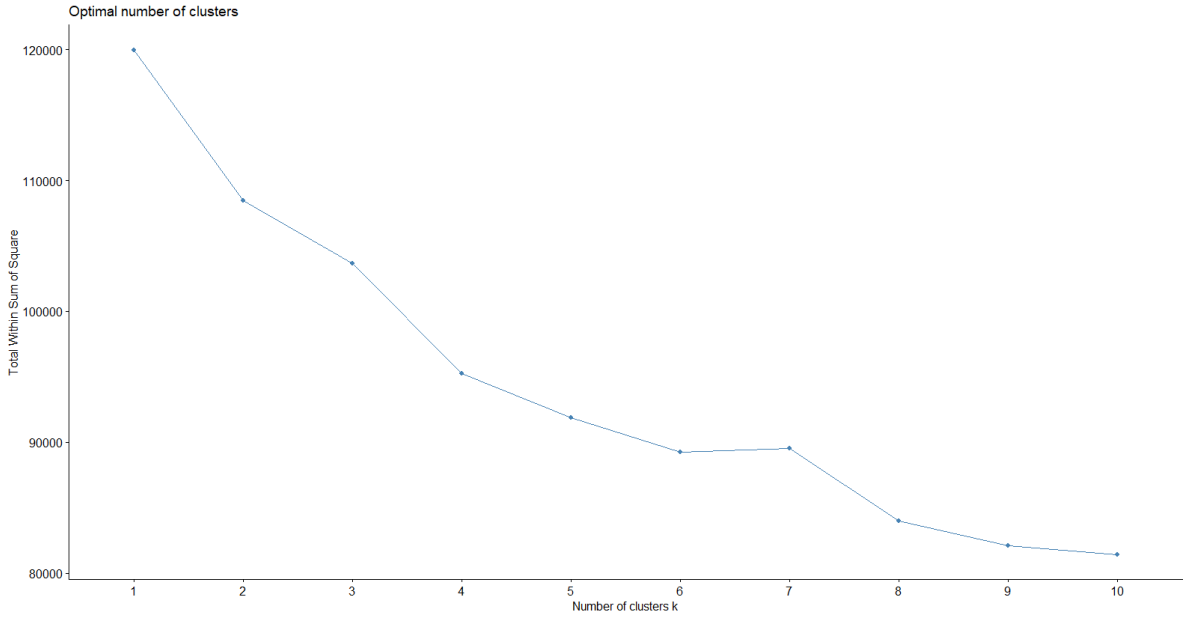
APPENDIX D - CLUSTERING

Binary and nominal variables are first removed from data before clustering is done. The variables are then scaled using the scale() function in R.

```
retention1 <- copy(retention)
retention_xnominal <- retention1[,c("Geography", "Gender", "Exited", "IsActiveMember", "PersonalAdvisor", "UnresolvedComplaint"):= NULL]
summary(retention_xnominal)
# use k means to form 5 clusters on the data
retention_xnominal$FinancialLiteracy <- as.numeric(retention_xnominal$FinancialLiteracy)
retention_xnominal <- scale(retention_xnominal)
```

The function fviz_nbclust() for kmeans, with the total within sum of square method, is used to determine the optimal number of clusters. However, the resultant plot from the function does not indicate any clear optimal number. Thus, we continued with the number of clusters as 5.

```
fviz_nbclust(retention_xnominal,FUNcluster= kmeans, "wss")
```



K-means clustering is then done using the kmeans() function. The cluster centers can be seen below.

```
> Cluster <- kmeans(retention_xnominal, 5)
> Cluster$centers
  CreditScore          Age      Tenure     Balance NumOfProducts EstimatedSalary FinancialLiteracy NumberOfComplaints
1  0.05272593 -0.121883550  0.01581024  0.68347762   -0.38015998     -0.86826927     -0.0712919678        -0.25289469
2 -0.05436737  0.712772747 -0.05209154  0.26223788   -0.18799442      0.01289695      0.0065946922        -0.06170664
3 -0.02120279  0.007337253  0.07548443 -0.05748569    0.08165601      0.05014740      0.0167261344         2.77528630
4 -0.01319717 -0.156804179  0.03797296 -1.15855959    0.62589194     -0.04659791      0.0008656491        -0.25262299
5  0.01184220 -0.210652106 -0.05175864  0.66236168   -0.32793718      0.92459862      0.0613513293        -0.26286079
  AverageOfCustomerFeedbackOnService LastContactByABanker TimeBetweenRegistrationAndFirstInvestment FrequencyOfContact
1                          0.3301393          -0.29831694                              0.002880748        0.198828223
2                         -1.4759136           1.21086442                             -0.015954244       -0.811429351
3                          0.0038261          -0.01720695                              0.014984633        0.006225212
4                          0.3333560          -0.23461726                             -0.022210157        0.173867348
5                          0.3388588          -0.29706307                              0.033862989        0.179744845
> |
```

The clusters are visualised using the fviz_cluster() function from the factoextra package. Principal component analysis is used to determine the Dim1 and Dim2 since the number of columns is >2.

```
fviz_cluster(Cluster, data = retention_xnominal,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#E83628", "#B26678"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
)
```

As can be seen above, the clusters are not very distinct and clear. There seems to be a lot of overlap between the clusters.

The clusters are then added as a variable into the original data and logistic regression is done on this dataset.

```
clustered_retention$cluster<-as.factor(Cluster$cluster)
```

Original model accuray:
Trainset: 0.9547013
Testset: 0.9556519

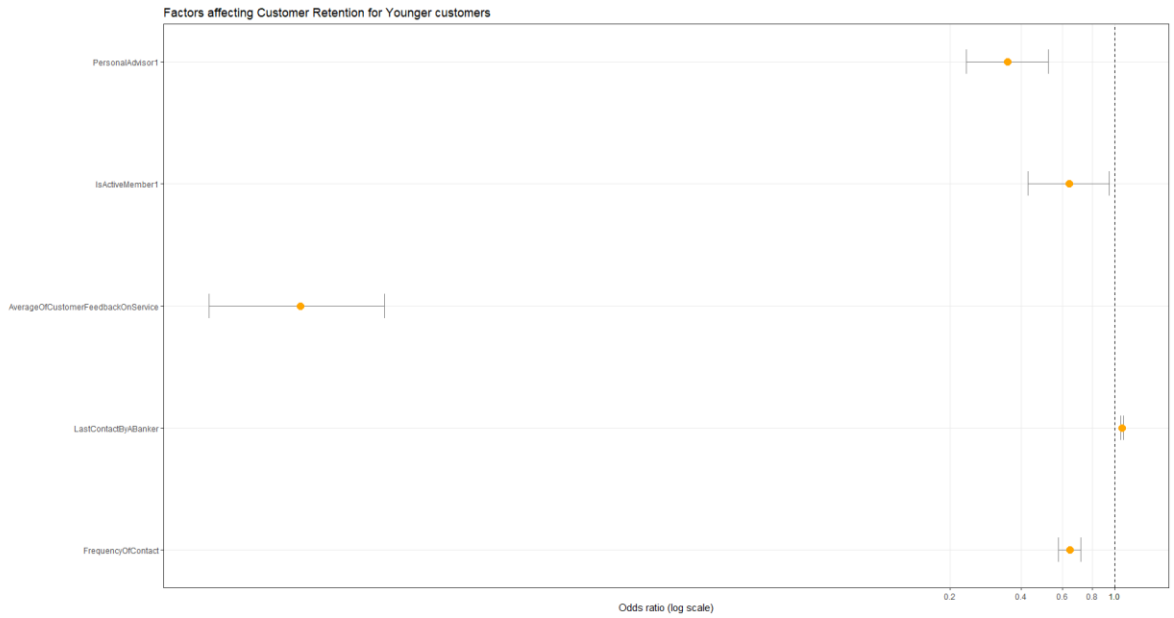
Accuracy of model with clusters:

Trainset: 0.9582738
Testset: 0.9566522

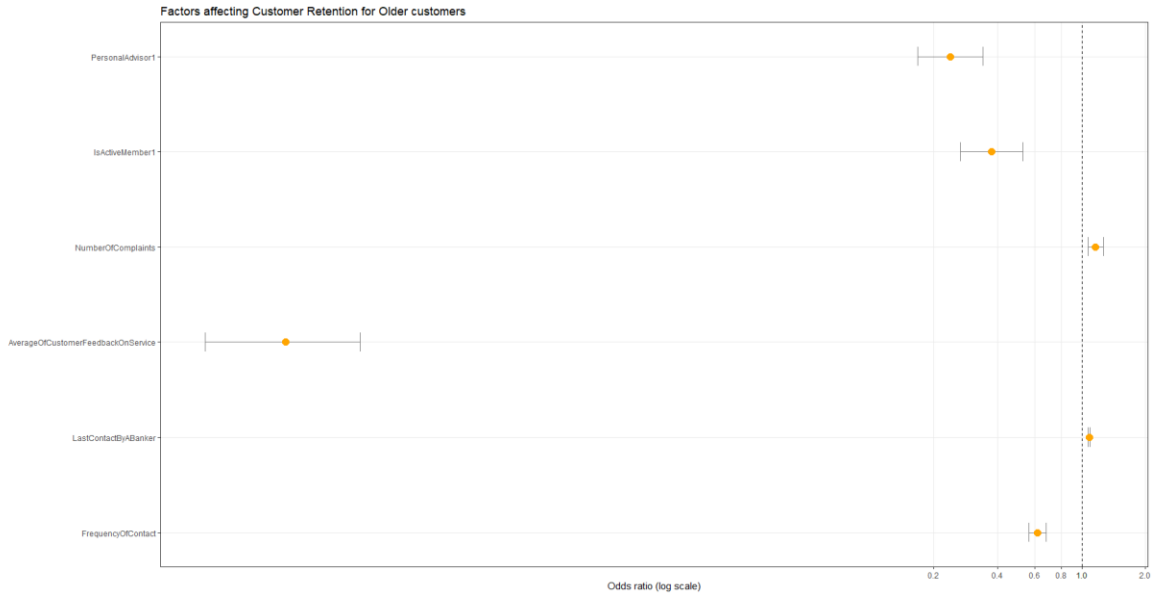
As can be seen, the accuracy is improved only slightly when the clusters are added. Thus, it can be concluded that there is not a clear clustering or a structure in the customer data.

APPENDIX E - VARIABLE ANALYSIS OF YOUNGER VS OLDER CUSTOMERS

The following graph shows the odds ratio of the various internal variables when a logistic regression model is run on a dataset with “Age” < 39.



The graph below shows the odds ratio of the various internal variables when a logistic regression model is run on a dataset with “Age” >= 39.



REFERENCES

Aren, S., & Aydemir, S. D. (2015, December 10). *The Factors Influencing Given Investment Choices of Individuals*. Retrieved 28 October 2020, from https://www.sciencedirect.com/science/article/pii/S1877042815056980

Bain and Company. (2001, September). *Prescription for cutting costs*. Retrieved 25 October 2020, from https://media.bain.com/Images/BB_Prescription_cutting_costs.pdf

Bedford, P. (2017, March 15). *Why is it SO difficult to measure Retention and Attrition accurately?* Retrieved 13 September 2020, from https://www.retentionguru.com/blog/why-is-it-so-difficult-to-measure-retention-and-attrition-accurately

Bernazzani, S. (2020, July 15). *Here's Why Customer Retention is So Important for ROI, Customer Loyalty, and Business Growth*. Retrieved 20 September 2020, from https://blog.hubspot.com/service/customer-retention

Bhandari, P. (2020, August 31). *Sampling Bias: What is it and why does it matter?* Retrieved 24 October 2020, from https://www.scribbr.com/methodology/sampling-bias/

Coval, J. D. & Moskowitz, T. J. (2001, August). *The Geography of Investment: Informed Trading and Asset Prices*. Retrieved 24 October 2020, from https://www.jstor.org/stable/10.1086/322088?seq=1

Cunningham, S., Ridley, H., Weinel, J. & Picking, R. (2020, April 22). *Supervised Machine Learning for Audio Emotion Recognition.* Retrieved 20 October 2020, from https://link.springer.com/article/10.1007/s00779-020-01389-0

ForceManager (2018, November 30) *Benefits of Customer Retention: Why it's Just as Important as Acquisition.* Retrieved 23 October 2020, from https://www.forcemanager.com/blog/benefits-of-customer-retention/

Harris, A. B. (2018, September 20). *How to Use Predictive Analysis to Improve Customer Retention.* Retrieved 27 October 2020, from https://www.business2community.com/strategy/how-to-use-predictive-analysis-to-improve-customer-retention-02119410

Hemalatha, S. (2019, April). *Factors Influencing Investment Decision of the Individual Related to Selected Individual Investors in Chennai City.* Retrieved 23 October 2020, from https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F10940486S419.pdf

JISC. *Audiovisual Research Data*. Retrieved 25 October 2020, from https://www.jisc.ac.uk/guides/audiovisual-research-data

Lebeid, M. (2018, August 8). *Misleading Statistics Examples – Discover The Potential For Misuse of Statistics & Data In The Digital Age.* Retrieved on 17 October 2020, from https://www.datapine.com/blog/misleading-statistics-and-data/

Lobato, J. M. H. (2017, July 20). *Neural networks with optimal accuracy and speed in their predictions.* Retrieved 27 October 2020, from https://towardsdatascience.com/neural-networks-with-optimal-accuracy-and-speed-in-their-predictions-d2cdc3b21b50

MacArthur, H. V. (2020, August 27). *Data And Diversity: How Numbers Could Ensure There's A Genuine Change For The Better*. Retrieved 23 October 2020, from https://www.forbes.com/sites/hvmacarthur/2020/08/27/data-and-diversity-how-numbers-could-ensure-theres-a-genuine-change-for-the-better/?sh=325c91d06907

Neff, A. (2018, January 3). *How Customer Service Must Adapt to the Largest Consumer Segment: Millennials*. Retrieved 10 October 2020, from https://www.icmi.com/resources/2018/how-customer-service-must-adapt-to-the-largest-consumer-segment-millennials

Pierrot, P. (2019, November 20). *Customer Retention in Financial Sector: Challenges and solutions*. Retrieved 21 October 2020, from https://www.victanis.com/blog/customer-retention-in-financial-services

Qualtrics. *How To Create An Effective Customer Retention Survey.* Retrieved September 13, 2020, from https://www.qualtrics.com/experience-management/customer/customer-retention-surveys/

Reichheld, F. F. and Sasser, W. E. (1990) 'Zero defections: Quality comes to services', Harvard Business Review (September–October), Vol. 68, pp. 105–111.

Riffenburgh, R. H. (2012). *Sampling Bias.* Retrieved 23 October 2020, from https://www.sciencedirect.com/topics/mathematics/sampling-bias

Saleh, K. (2017). *Customer Acquisition Vs.Retention Costs*. Retrieved 24 October 2020, from https://www.invespcro.com/blog/customer-acquisition-retention/

State of Inbound 2018 Global Report. (2018). Retrieved 21 October 2020, from https://cdn2.hubspot.net/hubfs/3476323/State%20of%20Inbound%202018%20Global%20Results.pdf

SurveyMethods. (2017, February 8). *Benefits and Weaknesses of Customer Satisfaction Surveys*. Retrieved 21 October 2020, from https://surveymethods.com/benefits-and-weaknesses-of-customer-satisfaction-surveys/

Waida, M. (2019, July 25). *Why Customer Retention Is More Important Than New Revenue.* Retrieved 21 October 2020, from https://www.wrike.com/blog/customer-retention-more-important-than-new-revenue/

Warwick-Ching, L. (2019, May 23). *Wealth managers adapt to appeal to female clients*. Retrieved 21 October 20202, from https://www.ft.com/content/585cc928-764d-11e9-bbad-7c18c0ea0201

Yu, M. P., Grustani, H. G., & Intano, M. S. (2017). FACTORS INFLUENCING CUSTOMER RETENTION AMONG BANKS. *Sci. Int. (Lahore), 29*(July-August), 729-732.