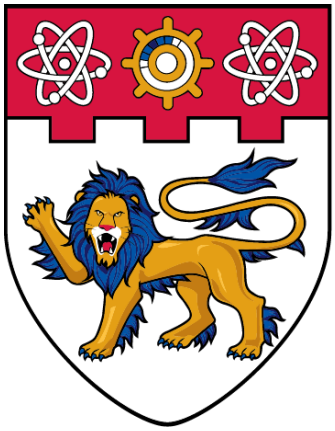


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 Analytics I: Visual & Predictive Techniques

Semester 1 - AY 2020 / 2021

Seminar Group 7, Team 5

Tutor: Prof. Sanli Pinar Darendeli

Name	**Matriculation Number**
Chen Jiaru	U1810473G
Myat Thu Soe	U1910462C
Oh En Qi Sherman	U1910471K
Venkat Subramanian	U1921075D

Table of Contents

1. Executive Summary

Problem Statement

White Rock is a 300-year-old asset management company formed in London with a strong presence in 32 markets globally. Despite strong growth in this industry, the increased competitiveness in this industry may lead to clients ending their relationship with the firm, which can be attributed to unmet expectations or a change to better alternative companies. This is known as churn. An increased churn will lead to lower revenues and higher costs as clients are the key revenue drivers of White Rock. In order to stay competitive, White Rock needs to improve on their deficiencies to avoid an increased churn. With today's technology, many companies turn to analytics as a powerful tool to create business insights. Analytics harnesses historical data for the purpose on creating future predictive applications. How can White Rock utilise analytics to determine the key factors of churn and the significant moments where churn occurs?

Proof of Concept

The *churn modelling* data, that is based on the context of a bank, is used for data exploration, analysis, preparation and creation of the **Churn model**. Using the model, White Rock can effectively i) Predict the moment of churn, ii) Predict the factors contributing to churn and iii) Improving their areas of deficiencies to meet the needs of their clients. The team utilised the data set on a few sets of assumptions. Given that the data is unbalanced, proportion-based visualisations were also used for data exploration. From the initial exploration, the NumOfProducts, Geography, Gender, Age, IsActiveMember and Balance were found to have a correlation with exiting the bank. These variables were then studied in detail and key observations were noted in this report. After cleaning and preparing data, the churn prediction models were then built using Classification and Regression Tree (CART) techniques as it helps identify the important variables and subsequently 'moments of impact'.

Further, a balanced trainset was created considering that the original dataset was imbalanced; the difference in the variable importance of the Churn model was observed, along with changes in accuracy, specificity and sensitivity. To evaluate the model's overall effectiveness, accuracy, sensitivity and specificity of the CART model was compared with that of the logistic regression techniques. The team concluded that the CART model trained using a balanced trainset has the most optimal overall accuracy, sensitivity and is the most appropriate model. It was concluded from the variable importance derived from our CART model that Age, NumOfProducts, CreditScorePerAge, NormalisedBalance and BalancePerProduct are the key factors to churn.

Benefits and Impact of Recommendation

The team recommends the Churn model to White Rock as it can be directly applied given that the variables in the dataset are mappable to that of White Rock's. However, White Rock needs to test for the accuracy and sensitivity of the model based on historical records, and if necessary, re-train the model to improve the results. The results can then be used to engineer a "customer retention program" to target potential churns.

Subsequently, A/B testing can be conducted to pinpoint the effects on churn that their business strategies bring about, thereby allowing them to adapt their program accordingly. In view that NumOfProducts is a pivotal variable for churn, the team recommends White Rock to employ pricing strategies such as lowering pricing registrations cost for easy entry to their products, while making other costs expensive to deter them from switching products. All in all, our model will set a foundation for White Rock to explore plenty of opportunities on top of our above recommendations and will undoubtedly give them the competitive edge in the asset management industry.

2. Business Proposition

2.1. Introduction to White Rock and its Business Environment

White Rock is a 300-year-old company formed in London with 32 markets globally, including 8 in Asia. It operates mainly as an asset manager, with institutional clients as well as intermediaries who distribute their investment products.

The asset management industry is experiencing rapid growth with a strong pipeline of investment assets which is set to grow from $64 trillion in 2014 to about $102 trillion by 2020, a compounded annual growth rate of 6%. (PricewaterhouseCoopers, 2020). In Singapore, the asset management industry expanded at a 14% compound annual growth rate from 2013 to 2018. The stable political and economic characteristics allows Singapore to be a conducive and thriving place for asset management activity, increasing market opportunities in the Asia Pacific (MAS, 2018).

To stay competitive and relevant in today's digital age, White Rock wishes to exploit internal data which they have collected from transactions, processes and activities monitoring to automate and make informed business decisions, with hopes to improve business processes across White Rock.

2.1.1. Business model analysis

White Rock's key activities encompass operations, distribution (sales and marketing), and investment management services. They manage and invest on behalf of clients, including high net-worth individuals, government entities, corporations, financial institutions and intermediaries (Ganti, 2020). White Rock then levy a fee in return for their services of growing their clients' portfolio and helping them mitigate risk. In essence, clients are the key revenue drivers of White Rock which makes providing the best value, services and ensuring client satisfaction and retention essential to White Rock.

2.2. Business Problem

As clients are the key revenue drivers of White Rock, the inability to retain customers prevents White Rock from reaching their revenue targets, increases unnecessary cost and brings about other unintended consequences detrimental to the company.

Therefore, it is in White Rock's best interests to analyse and understand the attributes of clients who leave, so that they can take steps to prevent clients from leaving and thereafter reduce its churn rate. Accordingly, the churn rate will be defined for this report and the various impacts and root causes of high churn will be explored.

2.2.1. Churn Rate

Churn rate, also known as the rate of attrition or customer churn, is defined as the rate at which customers stop doing business with an entity. (Frankenfield, 2020). In White Rock's context, it can be expressed as the percentage of clients who discontinue their business relationship with White Rock. The formula is as follows:

$$Churn\ Rate\ = \frac{Clients\ Lost\ in\ a\ Period}{Total\ Clients}$$

2.2.2. Impacts of High Churn

In the increasingly competitive asset management industry that White Rock is in, the implications of high churn rate can be extensive and can snowball into bigger problems such as a significant reduction in profits and lower employee productivity. The following are key negative impacts of high churn rates on the company.

1) Direct Financial Cost

The true cost of churn is far-reaching, as it not only involves a loss in future revenue but also has a significant impact on the future cost of the company. The true cost of churn can be observed in three aspects (ClientSuccess, 2018).

Firstly, it would be the revenue that the company would have earned for each year the customer was retained.

Next, it would be the potential for loss of income from upselling deals to the lost customers. According to the book Marketing Metrics, businesses have about 60% to 70% chance of selling to an existing customer while the probability of selling to a new prospect is only 5% to 20% (Hull, 2013). Consequently, each client churn will also result in a significant decreased probability of upsell or cross-sell opportunities.

Lastly, it would be the cost of acquiring new customers. The marketing team will have to redirect time and resources to bring in new leads, prospective customers and to re-attract customers that have been lost. Research shows that it can cost up to 5 times as much to acquire a new customer than it does to retain current ones (Wertz, 2018).

One way to quantify the direct cost of churn is through the Lifetime Value (LTV) formula. The customer LTV is the profit margin a company expects to earn over the entirety of their business relationship with the average customer (Karnes, 2020). To reap higher revenues, a business should aim to maximise the LTV (Alex, 2019). The LTV formula is as follows:

$$Lifetime\ Value\ =\ [Average\ Revenue]\ X\ [Number\ of\ Transactions]\ X\ [Retention\ Time\ Period]$$

Having a high churn rate implies a low retention period, which in turn reduces the lifetime value of the customer. This also means that the business needs to spend more to acquire new customers, thus increasing its Customer Acquisition Cost (CAC). To ensure continuous success of White Rock, it is important to minimise CAC and maximise LTV.

From the analysis above, it can be observed that customer churn indeed has a significant negative impact on the current and future profits of White Rock.

2) Reduction in Employee Satisfaction

When the churn rate is high, employees would have the added pressure to attract and retain every potential client. This applies particularly to the asset management industry, where employee remuneration is often tied to the fees levied on the client. Over time, the constant hunt for clients can be exhausting if they repeatedly fail to get new clients. Employees may become increasingly less motivated, causing their morale to dip. With poor remuneration and a lack of opportunities to build meaningful bonds with clients, employee satisfaction can fall.

Customer satisfaction strategies begin with employee satisfaction (Heskett, Sasser and Schlesinger, 1997). Employees must first be satisfied with their-work life before companies can motivate them to improve performance and productivity, which in turn impacts the customer experience. Therefore, a fall in employee satisfaction can lead to lower customer satisfaction. As a result, churn increases as more clients turn away from White Rock to other asset managers who can better cater to their needs (Menafn, 2020) and reduce the revenue growth and profitability of White Rock.

Figure 1: Effect of Employee satisfaction on Shareholder Value

2.2.3. Root Causes of High Churn

The reasons as to why clients leave a business can be attributed mainly to the following three aspects. (Retently, 2015)

1) Poor Client Onboarding Experience

Despite rapid advances in technology, client onboarding remains highly manual for most asset management firms (O'Neill, 2018). It is inevitable for human error to arise during the manual process of information collection and data transfer, which leads to mistakes, duplication and a lack of an efficient customer data directory. Often, repeated requests for the same information are made to the customer. This inefficiency may lead to client churn if competitors in the same industry can provide a smoother client onboarding experience.

2) Poor Customer Service - Failure to Provide Desired Outcomes

Clients typically engage asset managers to solve problems and to achieve their desired outcome. Asset managers determine what investments to make and avoid, to grow a client's portfolio (Ganti, 2020). Most clients who engage the asset manager would expect to achieve a certain level of positive returns from the investments. Hence, it is not unusual for clients to switch asset managers if the returns and their desired outcome are not met. Furthermore, successfully tackling client issues in the initial interactions is crucial granted that client churn can be reduced by 67% at this stage. (Afshar, 2017)

3) Decreasing Relevance of Services to Clients

The asset management industry is continually evolving in terms of regulations, fee pressures, margin compression, and changing client expectations (Eley, 2020). Failure to keep up with the recent trends and changes to provide the level of service that is most relevant to the clients' needs will lead to churn.

2.3. Business Solution

To reduce the overall client churn rate, White Rock would need to be able to pinpoint the 'moments of impact', allowing them to predict when clients will leave, and to take immediate action to mitigate this. Below are the general solutions that White Rock can employ to reduce their churn rate.

1) Identifying moments of impact and enhancing areas of deficiencies

The team defines the 'moments of impact' as the key factors that result in their clients switching to other asset managers. By identifying and analysing the moments of impact, White Rock will be able to understand the reasons behind a client's decision to leave them and thereafter take action to improve its areas of deficiencies and deliver a better experience to its clients. (Campbell, 2017)

2) **Formulating business strategies**

Aside from enhancing their areas of deficiencies, White Rock can also come up with new business strategies to retain their clients and if possible, capture more market share. There is a constant need to innovate to stay competitive in today's asset management industry. In order to retain customers, one key area that White Rock should focus on is to deliver better customer experiences and improve relationships with their customers. The implementation of such strategies can potentially increase its customer lifetime value and boost the competitiveness of White Rock as an asset manager.

2.3.1. Measuring the Effectiveness of Our Business Solution

To assess the effectiveness of the suggested solutions, White Rock can use the following measures:

1) **Churn Rate**

This is the most direct measurement of the effectiveness of the proposed solution considering that its main purpose is to reduce client churn. The lower the churn rate, the more desirable it is.

2) **LTV/CAC Ratio**

This ratio can estimate the direct financial impacts of our solution. The success or failure of our solution will decrease or increase the churn rate, which will in turn increase or decrease LTV and affect the CAC. An LTV/CAC ratio of less than 1.0 suggests that the company is destroying value while a ratio of more than 1.0 suggests that the company is creating value. Hence, a company should strive for a high ratio (CFI, 2017).

3) **Client Relationship Management (CRM)**

To ensure the long-term success and sustainability of the proposed solution, there is a need to ensure that the solution can continue to meet customer and employee expectations. This is especially so for a service firm like White Rock, where the reliance and trust of customers in White Rock plays a key part in their decision to stay. Therefore, apart from quantitative metrics, survey and feedback can be obtained from the relevant parties to assess the client and employee satisfaction (Juneja, 2010). These results can be used to improve client and employee experience.

3. Analytics Proposition

To tackle the business problem, the guiding questions are as follows:

1) Given the dataset, how can a statistical model which measures whether a client leaves at a certain time be developed?

2) Is the team able to generate a model which identifies the key factors of churn?

3) With the identified factors contributing to the churn, will the model be able to identify the significance of these factors, from:

a) a business perspective for the business to make strategic decisions to reduce their deficiencies in those areas, and;

b) a statistical modelling perspective to identify the moments of impacts, the significant 'splits' in a decision tree?

3.1. Analytics Research Methodology

The analytics aspect of this report can be broken down into three parts:

1) **Obtaining a data set for our business analytics problem**

   For the team to begin the analytics process, it was fundamental to obtain a dataset that provides context, relevance, usability to White Rock's business problem. The dataset should also have a sufficiently large sample size to make accurate estimations, and thereafter draw concrete conclusions that are valuable for White Rock's business case.

2) **Data cleaning and exploration of the data**

   Before building the model, the data needs to be cleaned and prepared. By doing so, the insights generated from our data are more accurate and reliable (Rouse, 2008). Inaccurate data analytics can lead to misguided decision-making which in turn have a severe impact on business performance. The cleaned data is then used to perform predictive analytics to identify variables with significant relationships (Bhatt, 2020).

3) **Employing Predictive Analytical techniques**

   For this report, CART and logistic regression techniques were employed. CART allows for the identification of the moments of impact of churn by observing the relative importance of variables. Subsequently, logistic regression techniques can be utilised to compare the accuracy, specificity and sensitivity of the Churn Model.

3.2. Analytics Outcome and Performance Measures

Given that the outcome variable, "Exited", is categorical, Confusion Matrix will be used to evaluate the performance of the Churn Model, which is built on the CART model and compared against the logistic regression model.

The confusion matrix allows for the calculation of the accuracy, sensitivity and specificity of the model. The sensitivity and specificity score are especially significant given that the dataset used is imbalanced. The team believes that sensitivity should also be the main evaluation metric as it will enable White Rock to identify a greater number of customers that might leave. This is so despite its potential trade-offs, such as being less accurate, and that White Rock might end up spending more on their retention program. Conversely, this will improve CRM with its clients.

4. Data Review and Exploration

4.1. Data Review

4.1.1. Dataset Selection

Churn_modelling dataset was used to build the Churn Model. There are two main reasons why the team deemed this dataset as suitable for the analysis.

Firstly, the dataset is set under the context of the bank industry. Hence, this is a similar representation of White Rock's clients given that the business model in the banking industry is similar to that of the asset management industry since both industries provide financial services to their clients. Thereafter, the context of variables in the bank churn data can be related to White Rock (see **Appendix A**).

Secondly, this dataset consists of 10,000 rows with 11 relevant attributes. Therefore, there are sufficient observations to make reliable predictions using the CART and logistic model. Ggplot2, Hmisc, ggpubr and arsenal packages were used to visualise and summarise the dataset.

4.1.2. Questions and Observations

The team notes that although the dataset is sufficiently large and comprises seemingly relevant variables, there are some apparent questions and observations that need to be addressed:

- If a customer has exited, why are they still labelled as being active, 1 for the categorical variable “IsActiveMember”?
- Why do some customers have balances greater than 0 although they have exited the bank?
- Only the number of products is provided as a variable under “NumOfProducts”. Could the type of products also affect whether a customer will exit?
- The data is a point-in-time dataset without any variables that contain the dates to which they entered or exited.

4.1.3. Key Assumptions

Definitive answers for the aforementioned questions were not found due to the lack of a metadata or data dictionary. Hence, the following assumptions were made when analysing the dataset in the latter stages:

1) The activity of the customer is dependent on their level of patronage of the bank services in a period.

2) The bank balances refer to the balance that the customers had before they subsequently exited the bank (for customers who exited the bank).

Product types affect churn rate as some product types have higher switching cost than others (Geckoboard, 2020). Furthermore, the date of entry and exit for a customer would be vital in calculating average retention period and the customer LTV, which could prove to be a useful key performance indicator (KPI). As such, the lack of relevant data poses a limitation granted their importance in analysing the churn rate. Nevertheless, the team will work with the limited dataset and evaluate the results based on the data at hand.

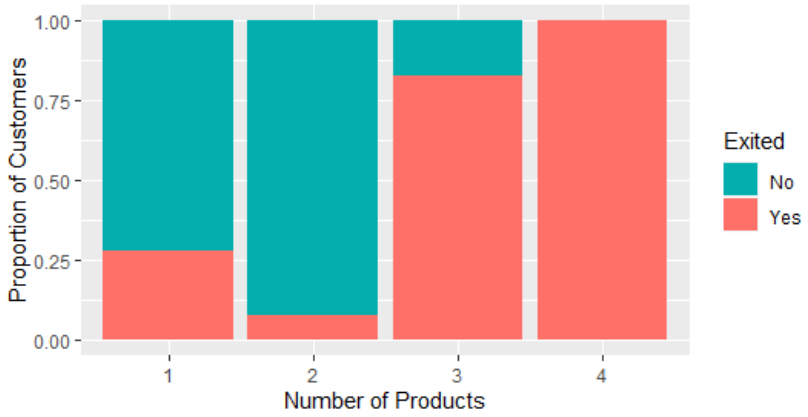
4.2. Data Exploration and Visualisation

4.2.1. Summaries and Surface Analysis

For the detailed statistical summary of the predictor variables against “Exited” and summary of the variables, see **Appendix B**. The summary of the findings is as follows:

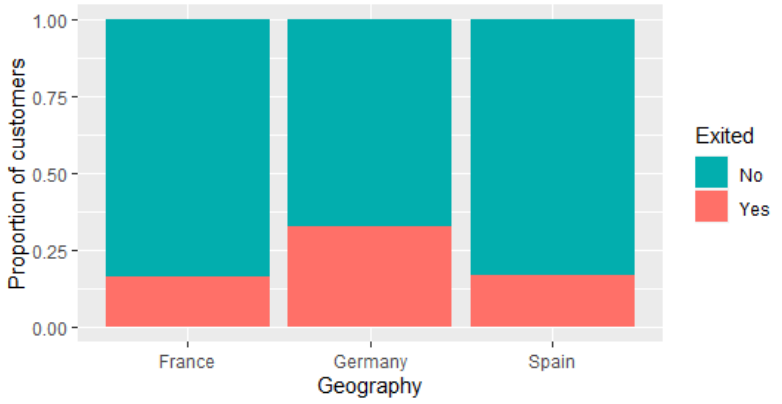
1) There are fewer customers who have exited (2,037 customers) than those who have not (7,963 customers) (see **Appendix C** for relevant graphs). There is a need to adjust for the difference in the proportion when exploring the data, by using visualisations that work with proportions/distributions such as percentage stacked bar charts, boxplots etc.

2) The number of products that the customer owns correlates with whether or not the customer will exit granted that the number of customers that owns either three or four products is a small sample size of 326, a meagre 3.26%.



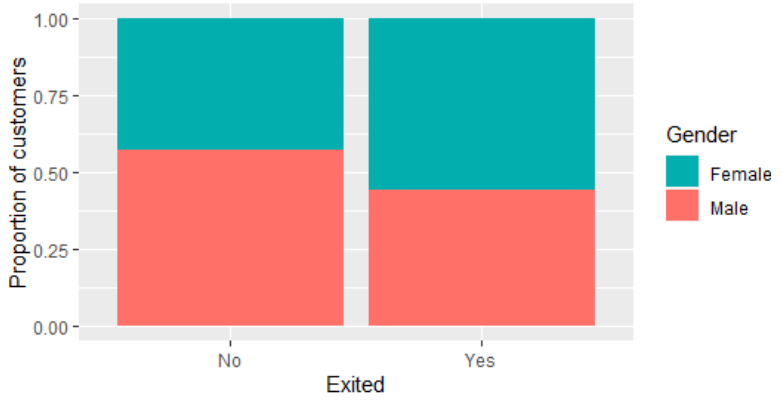
Graph 1: Impact of the number of products on customer exiting

3) Geography of the customer correlates with whether or not the customer will exit. (i.e. although German customers make up only 25.1% of the dataset, they make up 40.0% of those who exited).



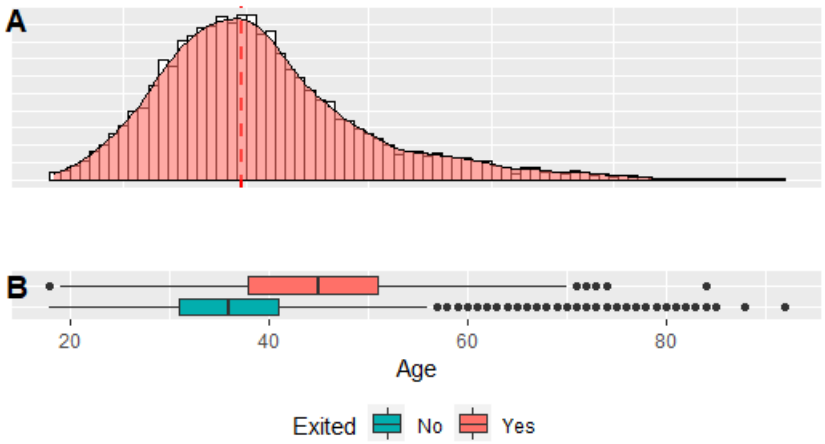
Graph 2: Impact of Geography on customer exiting

4) Gender of the customer also correlates with "Exited" (i.e. although female customers make up only 45.4% of the dataset, they make up 55.9% of those who exited).



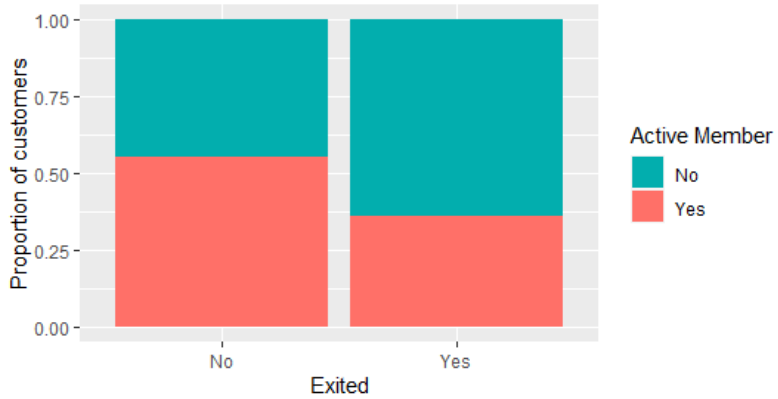
Graph 3: Impact of Gender on customer exiting

5) Age of the customer also displays a correlation with whether or not the customer will exit. The mean age of customer who exited was 44.8 while those who stayed had a mean age of 37.4 (see Graph 17 of **Appendix C**). Further, it is noteworthy that those who did not exit follows the distribution of the age of all customers.



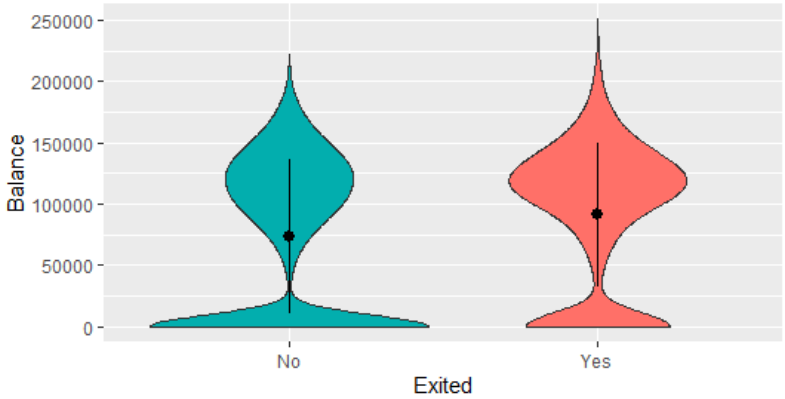
Graph 4: Distribution of Age on customer exiting

6) There is a higher proportion of active members who stayed at 55.5%, as compared to the proportion of active members who exited at 36.1%.



Graph 5: Impact of member's activity on customer exiting

7) Those who exited have much higher balances (averaging at 91,109) as compared to those who stayed (averaging at 72,745), indicating a correlation. However, with reference to the distribution, it appears that those who stayed are more spread around the base balance of 0, while those who exited have a higher distribution at around 120,000.



Graph 6: Impact of Balance on customer exiting

A bar plot of customers with 0 balance shows that 85.15% of this group of customers stayed (see Graph 15 in **Appendix C**). In the bank context, it could refer to customers who had created a bank account or opted for investment services but did not deposit any

funds for that purpose. Hence, the team will remove this group in the event it leads to an improvement in the analysis during the data exploration phase.

However, it should be emphasised that correlation does not imply causation. This begs the question of whether there is a common causal variable that is present in the mentioned predictor and outcome variables. Uncovering this information, although "unnecessary" since the CART model has an automatic variable selection feature, will ensure a more accurate model that detects the hidden causes. As such, White Rock can utilise this information better in implementing solutions or make improvements regarding their churn rate.

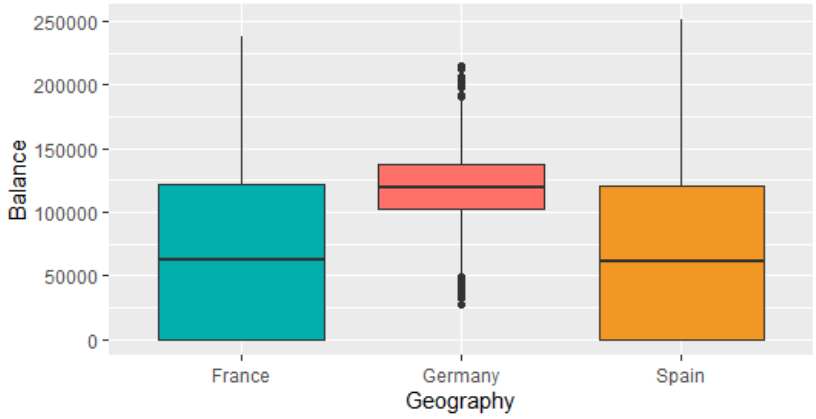
4.2.2. Variable specific analysis

1) Geography

The team suspects that there is a common causal variable that results in German customers having a higher likelihood of leaving as compared to customers from other countries. Hence, the guiding question for further exploration was whether other variables align with the correlation between Geography and Exited.

Observation 1

Distribution of balance by country reveals that German customers have the highest median balance. This is in alignment with the observation that those who exited have higher balances on average. Hence, balance is a potential common causal variable.



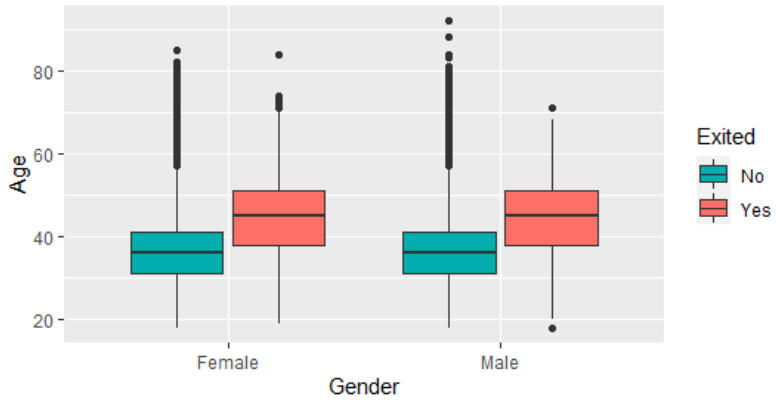
Graph 7: Distribution of Balance by Geography

2) Gender

The guiding question for further exploration (for gender) was whether any other variables align with the observation that more female customers leave the bank compared to male customers.

Observation 1

The boxplot of Age against Gender reveals that the departure was due to the age of the customer, rather than the gender.



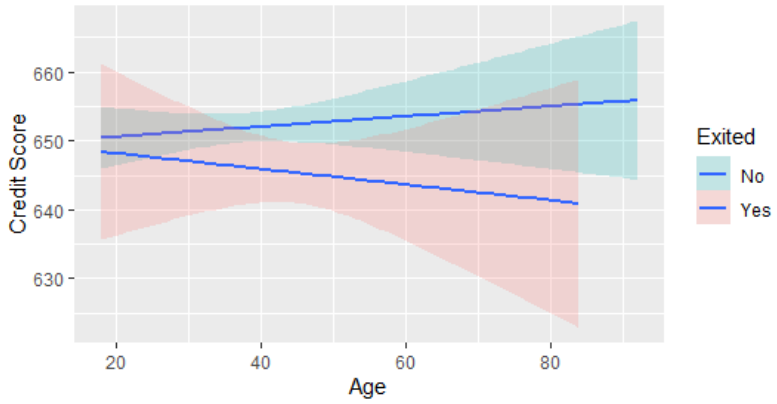
Graph 8: Impact of Age against Gender on Exited

3) Age

Though arguable that age comes with financial literacy, the team believes that there could be other factors apart from age that causes the customers to exit. Hence the guiding question for further exploration was whether any common causal variables align with the observation that there are more older customers who leave the bank.

Observation 1

The smoothed plot was utilised to plot Age against different variables. In the Credit Score against Age plot, observations were made that there was a negative correlation for customers who exited but positive for those who stayed.



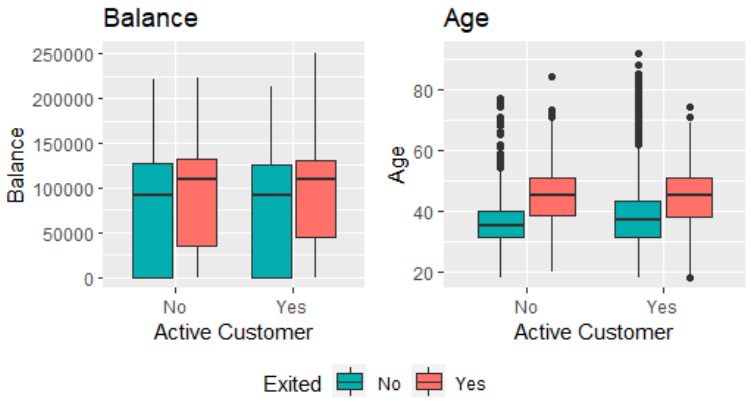
Graph 9: Impact of Credit Score against Age on Exited

4) Active Member

As observed earlier, the activeness of the member affects their departure. Nevertheless, the team tested out on whether the departure could be related to another variable.

Observation 1

Boxplots of Balance and Age against activeness of member reveals it is probable that the departure was due to the balance and age of the customer rather than the activity.



Graph 10: Distribution of Balance and Age against Active Customer on Exited

Observation 2

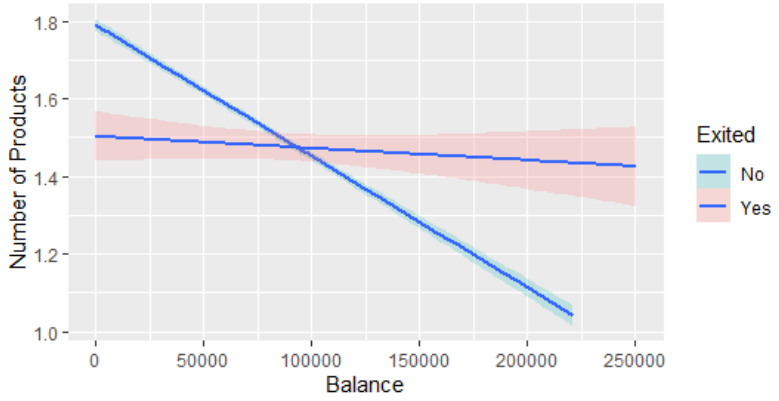
Plotting Balance Against Age (after conditional removal of 0 Balance) shows that customers that exited are mostly above the age of 43. What is even more intriguing is that the majority (94.89%) of those above 65 that stayed are active members (see Graph 16 in **Appendix C**). Hence, the activity of members could be the second level of filter for whether a customer will churn, after looking at the age or balance.

5) Balance

Although balance is likely one of the main indicators and cause of customer departure. The team wishes to explore the variable ("Balance") against other variables, and their respective effect on the departure.

Observation 1

The smoothed plot of the number of products against balance shows that there is a negative correlation for those who stayed, but little to no correlation for those who exited. This could be an indicator that balance adjusted for the number of products is the cause of customer leaving.



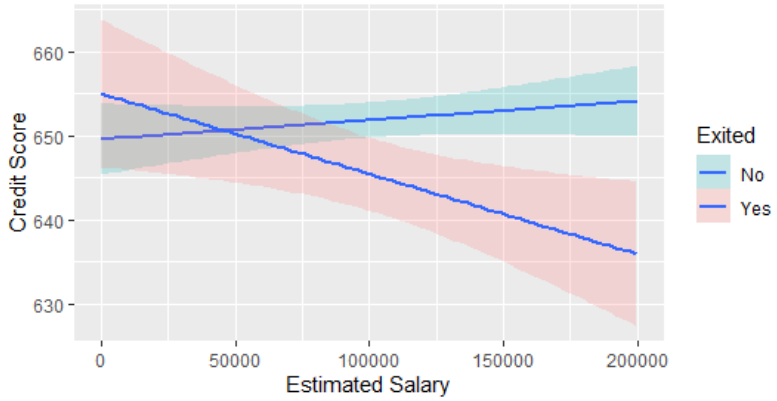
Graph 11: Impact of number of products against Balance on Exited

6) Estimated Salary

The initial statistical summary did not reveal any correlation between Estimated Salary and possibility of departure. Nevertheless, the team felt that the estimated salary was paramount to the bank given that it can identify the bank's high-value customers. Hence, a deeper analysis on the estimated salary was conducted.

Observation 1

The smoothed plot of Credit Score against Balance shows that there is a negative correlation for those who exited but positive correlation for those who stayed. Hence, this implies that the estimated salary adjusted for credit score is the cause of departure and not one variable alone.



Graph 12: Impact of Credit Score against Estimated Salary on Exited

5. Data Cleaning and Preparation

Data cleaning and preparation need to be conducted before building the model to ensure greater accuracy for our analysis and model. Here, the aforementioned problem of "correlation does not imply causation" will be addressed through feature engineering. In this step, data.table and the base packages were used to clean and prepare the data. Initially, the dataset consisted of 14 columns and 10,000 rows.

5.1. Data Cleaning

5.1.1. Dropping Irrelevant columns

The following three columns were dropped due to their lack of relevance:

1) RowNumber serves no purpose in this analysis.
2) CustomerId is a unique ID to identify customers and serves no purpose in this analysis.
3) Surnames are just customers' names and are unlikely to affect whether the customer exits.

5.1.2. Working with missing values

Generally, missing values are replaced with the mean, median, mode, other predictive values, or complete removal of the row. However, in this dataset, there are no missing values. Furthermore, even if there were missing values, the CART model handles them well due to the presence of surrogate splits which are activated on missing values.

5.2. Data preparation

5.2.1. Factorising Categorical Variables

Geography, Gender, HasCrCard, IsActiveMember and the predictor variable Exited are all nominal variables as the order does not matter. Instead, the numbers only serve as labels. Hence, they have been categorised without any order using the factor function.

5.2.2. Feature Engineering

As hinted above, the team will be conducting feature engineering on some predictor variables so that the model will better reflect causation. The features engineered are as follows:

1) A variable **normalisedBalance** is created by dividing the Balance of a customer by the mean balance of the Geography that they live in.
2) A variable **creditScorePerAge** is created by dividing a customer's Credit Score by their Age.
3) A variable **balancePerProduct** is created by dividing a customer's Balance by the Number of Products they own.
4) A variable **salaryCreditScoreRatio** is created by dividing a customer's Salary by their Credit Score.

5.3. Cleaned Dataset

The resulting cleaned and engineered dataset has 15 columns and 10,000 rows. It should be noted that the cleaned dataset was used for training both CART and logistic regression models.

6. Building the Model

6.1. Methodologies and Considerations

As mentioned earlier, CART and Logistic Regression models will be used to predict the outcome (categorical) variable "Exited". The dataset was split 70-30, for train and testset respectively, using R package caTools. Train-test split ensures that the trained model has improved performance on real-world data and decreased test prediction error. Another set of models was also built with a balanced trainset to predict the minority cases better.

To evaluate the CART models' overall effectiveness on the testset, the accuracy, specificity and sensitivity obtained from the confusion matrix were compared against that of the logistic regression models. The team then evaluated and selected the appropriate model that will fulfil our main evaluation criteria, higher sensitivity with relatively high accuracy.

6.2. CART Model

CART analysis was conducted on the balanced and unbalanced trainset. The rpart and rpart.plot packages were used to visualise the tree and identify the moments of impact. Further, the model was built using "class" for the parameter method as "Exited" as a categorical variable. Other benefits of CART models are that 10-Fold Cross-Validation (CV) is automatically applied and it can handle missing values through surrogate splits. In the creation of the balanced trainset, R package dplyr was used.

6.2.1. Unbalanced Trainset

Growing Out the Tree

The tree is grown to the maximum using the following lenient stopping criteria: (i) minimum split of 2 and complexity parameter (cp) of 0. However, the maximum tree has more than 80 splits. Hence, there is a need to prune the tree further to overcome overfitting and to simplify the tree.

Pruning the Tree

The tree was pruned based on the one standard error rule as our optimisation criteria. The simplest tree whose CV Relative Wrror was within the CV error cap (of one standard deviation)

was used as the optimal tree. The prune trigger is identified to be 0.0124 (see Graph 29 in **Appendix E**). After pruning the tree with the prune trigger, the optimal tree has only 8 splits.

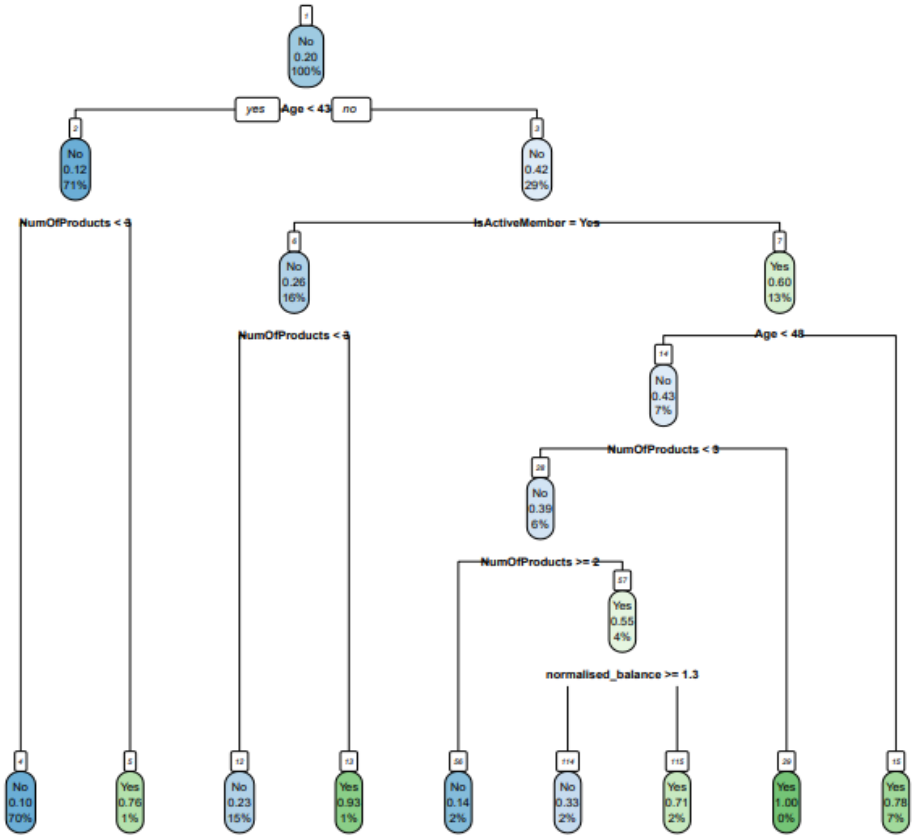


Figure 2: Pruned tree for the unbalanced trainset

Variable importance

The variable importance of the pruned tree shows that Age, NumOfProducts, IsActiveMember, and the engineered variables, creditScorePerAge and balancePerProduct were the most important factors in predicting the outcome (see Table 12 in **Appendix E**).

Prediction on the testset

The model was used to predict the testset and the following confusion matrix was produced:

		Predicted	
		No	Yes
Actual	No	2297	92
	Yes	358	253

Table 1: Classification tree for predicted testset using CART model trained on unbalanced trainset

6.2.2. CART Model for balanced trainset

Creating the balanced trainset

Within the trainset, the “Yes” and “No” cases for Exited were split up. 1426 rows from the “No” cases were then sampled (which was the number of “Yes” cases in the trainset) and combined with the “Yes” cases to create the balanced trainset.

Finding the optimal tree

This new balanced trainset is then used to train the CART model. Again, the tree was grown to the maximum, using the same lenient stopping criteria, and the cp values were plotted. The prune trigger was cp = 0.00496 (see Graph 31 in **Appendix E**). The tree was subsequently pruned using that cp.

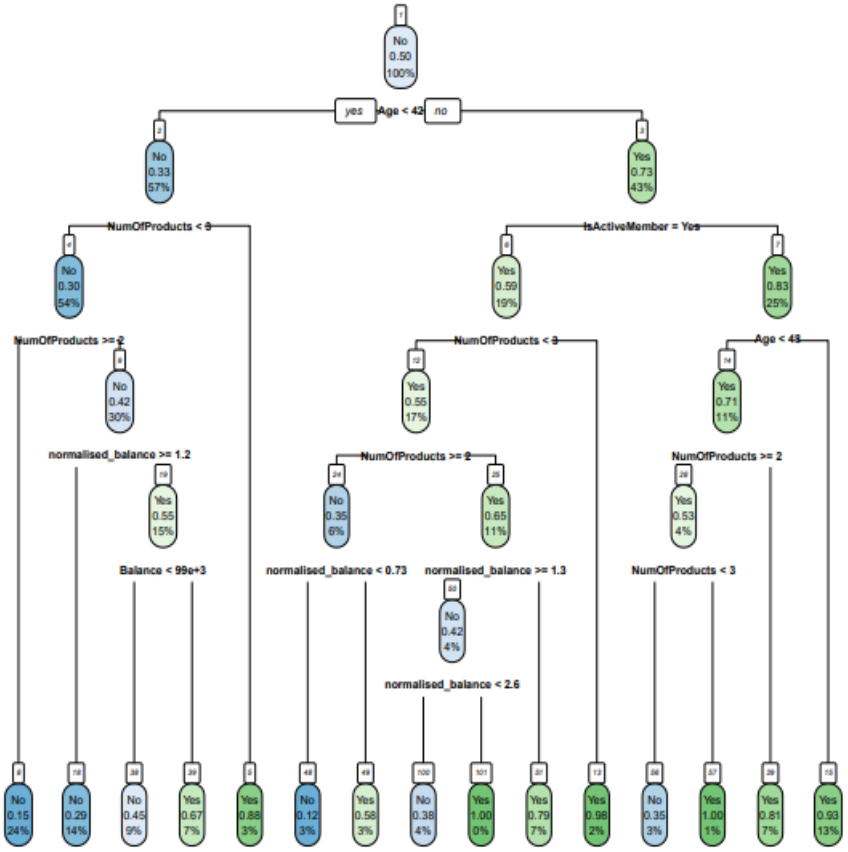


Figure 3: Pruned tree for the balanced trainset

Variable importance of the balanced CART model

Compared to the unbalanced pruned tree, it was observed that the importance of some variables in the balanced pruned tree was different. In the top 5 positions, there is a new variable, normalisedBalance that replaced IsActiveMember (See Table 13 in **Appendix E**).

Prediction on the testset

The confusion matrix of the model built on balanced trainset and predicted on the testset is as follows:

		Predicted	
		No	Yes
Actual	No	2008	381
	Yes	199	412

Table 2: Confusion for predicted testset using CART model trained on balanced trainset

6.3. Logistic Model

The Logistic model was subsequently used to compare against the CART model. Again, both balanced and unbalanced trainsets were used to build the model. For logistic regression, R Base Package function glm with parameter family equals binomial was used to build the model.

6.3.1. Logistic Model on unbalanced trainset

Finding the Optimal Logistic Model

The first logistic model was trained using the same unbalanced trainset. The model selection was automated using the step function which removes the variable that has the highest Akaike Information Criteria (AIC). AIC deals with overfitting and underfitting issues by ensuring that the trade-off between goodness of fit and the simplicity of the model is addressed. The resultant model retains key variables thereby not resulting in underfitting while removing variables that might cause overfitting. Comparing the odds ratio to the optimal logistic model obtained earlier, it was observed that the variables that capture the value of 1 in their confidence interval (and hence are statistically insignificant) were removed.

Prediction on the testset using standard threshold

The confusion matrix of the logistic model predicted on the testset using 0.5 threshold is as follows:

		Predicted	
		No	Yes
Actual	No	2311	78
	Yes	475	136

Table 3: Confusion matrix for predicted testset using Logistic model trained on unbalanced trainset with 0.5 threshold

Finding the optimal threshold using the CART model

To see whether the 50% threshold was suitable in differentiating the "Yes" and "No", the distribution of the probability of outcome was plotted for "Exited" cases (See Graph 33 in **Appendix E**). From the plot, it is observed that using a threshold of 0.5 is not good enough to distinguish the category clearly. Hence, the team attempted to find another threshold which would identify the cases best. The first split point of the CART model was used as a proxy for the optimal threshold (see **Appendix E** for further explanation). By using a value of **0.37** for the threshold, the sensitivity of the model greatly increased.

Prediction on the testset using the optimal threshold

		Predicted	
		No	Yes
Actual	No	2172	217
	Yes	365	246

Table 4: Confusion matrix for predicted testset using Logistic model trained on unbalanced trainset with 0.37 threshold

By using a lower threshold, the accuracy decreased (by 0.97%) and the sensitivity was increased (by 18.0%). However, a logistic model with high accuracy and sensitivity was still unobtainable.

6.3.2. Logistic Model on Balanced Trainset

Subsequently, the steps for the unbalanced trainset were repeated for the balanced trainset. Again, the step method was used, and the odds ratio of the significant variable were checked (see Image 4 in **Appendix E**). It was observed that statistically insignificant variables were removed. Next, the outcome was predicted using the threshold value of 0.5.

Prediction on the testset using standard threshold

The results using the standard threshold of 0.5 is as follows:

		Predicted	
		No	Yes
Actual	No	1683	706
	Yes	191	420

Table 5: Confusion matrix for predicted testset using Logistic model trained on balanced trainset with 0.5 threshold

There was a significant improvement in sensitivity when compared to the logistic regression model trained on the unbalanced trainset. However, there was a drop in the overall accuracy. Hence, the threshold was changed to one that would increase the sensitivity.

Prediction on the testset using the optimal threshold

The following confusion matrix is obtained from changing the threshold value to 0.55 (refer to **Appendix E** for further explanation):

		Predicted	
		No	Yes
Actual	No	1844	545
	Yes	235	376

Table 6: Confusion matrix for predicted testset using Logistic model trained on balanced trainset with 0.55 threshold

The overall accuracy increased by 3.9% but the sensitivity decreased by 7.2%. An increase in the accuracy of the model resulted in a more than proportional decrease in the sensitivity. Therefore, the team concluded that it was best to keep the threshold of the model at 0.5.

6.4. Model Evaluation

6.4.1. Evaluation Criteria

The evaluation metrics for each model are summarised in the table below.

Model	Overall Accuracy	Sensitivity	Specificity
CART unbalanced	85.0 %	41.4%	96.1%
CART balanced	**80.7%**	**67.4%**	84.1%
Logistic unbalanced (with 0.37 threshold)	80.6%	40.2%	90.9%
Logistic balanced (with 0.5 threshold)	70.1%	68.7%	70.4%

Table 7: Summary of the evaluation metrics for different models

From the table above, the team concluded that the CART model trained on the balanced trainset is the most suitable to predict the churn of customers given the main evaluation criteria of sensitivity and secondary being accuracy.

6.4.2. Key Factors

The project's main purpose was to identify the key variables which will act as the moment of impact in bank churn. However, CART and Logistic models classify the importance of variables differently. In the CART model, the variable importance refers to how well the specific variable

can classify the predicted variable with the lowest error. In logistic regression, the significant variables are determined by their p-value.

Given that the balanced CART model was identified to be most suitable, the important variables and its importance to White Rock are summarised below.

Key factors	Importance to White Rock
Age	As suspected, older customers are more likely to exit the bank. Hence, in formulating the relevant business solutions and marketing strategies to reduce churn, ensure that they are relevant to this particular demographic. Furthermore, trying to retain older customers would also have a higher return on value given that they are likely to have higher financial stability.
Number of Products	Customers who own three or more products are also more likely to leave. Hence, White Rock can target a tiered loyalty programme based on the number of products they own such that this group would be enticed to stay.
Credit Score Per Age	For this key variable, White Rock needs to make a cost-benefit analysis on the individuals with low credit score per age and determine if these customers' LTV is worth the credit risk that they pose as an individual.
Normalised Balance	White Rock can plan international programmes and marketing campaigns easily. This is vital since White Rock has 32 markets globally, with 8 in Asia; there is likely to be a disparity in the Balance amount when comparing Western countries against their Asian counterparts.
Balance Per Product	Again, although it is a surrogate split, Balance per product seems to be a better predictor of churn than Balance by itself. White Rock can use this ratio to identify the market segment that is likely to leave. It is also in White Rock interest to retain them as they are likely to be high-valued clients.

Table 8: Key variables and 'moments of impact' of churn

7. Implementation

7.1. Limitations and Future Works

1) As hinted earlier, the dataset itself has its limitations of being a point-in-time dataset and that it lacks variables that could be potential factors as to why customers exit, namely the type of products owned and the market condition. As such, the results are based on incomplete information and the model's results could be improved with the existence of additional information.

2) Our main evaluation criteria for the CART model was to achieve a high sensitivity rate to identify more clients that might exit. Although the adjustments of the trainset to balance the number of cases resulted in a higher sensitivity rate of **67.4**%, the accuracy rate also decreased from 84.9% in the unbalanced trainset to **80.7%** in the balanced trainset. The predictive accuracy and the sensitivity of the model can be improved by applying more advanced machine learning models such as bagging, boosting and random forests. Even Neural Networks could be applied in the identification of the clients who are more likely to leave. However, the trade-off here is complexity against accuracy which should be weighed carefully.

3) A lot of the customers who did not exit had a balance of 0. Although the model may be able to predict who leaves or stays, it is unable to quantify the value of the clients

who stay. Conversely, this highlights the need to retain the customers with high value as those who left are mainly high-value clients with a bigger balance. Hence, with the presence of more information such as timestamp on a client's entry and exit, a different model that predicts customer's LTV could be developed. This would also help White Rock identify which clients are worth keeping.

7.2. Recommendation

1) Assuming that White Rock's dataset has the same variables as the dataset used to train the model, the team will apply the trained CART model to identify whether the current White Rock's clients will leave or stay. The team recognises that there is also a need to test for the accuracy and sensitivity of our model based on the previous financial years' records. If the accuracy and sensitivity decrease significantly, the model needs to be re-trained.

2) White Rock can then engineer a "client retention program" that can target the customers who are predicted to leave. Besides, they can also bin these clients into groups based on their "value" (as denoted by the Number of Products, Balance per Product and Normalised Balance) and have tiered client loyalty programs based on the bin.

3) Subsequently, A/B testing can be conducted on the identified customers by splitting them into these groups:

- **a)** those with no client retention program (control group),
- **b)** those with client retention program but are not tiered, and
- **c)** those with client retention programs and are tiered

A/B testing (also referred to as split testing) is a method of comparing two versions (or more) of a program to identify the change in the control variable. A/B testing allows White Rock to pinpoint which changes affected their clients' behaviours and which did not (Optimizely, 2020). Hence, White Rock can adapt their strategies and programs accordingly.

4) The number of products seems to be the second most important variable for client churn. Therefore, White Rock could retain client with a higher number of products by increasing the switching cost while making the registration cost for their products cheap (Kamalaratnam, 2020). This way, clients are deterred from switching. However, this should be done sparingly as it may also deter new clients from registering.

7.3. Conclusion

Though the predictive accuracy and sensitivity of the CART model may not be as high as the other complicated models (i.e. bagging, boosting or random forests), they are easier to explain and understand. Furthermore, the model evaluation of CART versus Logistic model reveals that the CART model performed better overall in terms of accuracy and sensitivity, making it a suitable model.

The feature engineering step also proved to be relatively important in finding out the key variables and possibly producing causal variables given their position in the variable importance. Moreover, variable importance enables White Rock to identify the moments of impact of customer churn. By applying the model, White Rock can take targeted actions such as loyalty programmes on clients who are predicted to churn or come up with alternate plans like increasing switching cost. The effectiveness of the action can then be measured by the LTV/CAC ratio, churn rate and customer surveys.

8. References

Afshar, V. (2017, December 06). *50 Important Customer Experience Stats for Business Leaders*. Retrieved from: https://www.huffpost.com/entry/50-important-customer-exp_b_8295772

Alex. (2015, November 20). *The Three Leading Causes of Customer Churn*. Retrieved from: https://www.retently.com/blog/three-leading-causes-churn/

Alex. (2019, February 28). *12 Proven Tactics to Increase Your Customer Lifetime Value (CLV).* Retrieved from: https://www.retently.com/blog/increase-customer-lifetime-value/

Bhatt, V. (2020, April 16). *The Significance of Data Cleansing in Big Data*. Retrieved from: https://aithority.com/guest-authors/the-significance-of-data-cleansing-in-big-data/#:~:text=Data%20cleansing%20or%20scrubbing%20or,data%20are%20into%20the%20picture

Campbell, P. (2017, June 13). *How B2B And B2C Companies Solve Churn Differently*. Retrieved from: https://www.profitwell.com/recur/all/b2b-b2c-churn

Corporate Finance Institute. (2017, June 17). *CAC LTV Ratio - Customer Acquisition Cost and Customer Lifetime Value*. Retrieved from: https://corporatefinanceinstitute.com/resources/knowledge/valuation/cac-ltv-ratio/

Customer Churn Rate: KPI example. (2020, October 2). Retrieved October 22, 2020, from https://www.geckoboard.com/best-practice/kpi-examples/customer-churn-rate/

Eley, J. H. (2020, April 09). *Point by Point: Challenges in the investment management industry*. Retrieved from: https://www.thegoldensource.com/challenges-investment-management-industry-point-by-point/

Frankenfield, J. (2020, May 25). *Churn Rate.* Retrieved from: https://www.investopedia.com/terms/c/churnrate.asp

Ganti, A. (2020, February 25). *Asset Management.* Retrieved from: https://www.investopedia.com/terms/a/assetmanagement.asp

Hull, P. (2013, December 06). *Don't Get Lazy About Your Client Relationships*. Retrieved from https://www.forbes.com/sites/patrickhull/2013/12/06/tools-for-entrepreneurs-to-retain-clients/

James L. Heskett, W. Earl Sasser and Leonard A. Schlesinger (1997). *The Service Profit Chain*, 1997).

Juneja, P. (2010, July 14). *Customer Relationship Management*. Retrieved from: https://www.managementstudyguide.com/customer-relationship-measurement.htm

Kamalaratnam, J. (2020, May 11). *Retention Strategies in Retail Banks - Xerago Blog*. Retrieved October 23, 2020, from https://www.xerago.com/blog/retention-strategies-in-retail-banks/

Karnes, K. (2020, July 30). *Customer Lifetime Value: What is it and How to Calculate*. Retrieved from: https://clevertap.com/blog/customer-lifetime-value/

MarketingMind. (2018, October). Customer Satisfaction and Employee Satisfaction. Retrieved from: https://www.ashokcharan.com/Marketing-Analytics/~cs-customer-and-employee-satisfaction.php#E6.2

Menafn. (2020, September 27). *Why customer retention is vital for the growth and success of a business*. Retrieved from: https://menafn.com/1100863680/Why-customer-retention-is-vital-for-the-growth-and-success-of-a-business

Monetary Authority of Singapore. (2018). *2018 Singapore Asset Management Survey The Gateway To Asset Management Opportunities In Asia.* Retrieved from: https://www.mas.gov.sg/-/media/MAS/News-and-Publications/Surveys/Asset-Management/Singapore-Asset-Management-Survey2018.pdf

O'Neill, K. (2018, September). *Solving the 4 Key Challenges Facing Asset Managers.* Retrieved from: https://www.fenergo.com/resources/blogs/challenges-facing-asset-managers.html

PricewaterhouseCoopers. (2020). *Asset Management 2020: A Brave New World.* Retrieved from: https://www.pwc.com/gx/en/industries/financial-services/asset-management/publications/asset-management-2020-a-brave-new-world.html

Rouse, M. (2008, March). *What is garbage in, garbage out (GIGO)?* Retrieved from: https://searchsoftwarequality.techtarget.com/definition/garbage-in-garbage-out

Sasser, W. E., Schlesinger, L. A., & Heskett, J. L. (1997). *Service Profit Chain*. Riverside: Free Press.

The True Cost of Customer Churn - Part 1. (2018, January 10). Retrieved from https://www.clientsuccess.com/blog/true-cost-customer-churn-part-1/

Weavee. (2013, October 30). *The other sides of finance: asset management.* Retrieved from: https://www.weavee.co/articles/asset-management/introducing-asset-management/the-other-sides-of-finance-asset-management

Wertz, J. (2018, September 13). *Don't Spend 5 Times More Attracting New Customers, Nurture the Existing Ones.* Retrieved from: https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/

9. Appendix

Appendix A

No.	Variable (Bank Context)	Variable (White Rock Context)
1	Exited	Information on whether a particular client is still with White Rock
2	EstimatedSalary	Net Worth of a particular client
3	IsActiveMember	Level and duration of patronage
4	HasCrCard	Credit facility
5	NumOfProducts	The number of White Rock' investment products that were purchased by a particular client
6	Balance	Client's assets under White Rock
7	Tenure	Duration in which a client has been with White Rock
8	Age	Age of the individual clients Age of the institutional clients
9	Gender	Gender of the individual clients
10	Geography	Country of origin of White Rock's clients
11	CreditScore	Credit rating of White Rock's clients
12	Surname	Name of White Rock's clients
13	CustomerId	The Unique Entity Number (UEN) of White Rock's institutional clients
14	RowNumber	Row Number

Table 9: Mapping of the Variables from dataset to White Rock's context

Appendix B

		0 (N=7963)	1 (N=2037)	Total (N=10000)	p value
Credit Score	Mean (SD)	651.853 (95.654)	645.351 (100.322)	650.529 (96.653)	0.07
	Range	405.000 - 850.000	350.000 - 850.000	350.000 - 850.000	
Geography	France	4204 (52.8%)	810 (39.8%)	5014 (50.1%)	<0.001
	Germany	1695 (21.3%)	814 (40.0%)	2509 (25.1%)	
	Spain	2064 (25.9%)	413 (20.3%)	2477 (24.8%)	
Gender	0	3404 (42.7%)	1139 (55.9%)	4543 (45.4%)	<0.001
	1	4559 (57.3%)	898 (44.1%)	5457 (54.6%)	
Age	Mean (SD)	37.408 (10.125)	44.838 (9.762)	38.922 (10.488)	<0.001
	Range	18.000 - 92.000	18.000 - 84.000	18.000 - 92.000	
Tenure	Mean (SD)	5.033 (2.881)	4.933 (2.936)	5.013 (2.892)	0.162
	Range	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	
Balance	Mean (SD)	72745.297 (62848.041)	91108.539 (58360.795)	76485.889 (62397.405)	<0.001
	Range	0.000 - 221532.800	0.000 - 250898.090	0.000 - 250898.090	
NumOfProducts	Mean (SD)	1.544 (0.510)	1.475 (0.802)	1.530 (0.582)	<0.001
	Range	1.000 - 3.000	1.000 - 4.000	1.000 - 4.000	
HasCrCard	0	2332 (29.3%)	613 (30.1%)	2945 (29.4%)	0.475
	1	5631 (70.7%)	1424 (69.9%)	7055 (70.5%)	
IsActiveMember	0	3547 (44.5%)	1302 (63.9%)	4849 (48.5%)	<0.001
	1	4416 (55.5%)	735 (36.1%)	5151 (51.5%)	

EstimatedSalary	Mean (SD)	99738.392 (57405.587)	101465.678 (57912.418)	100090.240 (57510.493)	0.226
	Range	90.070 - 199992.480	11.580 - 199808.100	11.580 - 199992.480	

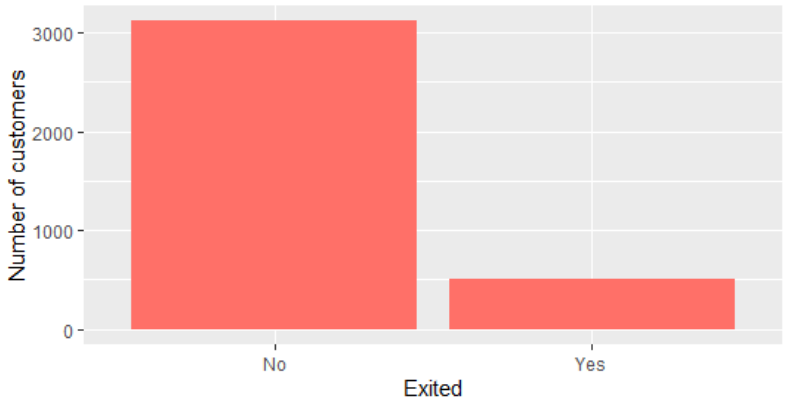
Table 10: Summary of the predictor variables against the outcome variable ("Exited")

	CreditScore	**Balance**	**EstimatedSalary**	**Age**	**Tenure**	**NumOfProducts**
Min	350.0	0	11.58	18.00	0.000	1.00
1Q	584.0	0	51002.11	32.00	3.000	1.00
Median	652.0	97199	100193.91	37.00	5.000	1.00
Mean	650.5	76486	100090.24	38.92	5.013	1.53
3Q	718.0	127644	149388.25	44.00	7.000	2.00
Max	850.0	250898	199992.48	92.00	10.000	4.00
	Gender	**HasCrCard**	**IsActiveMember**	**Exited**		**Geography**
No	4543	2945	4849	7963	**France**	5014
Yes	5457	7055	5151	2037	**Germany**	2509
					Spain	2477

Table 11: Summary of all the variables

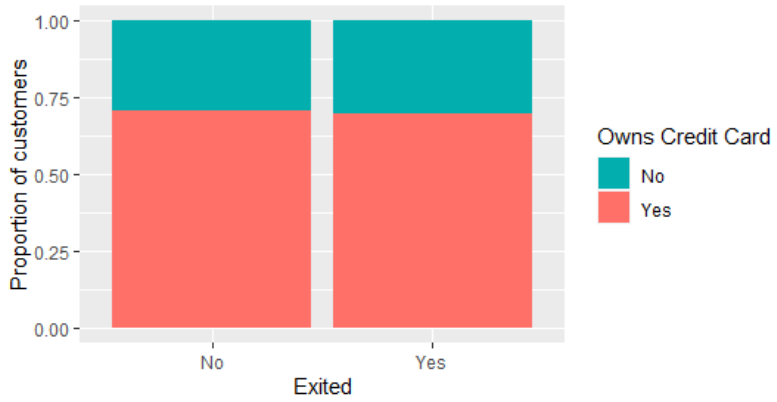
Appendix C

The following graph shows the imbalance in the proportion of outcome variables “Exited”. This needs to be taken into account when plotting graphs and also in training the model.



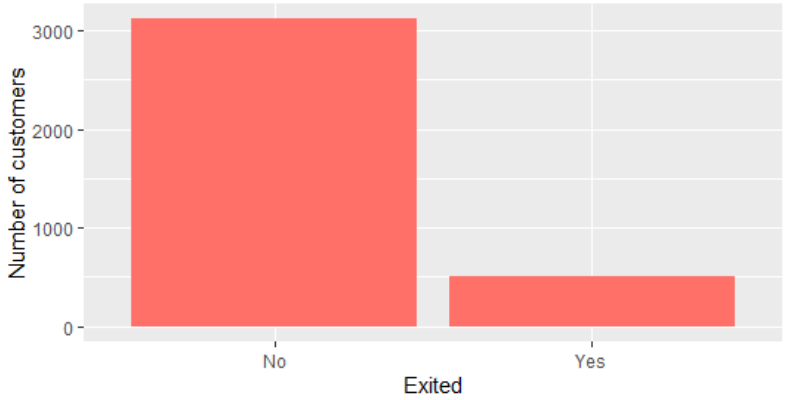
Graph 13: Unbalanced proportion of outcome variables “Exited"

The following graph shows that there is no correlation between HasCrCard and Exited.



Graph 14: Correlation between HasCrCard and Exited

The following graph shows the imbalance in the proportion of outcome variables “Exited” when the balance of the customer is 0.

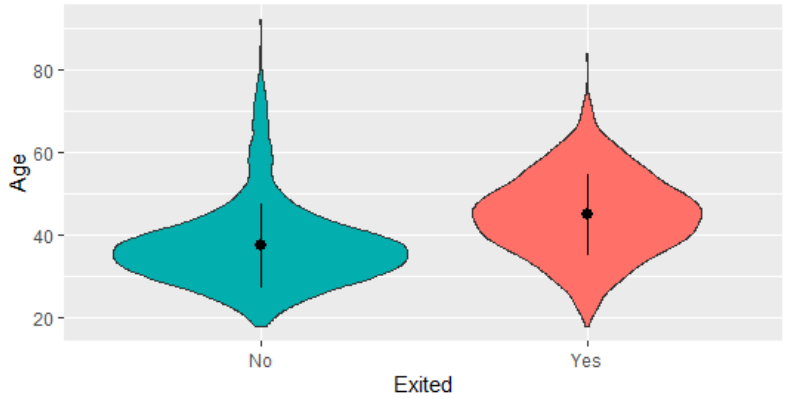


Graph 15: Unbalanced proportion of outcome variables “Exited” when Balance is 0

The following graph shows that the majority (94.89%) of customers above 65 (after conditional removal of 0 Balance) that stayed are active members.

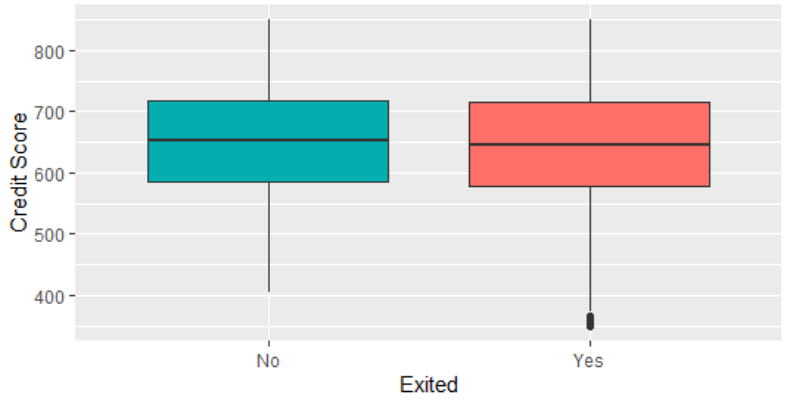
Graph 16: Impact of Active Member on Exited for customers aged above 65

The following graph shows the age of the customer displays a correlation with whether or not the customer will exit.



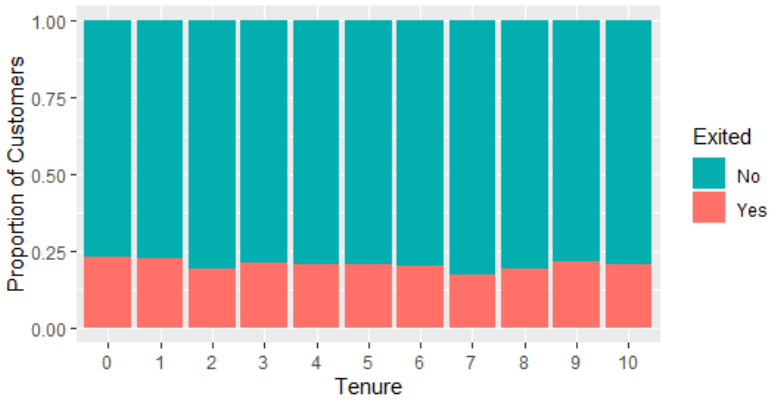
Graph 17: Impact of Age on Exited

The following graph shows that Credit Score has little to no correlation with whether or not a customer exit.



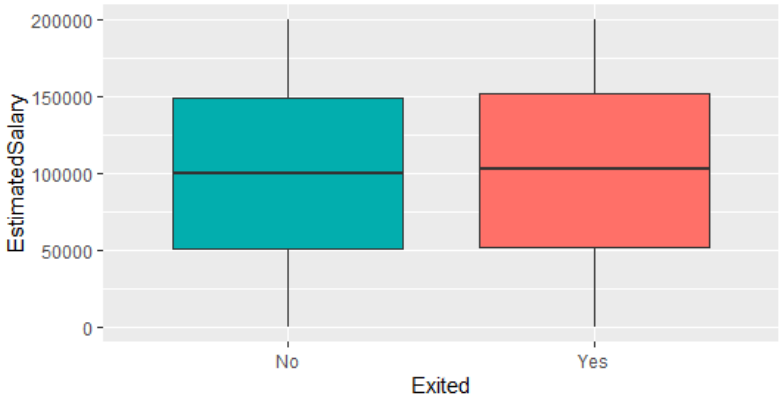
Graph 18: Impact of Credit Score on Exited

The following graph shows that tenure has little to no correlation with whether or not a customer will exit.



Graph 19: Impact of Tenure on Exited

The following graph shows that a customer's estimated salary has little to no correlation with whether or not a customer will exit. However, it should be noted that a deeper analysis was conducted as the team feels that clients with high estimated salary will represent the high-value clients and is worth investigating.

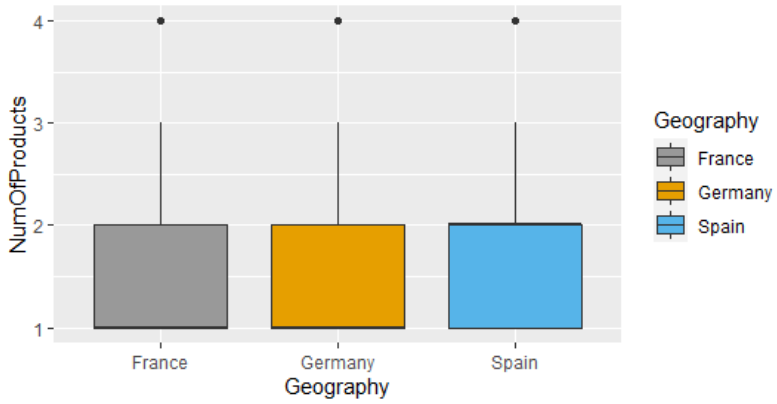


Graph 20: Impact of Estimated Salary on Exited

Appendix D

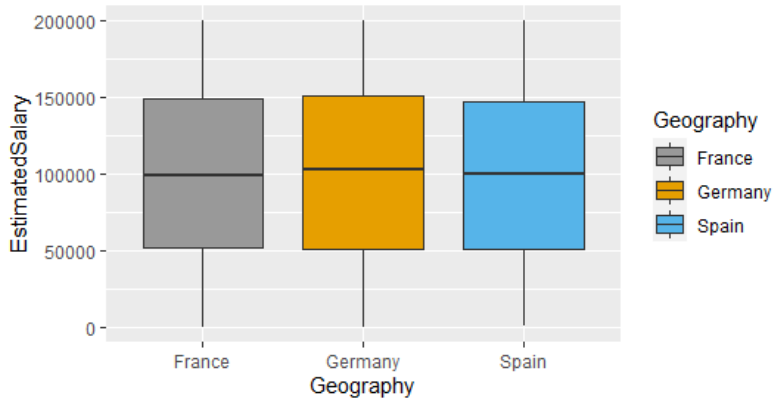
This section includes graphs from the variable specific analysis that produced little to no insight.

The following graph shows the distribution of the NumOfProducts against Geography. Although Spain had the highest median for 'NumOfProducts, Spain was not the Geography with most people leaving the bank. Hence the insight was irrelevant.



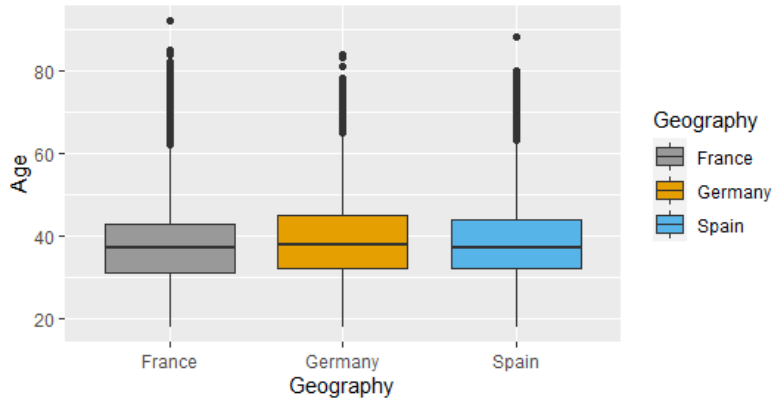
Graph 21: Distribution of the number of products against Geography

The following graph shows the distribution of EstimatedSalary against Geography. There was little to no discrepancy in the distribution. Hence, the insight was insignificant.



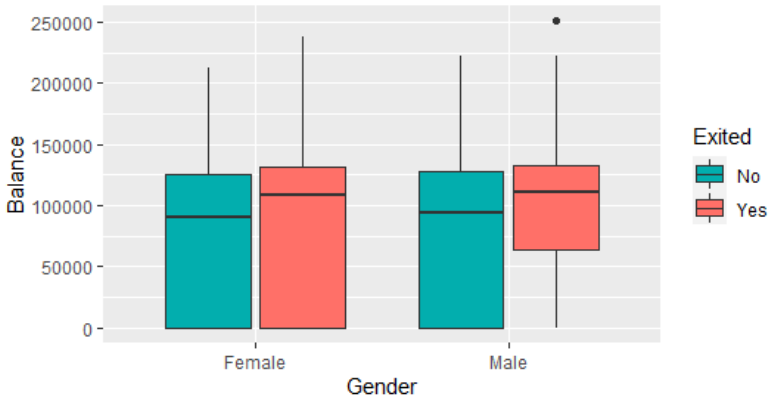
Graph 22: Distribution of Estimated Salary against Geography

Again, the plot of the distribution of Age against Geography showed little to no difference. Hence this information was omitted.



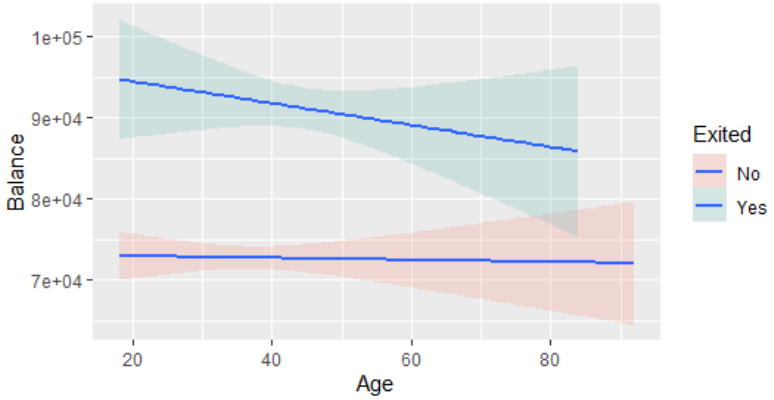
Graph 23: Distribution of Age against Geography

For Gender specific analysis, distribution of Balance by Gender on Exited shows that there is no "skew" in the distribution of Balanced on Exited for females but there was an obvious skew for males. However, this was a mixed result and has little significance to our model.



Graph 24: Distribution of Balance against Gender on Exited

For Age specific analysis, the smoothed plot of Balance against Age on Exited shows that the difference between “Yes” and “No” for Exited was merely the balance.



Graph 25: Impact of Balance against Age on Exited

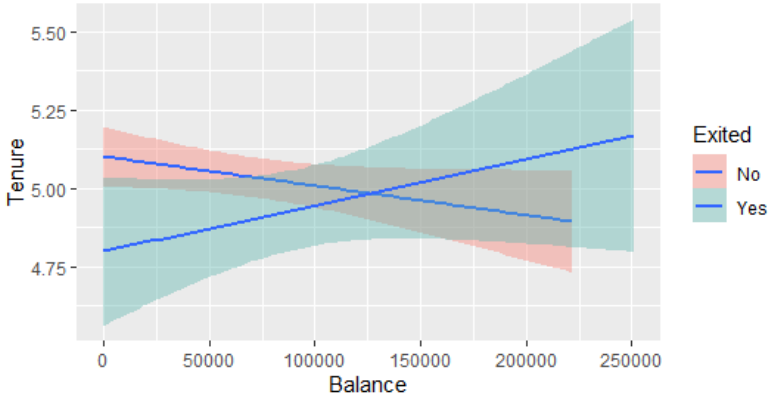
For Estimated Salary specific analysis, the point plot of the number of products against Estimated Salary shows that the ‘number of products’ was the sole factor in the prediction of Exited.

Graph 26: Impact of Number of Products against Estimated Salary on Exited

The smoothed plot of Balance against Estimated Salary on Exited also shows that Balance was the sole factor for the prediction of Exited.

Graph 27: Balance against Estimated Salary on Exited

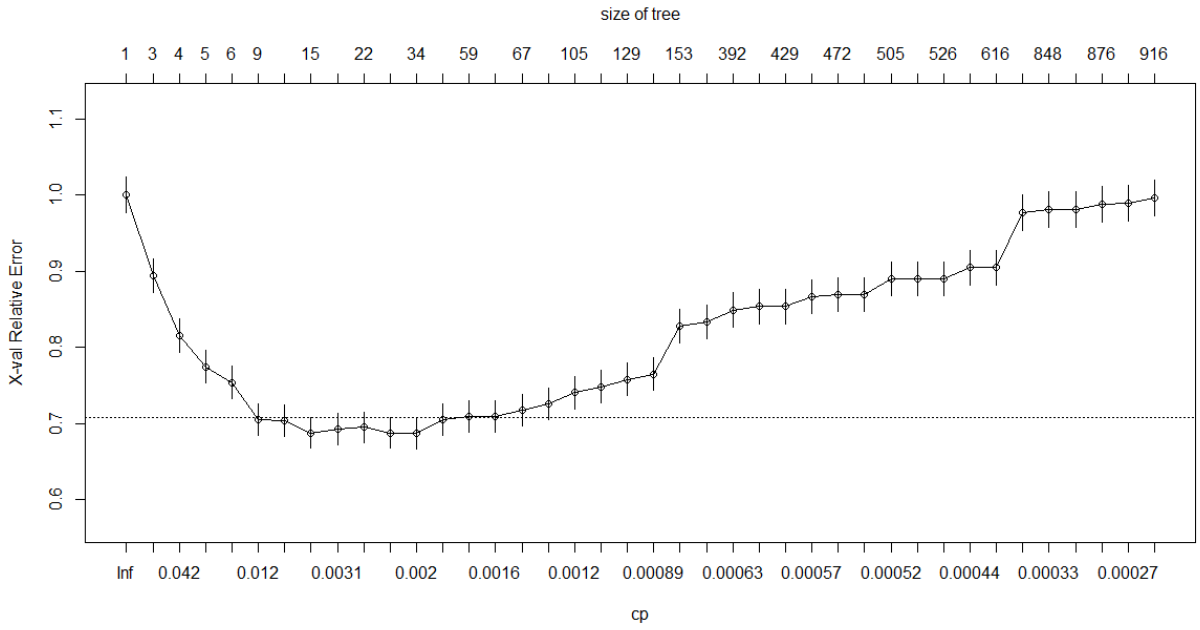
In Balance specific analysis, although there is a difference in the correlation between Tenure against Balance on Exited, there is a huge overlap between the two. Hence, the insight was not included.



Graph 28: Impact of Tenure against Balance on Exited

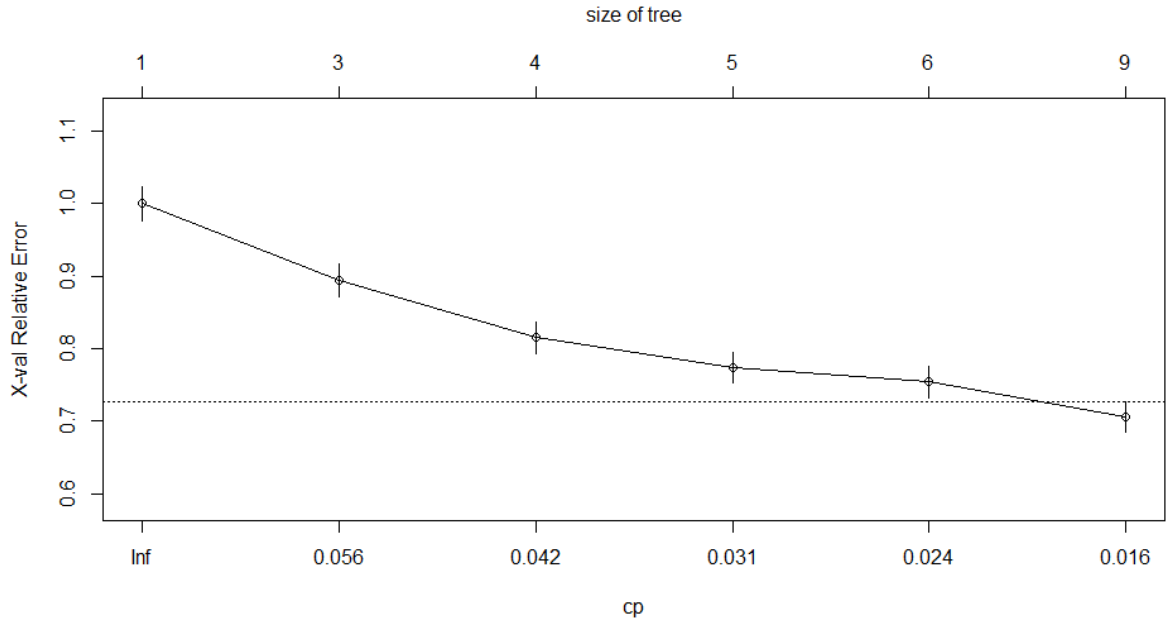
Appendix E

The following graph shows the cross-validation (CV) relation error against cp values obtained from the maximum tree that was trained on the unbalanced trainset.



Graph 29: CV Relative Error against cp for the maximum tree trained on unbalanced trainset

The following graph shows CV Relative Error against cp for the pruned tree with prune trigger of 0.0124 trained on unbalanced trainset.



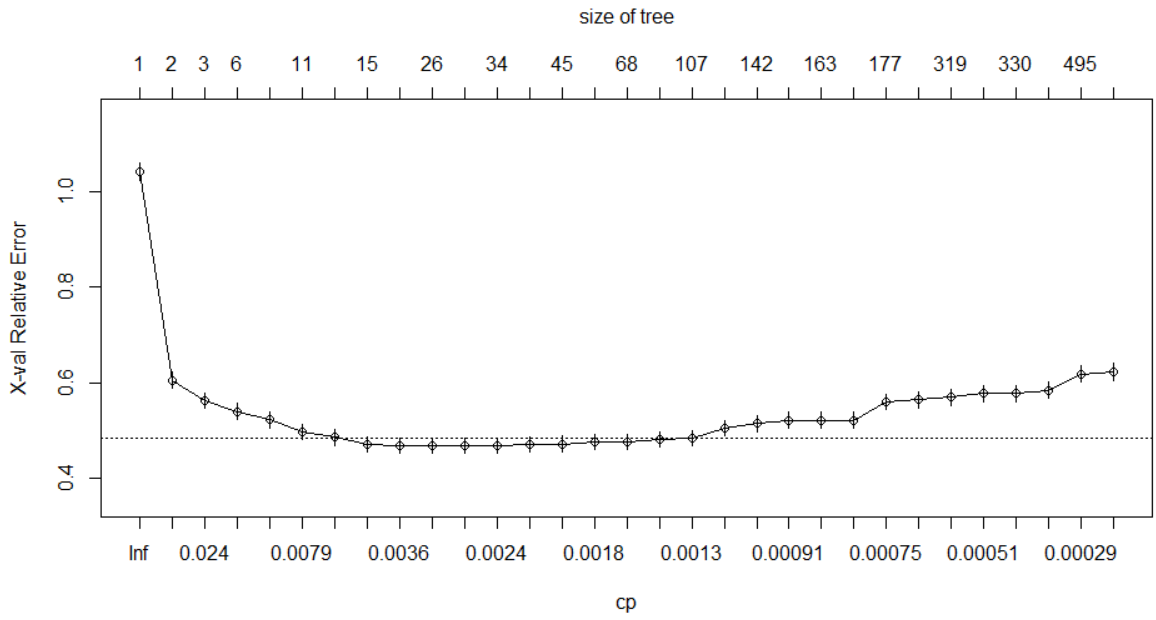
Graph 30: CV Relative Error against cp for the pruned tree trained on unbalanced trainset

The following table shows the variable importance for the pruned tree that was trained with the unbalanced train dataset.

Age	328.4470432
NumOfProducts	191.4582953
creditScorePerAge	160.0799193
IsActiveMember	118.8775349
balancePerProduct	28.3887948
normalisedBalance	25.2634176
Balance	16.8282073
Geography	11.9919414
CreditScore	2.5689522
Tenure	2.3208196
EstimatedSalary	1.4663283
salaryCreditScoreRatio	0.3217469

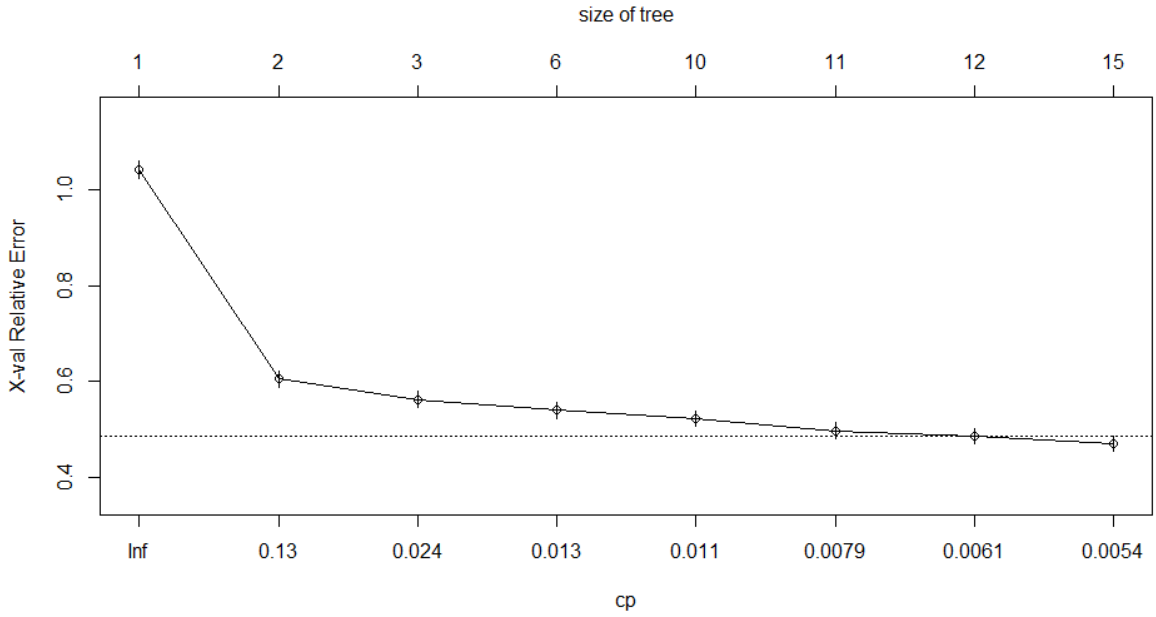
Table 12: Variable Importance of the independent variables for the pruned tree trained on unbalanced trainset

The following graph shows the cross-validation (CV) relation error against cp values obtained from the maximum tree that was trained on the balanced trainset.



Graph 31: CV Relative Error against cp for the maximum tree trained on balanced trainset

The following graph shows CV Relative Error against cp for the pruned tree with prune trigger of 0.00496 trained on balanced trainset.



Graph 32: CV Relative Error against cp for the pruned tree trained on balanced trainset

The following table shows the variable importance for the pruned tree that was trained with the balanced train dataset.

Age	247.379811
NumOfProducts	182.068084
creditScorePerAge	149.248142
normalisedBalance	103.687821
balancePerProduct	101.800559
Balance	69.340419
IsActiveMember	37.217649
Geography	33.451861
EstimatedSalary	6.791583
CreditScore	3.725768
salaryCreditScoreRatio	2.566652
Tenure	1.326127

Table 13: Variable Importance of the independent variables for the pruned tree trained on balanced trainset

The following image depicts the confidence interval of all the independent variables in the initial logistic model trained on the unbalanced trainset.

```
> OR.CI
                                2.5 %    97.5 %
(Intercept)                0.03172966 0.3497212
CreditScore                1.00421741 1.0077516
GeographyGermany           1.54672098 4.4411861
GeographySpain             0.94508342 1.3153279
GenderMale                 0.53182689 0.6876177
Age                        0.99205812 1.0290409
Tenure                     0.97347008 1.0175299
Balance                    0.99998963 1.0000082
NumOfProducts              0.76361764 1.1093838
HasCrCardYes               0.89162015 1.1827736
IsActiveMemberYes          0.29328306 0.3854319
EstimatedSalary            0.99998321 0.9999960
normalised_balance         0.65779602 2.0211835
normalised_creditscore     0.78872511 0.8751599
balance_per_product        0.99999769 1.0000061
salary_credit_score_ratio 1.00286941 1.0108868
```

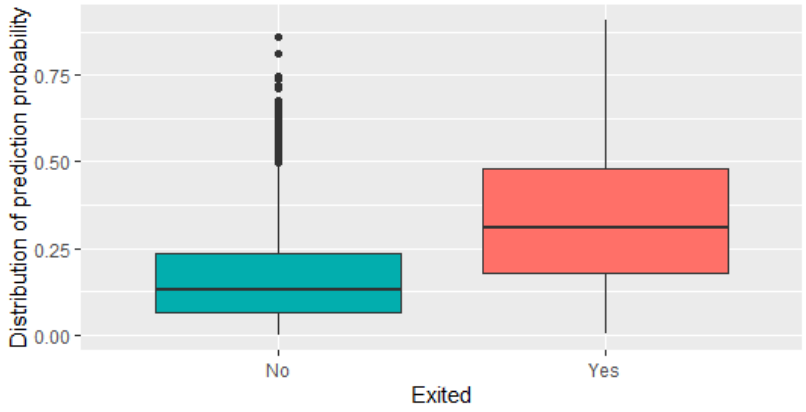
Image 1: Confidence interval of all the variables in the unbalanced trainset

The following image depicts the confidence interval of the significant independent variables in the optimal logistic model that was trained on the unbalanced trainset. It can be observed that none of the confidence intervals contains 1.

```
> OR.CI
                                2.5 %    97.5 %
(Intercept)                0.06272569 0.3258802
CreditScore                1.00536649 1.0079870
GeographyGermany           1.98758535 2.6985423
GeographySpain             0.94237487 1.3107234
GenderMale                 0.53285359 0.6887196
IsActiveMemberYes          0.29785883 0.3895099
EstimatedSalary            0.99998319 0.9999960
normalised_creditscore     0.79392248 0.8229984
balance_per_product        1.00000224 1.0000046
salary_credit_score_ratio 1.00289920 1.0109016
```

Image 2: Confidence interval of all the significant variables in the unbalanced trainset

The following graph shows the distribution of the probability of customer exiting on 0.5 threshold for unbalanced trainset.

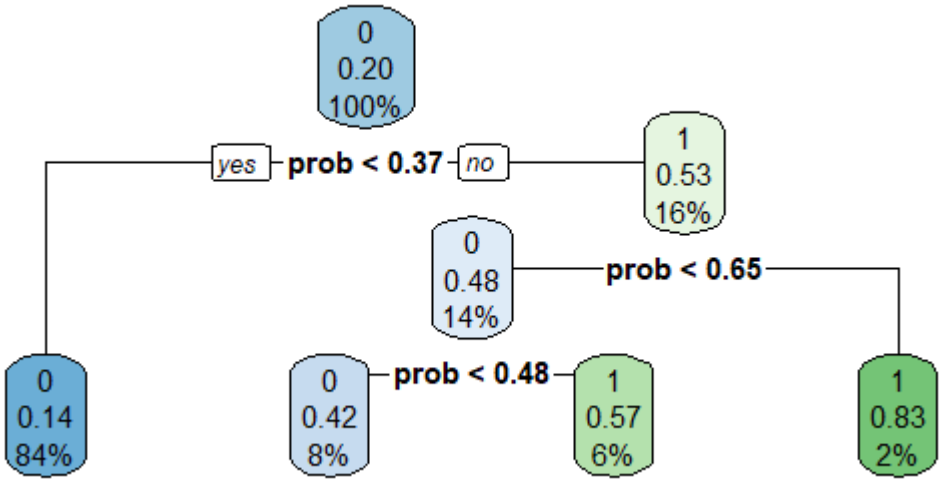


Graph 33: Probability of customer exiting for unbalanced trainset with 0.5 threshold

How the optimal threshold was selected:

To identify the optimal threshold, the team ran a CART model using the predicted probability values that were obtained from the logistic model to predict the exited cases. The tree was grown to its maximum and subsequently pruned using the one standard deviation rule to obtain the optimal tree. The pruned tree will classify the exited cases using its predicted probability values. The optimal threshold for the logistic model was then set as the first split in the CART model. By using this approach, the optimal threshold can be identified.

The following plot shows the identification of the optimal threshold for the logistic model trained on unbalanced trainset.



Graph 34: CART model to identify the optimal threshold for the logistic model trained on unbalanced trainset

The following image depicts the confidence interval of all the independent variables in the initial logistic model trained on the balanced trainset.

```
> OR.CI
                                  2.5 %     97.5 %
(Intercept)                  0.01632919  0.3675256
CreditScore                  1.00139235  1.0058972
GeographyGermany             0.89028238  4.6336460
GeographySpain               0.89166160  1.3583789
GenderMale                   0.54630802  0.7628920
Age                          1.02129131  1.0740314
Tenure                       0.97030601  1.0277436
Balance                      0.99998964  1.0000185
NumOfProducts                0.70925052  1.0573963
HasCrCardYes                 0.82559495  1.1958864
IsActiveMemberYes            0.32370966  0.4548181
EstimatedSalary              0.99998260  0.9999995
normalised_balance           0.41171417  2.3902681
normalised_creditscore       0.86308003  0.9741015
balance_per_product          0.99999436  1.0000045
salary_credit_score_ratio    1.00099319  1.0115591
```

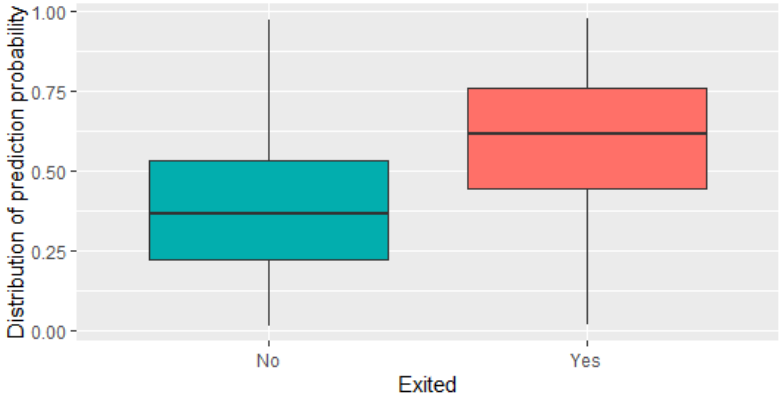
Image 3: Confidence interval of all the variables in the balanced trainset

The following image depicts the confidence interval of the significant independent variables in the optimal logistic model that was trained on the balanced trainset. It is observed that none of the confidence interval contains 1.

```
> OR.CI
                                  2.5 %    97.5 %
(Intercept)                  0.01622758 0.3394242
CreditScore                  1.00140191 1.0059014
GeographyGermany             1.65680861 2.5322705
GeographySpain               0.89337198 1.3597904
GenderMale                   0.54630036 0.7626162
Age                          1.02127401 1.0739605
Balance                      1.00000193 1.0000050
NumOfProducts                0.77798692 1.0003303
IsActiveMemberYes            0.32400992 0.4549076
EstimatedSalary              0.99998258 0.9999994
normalised_creditscore       0.86301364 0.9739790
salary_credit_score_ratio    1.00101633 1.0115698
```

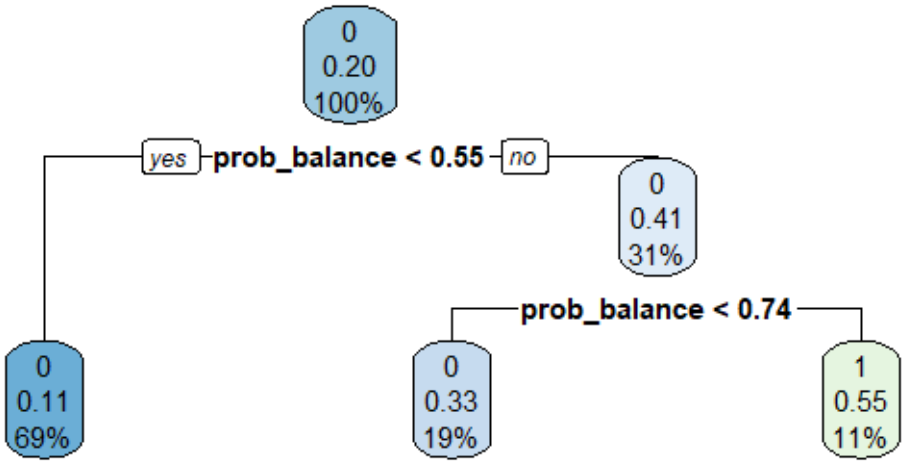
Image 4: Confidence interval of all the significant variables in the balanced trainset

The following graph shows the distribution of the probability of customer exiting on 0.5 threshold for balanced trainset.



Graph 35: Probability of customer exiting for balanced trainset with 0.5 threshold

The following plot shows the identification of the optimal threshold for the logistic model trained on balanced trainset.



Graph 36: CART model to identify the optimal threshold for the logistic model trained on balanced trainset