

BC2406 ANALYTICS I: VISUAL AND PREDICTIVE TECHNIQUES

SEMESTER 1 AY 2020/2021

SEMINAR GROUP 2 TEAM 8

|  |  |
| --- | --- |
| Name | Matriculation No. |
| WANG WAYNE | U1920270G |
| CHEE WEI KIAT COLIN | U1910457D |
| NEO SAY YIN FLORENCE | U1910346G |
| DALINI VICKTORIA RAJANDHRAN | U1710158A |

# 

**Contents**

[Executive Summary 4](#_Toc55131662)

[Background 5](#_Toc55131663)

[1. Defining the Business Problem 5](#_Toc55131664)

[1.1 Importance of Client Retention 5](#_Toc55131665)

[1.2 Difficulties of Client Retention 6](#_Toc55131666)

[1.3 Important Trends 7](#_Toc55131667)

[1.3.1 Changing Consumer Patterns 7](#_Toc55131668)

[1.3.2 Changing Demographic Patterns 7](#_Toc55131669)

[1.4 Costs of inefficient client retention 8](#_Toc55131670)

[2. Opportunities to improve client retention 8](#_Toc55131671)

[2.1 Churn rate 8](#_Toc55131672)

[2.2 Client Retargeting 9](#_Toc55131673)

[3. Proposed Solution 9](#_Toc55131674)

[3.1 Preparation of Datasets 11](#_Toc55131675)

[3.2 Predictive Techniques and their Basis of Comparison 12](#_Toc55131676)

[3.3 Churn Model 13](#_Toc55131677)

[3.3.1 Churn Model - Logistic 14](#_Toc55131678)

[3.3.2 Churn Model - CART 15](#_Toc55131679)

[3.3.3 Churn Model - Conclusion 16](#_Toc55131680)

[3.4 Product Acceptance Model 16](#_Toc55131681)

[3.4.1 Product Acceptance Model - Logistic 16](#_Toc55131682)

[3.4.2 Product Acceptance Model - CART 18](#_Toc55131683)

[3.4.3 Product Acceptance Model - Conclusion 18](#_Toc55131684)

[4. Limitations of Proposed Solution 19](#_Toc55131685)

[4.1 Imbalanced Datasets 19](#_Toc55131686)

[4.2 Unable to Consider Qualitative Reasons 20](#_Toc55131687)

[5. Optimisation of the Client Retention Priority Model 20](#_Toc55131688)

[6. Conclusion 21](#_Toc55131689)

[References 22](#_Toc55131690)

[Appendix A: Datasets and Data Dictionary 23](#_Toc55131691)

[Appendix B: Data Exploration for Churn Dataset 27](#_Toc55131692)

[Appendix C: Data Cleaning for Personal Loan Dataset 29](#_Toc55131693)

[Appendix D: Data Exploration for Personal Loan Dataset 30](#_Toc55131694)

[Appendix E: Churn Logistic Model 31](#_Toc55131695)

[Appendix F :Product Acceptance Logistic Model 36](#_Toc55131696)

[Appendix F-3: Optimal Product Acceptance CART Model 40](#_Toc55131697)

[Appendix F-4: Variable Importance for Product Acceptance Model 40](#_Toc55131698)

# Executive Summary

An average 65% of a company’s revenue comes from existing clients and a 5% increase in customer retention leads to a 25%-90% increase in revenue. Hence, it is of utmost importance to improve White Rock’s client retention strategies. This report explains the general difficulties with client retention and provides solutions to decrease churn rate and improve tenure with the organization while keeping revenue in mind.

In recent years, there has been strong cost cutting pressures within asset management institutions. From consumers turning to firms which can offer them the lowest processing fees, to a huge increase in regulatory costs, profit margins have thinned. To curb this problem, it is important that companies optimise their resource allocation. To create the best value for themselves, firms should identify areas where their invested resources can bring the highest returns.

With recent technological advancements in machine learning and artificial intelligence, data can be collected to construct insightful solutions to track and predict whether clients will churn and improve churn rate through predictive analysis on variables affecting churn. We used two datasets from Kaggle: Bank Churn and Personal Loan Classification. Our proposed solution (Client Retention Priority Model) integrates the two models to give White Rock an avenue to prioritize which clients in their current client base to better allocate resources and efforts for retention. This model is designed to determine a client’s prioritisation category (High Priority / Low Priority / No Priority) when it comes to allocating resources. However, there are limitations to the effectiveness of the Client Retention Priority Model. Firstly, the model is subjected to biases from the datasets used, giving rise to poor decision making. Secondly, the model is unable to consider qualitative factors such as when a client goes through a critical life event, which evidently is a strong predictor of churn.

The implementation of the proposed model not only aims to provide a long-term solution but to redefine traditional strategies of client retention through the use of data analytics. The Client Retention Priority Model seeks to mine the huge amount of internal data collected by White Rock from its 32 markets to ultimately recommend which clients to allocate significantly more resources to and which clients to allocate minimal resources to base on the individual client’s priority level. This saves monetary and non-monetary resources which can be better allocated in other areas of the business while increasing client retention and profitability of the company.

# Background

In recent years, the face of the Asset Management Industry has seen many disruptions (Waite, Massa, & Cannon, 2019). With the uprise in FinTech, consumers are spoiled for choice with more options now than ever before. Many are shifting their investments to cheaper index funds where they can enjoy similar margins at a fraction of the processing fees. To make things worse, Asset Wealth Managers are faced with increasing costs from complying with stricter regulatory requirements introduced in recent years. The trend of thinning profit margins is worrying. The decrease in market share, coupled with increased costs have forced the organisations to seek innovation through technology and turn to data-driven insights to provide better returns for their consumers. Currently, there are many asset management analytics tools in the market which are focused on generating greater revenue for the company and returns for consumers.

# 1. Defining the Business Problem

Being an incumbent in the market, White Rock (WR) has amassed a large consumer base through the years. As a revenue-driven organisation, it is extremely vital for them to secure their market share. However, retaining more clients and cutting costs while doing so through better resource allocation is equally important in ensuring healthy profit margins. We believe that this is an area to be explored. To achieve this, minimising client churn rate with efficient resource allocation and to increase client tenure is a potential solution. Streamlining and simplifying client retention processes will also minimise touch points across departments, allowing for a reduction in processing costs and time lags due to miscommunications. This will lead to cost-savings and an overall improvement in consumer experience (Jones, n.d.).

## 1.1 Importance of Client Retention

Currently, many companies are focused on client acquisition to achieve growth. Over the past five years, asset management institutions have increased their Sales and Marketing efforts by over 50% to achieve this. While this remains the main cost driver, the additional dollar generated from such efforts is not increasing proportionately.

On the contrary, most organisations underestimate the value client retention efforts can bring. It is often under-emphasized, if not neglected. As [65% of a company’s business](https://smallbiztrends.com/2016/10/customer-retention-statistics.html) (Mansfield, 2016) comes from existing clients, it’s important to focus on the clients who are already engaged with the company. Maintaining a pool of loyal clients can drive sustainability and growth in the following ways:

1. Satisfied clients are likely to recommend the products and services to their friends. Word-of-mouth referral is deemed the most cost-efficient marketing strategy as the company does not incur additional marketing costs in the process. Furthermore, consumers are more willing to try a new product when recommended by a friend, as there is a element of trust and the product has been tried and tested.
2. Existing clients are likely to open accounts and take up products on behalf of their offsprings. In the future, the new generation would inherit these accounts and are likely to continue investing with the company. WR stands to benefit from having a ready batch of clients who they can also tap on to reach out to the younger generations who are fresh to asset management products, as opposed to directly competing with the smaller Fintechs to capture the attention of the younger consumers.
3. Lastly, loyal clients understand the value of the products and services of WR. Thus, they are likely to make repeat purchases at premium prices. According to Gartner (Customer Experience: Increase customer loyalty and retention with smart CX strategy, 2020), there is a higher chance (60% to 70%) of selling new products to an existing client while the probability of selling to a new client is only 5% to 20%. They trust WR to make the best decisions for them and are less skeptical when investing in new products. Furthermore, they are open with their feedback and this is a valuable resource for the company to understand areas of improvement.

## 1.2 Difficulties of Client Retention

In today’s complex world, retaining clients is a challenging task and offering clients a great experience alone, will not help accomplish this task. In order to better provide for our clients and to improve retention rates, we need to identify triggers and predict the probability of clients leaving.

Often, companies face strong competition - from companies who can seemingly provide more to consumers. Companies that are able to deliver products which effectively meet consumers’ needs will have a clear competitive edge. This is visible by the shift in consumer preference to smaller Fintech companies (2019 Global Wealth Management, 2019). They are able to provide personalisation of services and give consumers more transparency and control over their investments. The smaller consumer base plays to their advantage as they are able to have close continual engagements with their clients. This gives them an edge in identifying changing needs in the market and designing suitable products ahead of their bigger competitors.

For big players like WR, they tend to be on the losing end when it comes to client engagement as it is difficult and costly to establish such a connection with each of their existing clients. To improve this situation, we can identify clients to engage with and understand their concerns. These clients are likely to be those who are on the brink of exit. We hope that identifying this group early will give us enough time to engage with them and understand their concerns and needs. Through addressing their concerns, we hope to be able to retain them for a longer period.

## 1.3 Important Trends

### 1.3.1 Changing Consumer Patterns

Based on a recent study conducted by EY, it is projected that one-third of wealth management clients will switch providers in the next three years (Global, 2019 ). Firms are likely to see the departure of 39% of Ultra-high-net-worth (UHNW) clients, who are seeking to diversify their products (Vincent, Goldstein, & Cunniff, n.d.). This is an area of concern as investors are bringing their wealth away. Once they have transferred their assets away, it will be difficult to convince them to return. Although with this shift, we can expect to acquire new clients, it will pose a costly process as WR has to go through many processes to understand the risk profiles and needs of the new consumers. Not forgetting having to file heaps of papers to adhere to regulatory requirements. Thus, it is now more important than ever to be able to identify this group of clients and reinvest efforts to retain them.

### 1.3.2 Changing Demographic Patterns

In the coming years, many heirs will inherit their wealth from the generation of baby boomers. In the US alone, $58.1 trillion in assets is estimated to be passed down to the next generation from 2007 to 2061. Past trends have shown that 90% of heirs are likely to use a different wealth management advisor from their parents (Vincent, Goldstein, & Cunniff, n.d.). The key to prevent the younger generation from shifting their assets away is to build multi-generational relationships. For a start, it is important to ensure that client retention efforts with the parent generation are sustained. We should continually engage with them to prevent them from churning before they pass down their assets. Such engagements give us an avenue to reach out to their children, giving us an upper edge in retaining this young pool of investors and their inherited wealth.

### 1.4 Costs of inefficient client retention

All in all, client retention has to be a continual sustained effort. When dissatisfied clients leave, they take with them resentment and bad experiences. In most cases dissatisfied clients are unlikely to return and would actively discourage others from patronising the organisation. This brings about reputational damages which will require time and money in marketing efforts. It is no easy feat to earn back the trust of lost clients. Coupled with the current rise in competing companies and products, clients recognise that there are an abundant of alternatives for them to choose from. Thus, it may be more difficult than ever to win back this group of clients.

# 2. Opportunities to improve client retention

## 2.1 Churn rate

Client churn occurs when an existing client stops doing business with an entity. Especially in the asset management industry, with so many different options clients can choose from to invest their assets, financial institutes are increasingly competitive when it comes to trying to keep their clients continue using their services.

Retaining existing clients is very important but there is only so much financial institutions can do when clients stop using their services and as such, predicting when or why they churn becomes crucial in improving future clients’ retention. The earlier and more accurate the prediction, more costs are saved in client engagement. In fact, banks can avoid up to 11% of churn just by being more proactive with client service at an early enough time period (Sethuraman, 2019 ).

So, how exactly should we approach churn prediction? The most intuitive method is to conduct comprehensive feedback through client surveys to understand their experiences and either directly or indirectly determine whether the clients are starting to use competitor’s services instead. However, surveys are costly to implement and some clients may provide non-truthful responses resulting in response bias (Jovancic, 2019). Collecting and processing such data is a long drawn process. It is likely that by the time we are able to sense a client churning, they would have already left.

The best approach lies in using advanced data science and machine learning techniques to develop models which learn from data involving past client behaviours to find triggers which will contribute to churning and use this to predict future occurrences and odds of client churn so as to improve client retention (Malik, 2018).

## 2.2 Client Retargeting

For our purpose, client retargeting refers to the identification of current clients who we should target our marketing efforts at. Understanding the likelihood of a client churning and the possible reasons for churn will set the foundation for client retargeting. When we identify a possible pool of clients who are likely to churn, we can look at client retargeting. The idea is to retarget these high churn potential clients with new products to rejuvenate their client life cycle with the bank.

The Qualtrics Banking Report (Banking Customer Experience Program, 2019) indicated that the top 3 reasons why bank clients churn is due to ‘Poor Service’ , ‘Poor Communication’ and ‘Poor Product Match’. However, retargeting clients through selling new products allows the bank to reconnect and communicate with their clients. It also provides an opportunity for the bank to provide a better product for the client to improve product match. Also, clients perceive client retargeting methods as a form of exclusive service which addresses the main issue of why they churn.

# 3. Proposed Solution

We can develop a model (Churn Model) to identify clients who are likely to churn based on certain variables. Through filtering out these clients, sales managers can have a better view on who they should spend more resources to retarget. From the developed model coupled with expert opinion, we found “Number of Products” to be a significant variable in influencing churn. This is evident from our findings where the CART model’s variable importance for age is the highest, followed by number of products, then active membership (Appendix E-5). Furthermore, the odds ratio from the logistic regression suggests that the more number of products owned by the client, the less likely he will churn (Appendix E-1.5). Since the number of products is the variable of highest importance that is influenceable by the financial institution after age, we can spend resources promoting products to clients who are predicted to churn to induce them to act otherwise.

Keeping in mind that it is important for WR to selectively target clients for more efficient resource allocation, we further leveraged on this information and took a step further to develop a ‘Product Acceptance Model’. This model aims to predict the likelihood of one taking up a product with WR. Through this model, we can further streamline clients to identify those who are of higher priority: likely to churn but are predicted to accept additional products from WR. In turn, reducing the likelihood of them churning.

Using both models sequentially, we can derive their likelihood of churn and product acceptance. The ‘Client Retention Priority Model’ integrates both results to give WR a accurate representation on specific individuals and their priority level.

|  |
| --- |
| ***Client Retention Priority = Churn + (Churn x Acceptance)*** |

Based on whether a client will churn (Churn outcome = 1) we then consider whether they are of higher priority based on whether they will accept a new product sold to them (Acceptance outcome = 1). Hence we add the Churn outcome with the outcome of them accepting a new product and the outcome of them churning (Churn x Acceptance), as, if they will not churn (Churn outcome = 0) there is no point prioritising them.

The outcome of the model will be a score ranging from 0 to 2, which will represent the priority of the client when it comes to allocating resources for retention. Below shows the outcome of this model based on the Churn Rate and Product Acceptance Model which reflects the priority score a client may be assigned. We hope that through this model, resources can be channeled more efficiently, and returns per dollar invested will be improved.

|  |  |  |
| --- | --- | --- |
| **Churn** | **Acceptance** | **Priority Indicator** |
| 0 | 0 | **0** (No priority) |
| 0 | 1 | **0** (No priority) |
| 1 | 0 | **1** (Low priority) |
| 1 | 1 | **2** (High priority) |

Subsequently, the model priority score will be used to segment the consumers into 3 tiers. The interpretation of the scores are as such:

(i) High Priority Clients (Score = 2) a significant increase in man hours in terms of customer service and money can be allocated to retain these clients;

(ii) Low Priority Clients (Score = 1) resources may still be allocated to increase the probability of retaining such clients but WR should spend an average amount of resources as they are predicted to ultimately churn and not accept a new product; and

(iii) No Priority Clients (Score = 0) WR should allocate minimal resources in the form of business-as-usual service as they are predicted not to churn regardless of whether they accept another product.

## 3.1 Preparation of Datasets

To demonstrate the implementation of our solution, we used two datasets found on Kaggle: (1) Bank Churn and (2) Personal Loan Classification.

(1) - To develop the Churn model, we adapted the Bank Churn dataset which contains clients’ details including demographic information such as age and gender, as well as bank-related information such as balance and number of products with the bank. The target variable is a binary variable reflecting whether the client has left the bank or continues to be a client.

(2) - To develop the Product Acceptance model, we adapted the Personal Loan Classification dataset which also contains clients’ details like demographic information such as age and education level, as well as bank-related information such as whether the client uses internet banking facilities or whether the client uses a credit card issued by the bank. The target variable is a binary variable reflecting whether the client accepted the personal loan offered in the last campaign or not.

For both datasets, we prepared the necessary adjustments to the variable columns by identifying the appropriate continuous and categorical variables. We removed columns such as “customerID” and “rownumber” so that proper analysis can be done on the dataset.

In terms of data cleaning steps, we did a preliminary run to check for duplicate and missing values (Appendix B.1). The Bank Churn dataset was rather clean and no additional cleaning steps were required. For the Personal Loan Classification dataset, some data cleaning steps were required as some clients had negative values for years of professional experience. With no proper description given to address the interpretation of negative experience, we had to assume that those were erroneous rows which we performed the necessary data cleaning steps (as shown in Appendix C).

We believe that most of the variables used in both datasets can be reasonably easily accessible to any financial institutions as the onboarding process of such an organization usually requires the clients’ demographic data to be collected and institution-related data to be created. As such, our client retention strategy involving the Client Retention Priority Model can be readily applied to most asset management organizations like White Rock.

Once the data that has been cleaned, we conducted a preliminary data exploration to better understand the dataset. Using the ggplot package within R, we plotted a correlation matrix to analyse the strength of relationships between the output and various input variables.

|  |  |
| --- | --- |
| Correlation Matrix for Churn Dataset | Correlation Matrix for Personal Loan Dataset |

From the matrix of both datasets, we could preliminarily conclude that there was correlation between our selected output variable and the other variables. This was important in allowing us to confirm the suitability of both datasets in forming the foundation of our data model. Other plots and graphs analysed during this process are as presented in Appendix B (Bank Churn) and Appendix D (Personal Loan Classification).

## 3.2 Predictive Techniques and their Basis of Comparison

Though there is a wide variety of predictive techniques we could use, we chose to compare between the use of logistic regression and CART (Classification and Regression Tree) models. These two predictive techniques were chosen as they are simple, easy to understand and implement. Both predictive techniques are suitable for developing our models on Churn Rate and Product Acceptance, and we recommend WR to produce the respective models by using both predictive techniques. After a holistic comparison between the model accuracies and the practicality of implementation, WR can choose their preferred predictive technique for each model.

The general approach to make a decision between the CART and logistic model is detailed below. After obtaining the optimal models for each predictive technique, we:

1. **Compare between Logistic and CART accuracies**

Firstly, we look at the overall accuracy of each model. If one model has a significantly lower misclassification error compared to the other, we would be significantly more inclined towards the model with the higher accuracy. The next indicator would be the frequency and type of misclassification error that has occurred - lower false positives or false negatives reflects a model with higher accuracy.

1. **Consider the practicality of implementation of each model**

Each predictive model has its own strengths and weaknesses. For example, CART uses surrogates when dealing with missing or NA values. The logistic model requires users to omit all missing values, or to perform further data cleaning by filling up missing values using other prediction means. Employees using the Client Retention Priority Model may prefer CART, as it is able to provide a visual representation of the classification process, and the simplicity and ease of use over the use of logistic regression as they lack the technical understanding of the technique used. However, the logistic model would be preferred in some cases. For instance for the Product Acceptance model, using the logistic model can help filter out clients which are missing the requirements easily and omit from consideration automatically.

With the above considerations in mind, WR can then conclude on which model to use by weighing both the accuracy and practicality of each model.

## 3.3 Churn Model

The predicted Y-variable for the model is whether the client has left the financial institution or continues to be a client. This is based on the “Exited”’ variable where 1 represents a client who has exited and 0 represents a client who has not exited. The dataset used for this model comprises 10,000 individual client records. We then developed a Churn Model using different techniques to predict the outcome variable and identify the clients who are likely to churn from White Rock and determine the significant variables which affect that outcome.

### 3.3.1 Churn Model - Logistic

For the Logistic Regression Model, we first did a train-test split of the dataset using a train-test split ratio of 70:30. Next, we ran the logistic regression model using the train set data with all the variables excluding the insignificant variables in the dataset such as “Rownumber”, “CustomerId” and “Surname”. This allowed us to determine which variables are statistically significant based on the p-value. (Appendix E-1.1)

The preliminary results of the logistic regression suggest that the estimated salary on an individual is a weak predictor for whether he will churn as the Odds Ratio is close to 1. However, expert opinion suggests that estimated salary should be a relatively strong determinant of churn. To verify this, we reran the model using a new variable representing the unit changes of estimated salary by 1 standard deviation. (Appendix E-1.2) The new results prove that an increase in 1 s.d. of estimated salary only increases the odds of churn by a factor very close to 1 (Appendix E-1.3) and thus is indeed a weak predictor of churn.

To build a stronger case as to which variables should be removed, we conducted a backward elimination to automatically remove statistically insignificant variables like the tenure of the client and whether the client owns a credit card. (Appendix E-1.4) On the same note, this backwards elimination process also removed estimated salary which further backs the case that this variable is a weak predictor. The final variables included in the logistic regression model is the following:

|  |
| --- |
| *glm(Exited ~ CreditScore + GeographyGermany + GeographySpain + GenderMale + Age + Balance + NumOfProducts + IsActiveMember1)* |

We checked for any multicollinearity problems using the vif function on the final model to get the Generalized Variance Inflation Factors. Based on the output, there appears to be no multicollinearity problem between the variables (Appendix E-1.6) and we are clear to proceed to test the model.

Comparing the predicted and actual outcome variables of the train set and test set, the overall predictive accuracy of the model using the train and test set were 81% and 82% respectively.

A confusion matrix was generated to further analyze our results. Considering the main purpose of this model is to predict clients who are likely to churn, our interest lies in (i) the True Positive Rate (reflecting clients who are predicted to churn and actually did churn), which will be our main pool of clients for retargeting and (ii) false positive rate (clients who are predicted to churn but actually did not churn), which will present wasted resources in the client retention process.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| **0** | **1** |
| **Actual** | **0** | 2306 | 83 |
| **1** | 469 | 142 |

As displayed in the table above, the true positive rate = 23.24% and false positive rate = 3.47% (Appendix E-2)

### 3.3.2 Churn Model - CART

For the CART model, unlike for the logistic model, we used all the variables and first grew the tree to its maximum by setting minimum split = 2 and cp = 0 (Complexity Parameter). After growing the tree to the maximum, we then found the optimal cp and pruned the tree by choosing the model with the optimal cp (Appendix E-2 ). Using the CART model developed using the train-set, we calculated the model accuracy using the test-set which is 0.8522857.

A similar confusion matrix was generated to assess the True Positive and False Positive rates, as done is 3.3.1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| **0** | **1** |
| **Actual** | **0** | 2299 | 90 |
| **1** | 330 | 281 |

Calculating from the values in the table above, the true positive rate = 45.99% and false positive rate = 3.77%. Refer to Appendix E-4 for a detailed breakdown.

### 3.3.3 Churn Model - Conclusion

The overall test set accuracy for the Churn Model using logistic regression = 0.8084286 while that for the model developed using CART = 0.8522857. The overall accuracies of both models are comparable, and we can conclude that using CART results in a higher model accuracy.

Next, we compared the percentage of True Positive Rates and False Positive rates of both models. To better meet our purpose of identifying customers who churn, it is important that the model has a higher True Positive Rate and a lower False Positive rate (as described in 3.3.1). This is observed in the CART results where the True Positive Rate was 22% higher and the false positive rate was 0.3% higher to that of the Logistic Regression Model. This makes CART a more reliable model in meeting our intended purpose. The list of clients for potential retargeting is more accurate. Employees can then spend their time to further understand their needs and deliver better products to reduce the likelihood of churning. While we are looking for a model with a lower false positive rate, the higher value reflected by CART is relatively insignificant. Looking at the results, we would still recommend the CART model as it has a much higher positive rate compared to the logistic model.

## 3.4 Product Acceptance Model

Once we have determined which clients are likely to churn based on the Churn Model in 3.1, we look at how to determine which clients should be retargeted. Our dataset comprises 5000 individual client data which includes variables such as Age, Income, Education Level, Credit card spending, etc. On this note, we have developed a Product Acceptance Model to identify which clients are likely to accept a new product purchase from WR and to also determine the significant variables affecting the outcome.

### 3.4.1 Product Acceptance Model - Logistic

The predicted Y-variable is whether or not the client will accept the personal loan. This is based on the column “Personal Loan” where 0 represents a client who does not accept the personal loan and 1 represents a client who accepts the personal loan. The train-test split ratio used is 70:30. Next, we ran the logistic model using the train set data with all the variables included to determine significant variables based on the p-value. (Appendix F-1.1) From this, we can first identify that the variable of Age, Experience and owning a securities account is not statistically significant.

To build a stronger case as to which variables should be removed, we conducted backward elimination to automatically remove insignificant variables such as Experience and Age. (Appendix F-1.3). Conducting backward elimination also helps to avoid creating a complex model. Next, we further analysed the odds ratio of the model and identify variables to possibly be removed if their OR is close to 1. (Appendix F-1.6) We observed that the OR of Income and CCAvg (Average credit card spending) is close to 1 but expert opinion suggests that Income and spending are strong determinants of demand. To verify this, we did the following calculations (OR of Income ^ standard deviation of Income). Based on the results in Appendix F-1.7, we can conclude that Income and CCAvg are not a weak predictor of Y-variable. On the same note, the backward elimination method kept CCAvg and Income which further backs the case that these 2 variables are significant.

The final variables included in the logistic regression model is the following:

|  |
| --- |
| *glm (Personal Loan ~ Income + Family + CCAvg + EducationGraduate + EducationAdvanced/Professional + CD Account + Online + CreditCard* |

With the model above, we checked for multicollinearity problems using the vif function to generate the Generalized Variance Inflation Factors. Based on the output, there is no multicollinearity problem between the variables. (Appendix F-1.8)

Similarly, a confusion matrix was generated to further analyze our results. Considering the main purpose of this model is to predict clients who are likely to take up an additional product, our interest lies in (i) the True Positive Rate (reflecting clients who are predicted to and actually did accept the products), to ensure that resources allocated will be maximised and (ii) false positive rate (clients who are predicted to accept the products but actually did not), to ensure that minimal resources are wasted on clients who will not take up our products.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| **0** | **1** |
| **Actual** | **0** | 1340 | 16 |
| **1** | 51 | 93 |

As displayed in the table above, True Positive rate = 62.00% and False positive rate = 10.67% (Appendix E-2)

### 3.4.2 Product Acceptance Model - CART

For the CART model, the predicted Y variable is whether the client will accept the personal loan from the bank. Similar to the development of the Churn CART model, all variables were included to allow the tree to grow to its maximum by setting minimum split = 2 and cp = 0 (Complexity Parameter). We then found the optimal cp and pruned the tree by choosing the model with the optimal cp.

In assessing the optimal CART model in Appendix F-3, we found the variables in the split points to be consistent with the significant variables included in the logistic regression model. This shows a consistency in assumptions made and significance of variables in the development of both models.

Next, we calculated the variable importance of each variable to further assess the consistency of the variables chosen (Appendix F-4). We noted that “Education” and “Income” are the most important variables. This is consistent with the logistic model where “Education” and “Income” are the most significant based on p-value.

A similar confusion matrix was generated to assess the True Positive and False Positive rates, as done in the logistic model. We noted that from the table below, the True Positive Rate is at 83.33% and the False positive Rate is at 0.6%. This means that there is a 83.33% chance the model predicts that the client will accept a new product from WR and the client will actually accept it. There will also be a 0.6% chance of the model predicting the client will accept a new product but actually does not accept. This means our model is reasonably reliable.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| **0** | **1** |
| **Actual** | **0** | 1347 | 9 |
| **1** | 16 | 128 |

### 3.4.3 Product Acceptance Model - Conclusion

The overall accuracy for the Product Acceptance Model using logistic regression for the test set is 0.9553333 while that for the model developed using CART is 0.9833333. We note that the CART model shows a slightly higher model accuracy.

Next, we compared the percentage of True Positive Rates and False Positive rates of both models. To better meet our purpose of predicting which client would accept a new product, it is important that the model has a higher True Positive Rate and a lower False Positive Rate so as to ensure resources are allocated effectively. This is observed in the CART model result where the True Positive Rate is 20% higher and the False Positive rate is 10% lower than that of the Logistic model. This makes the CART model a more reliable model in meeting our intended purpose. In comparing both results, we are able to invest more efficiently and minimise waste of resources with the results produced by CART.

Furthermore, it is important to take into consideration other aspects of the model to determine which model to be used. From the user perspective, the results are likely seen and used by sales managers within the company, they are unlikely to be analytics trained. Thus, as compared to a logistic regression equation, the decision tree produced by CART allows for an easier interpretation and understanding since it is a series of decisions at each split. Moreover, CART has a surrogate function to automatically handle missing values in a dataset. Whereas for the logistic model, rows with missing values have to either be replaced with an appropriate value or delete the row entirely before the model can be trained. Hence, the CART model is also better for training.

Given the considerations mentioned above, implementing the CART model will serve as a more accurate and easy to understand model for the user.

# 4. Limitations of Proposed Solution

Based on our proposed solution, we have pinpointed a few mitigating factors that will limit the effectiveness of the proposed solution.

## 4.1 Imbalanced Datasets

The Churn and Product Acceptance datasets used are observed to have imbalanced proportions of positive cases to negative cases of 20:80 and 10:90 respectively. The consequence of such disproportions is that predictive models will not perform as optimally compared to using perfectly balanced datasets (50:50 positive and negative cases).

When trained using an imbalanced dataset, the model will form a bias towards the majority case. This is due to the model not having ample data to make accurate predictions on the minority case, resulting in the model predicting the majority case more often with a high level of accuracy. Such a high level of accuracy will cause the bias to possibly be overlooked. Such bias will limit the effectiveness and accuracy of the models and may result in poor-decision making.

However, in the finance industry, the average annual churn rate for clients is at 25% (Mazareanu, 2020) similar to the Churn dataset proportion of positive cases at 20%. Likewise, the probability of selling a product to an existing client is 60% but the probability of selling to an existing client who is likely to churn will be significantly smaller. The datasets may be imbalanced but it does help to show a more realistic scenario which WR would encounter in their own real dataset.

## 4.2 Unable to Consider Qualitative Reasons

A study done by Ernst & Young concluded that asset management clients tend to switch during critical life changes (Nanayakkara, Birkin, & Wightman, 2019). The research concluded that over the past 3 years, 61% of clients churn due to a change in job, 50% of clients switch when having a child or getting married, 43% of clients switch when divorcing and 40% switch when retiring.

Another major reason for clients to churn is during intergenerational wealth transfer. 50% of client beneficiaries have churned after a transfer of wealth. On that note, our proposed solution is unable to consider such life events to predict the probability of client churn.

However, our model can be complimented by developing an artificial intelligence to learn client data that financial institutions possess as well as client public data through social media. This would help to develop a more complete profile of the client through using social media data in an attempt to pre-empt potential churn probability based on life events that the client may have posted on social media.

# 5. Optimisation of the Client Retention Priority Model

The Client Retention Priority Model aims to be a tool in which WR can develop into its optimal client retention strategy.

An added advantage of the model is the use of geographical data in predicting a client’s churn. The final priority score takes into account the proportion of high priority clients across the operating regions. This will be particularly useful as WR operates in 32 markets globally with a shared budget in certain markets (i.e. SEA’s budget is planned at $X million to be split among markets). Furthermore, we can understand the odds ratio of a client in market A compared to a client in market B. This will form a basis for management to understand how much costs to be allocated to the various markets and will serve as a good justification for any budget adjustments.

On the whole, the group does not foresee huge costs and difficulties in the implementation of the model. The main data source of the retention model is basic client information which can be found from the internal database of WR. Depending on how different markets store their data, a possible area where additional costs may be incurred may lie in the standardisation of data reports for processing. Since no external data is being used, we will not expect high costs to be incurred in the data gathering and cleaning process. Thus, making this model self-sustainable.

On the same note, the Client Retention Priority Model is made up of two CART models. The ease of use of the CART model adds to the optimisation capability of the solution to be used by any user. The model has the potential to communicate insights more efficiently by automating the output (Client ID, priority score and amount of resources to be allocated for a specific client) so that WR reduces the need to interpret results. Such reports can be generated yearly and serve to track any changes in priority score between the current and the last run. This change may be due to an increase in accuracy of the model or the reflection of successful client retention efforts.

# 6. Conclusion

Data and analytics have seen strides in the past few years in the finance industry. With the availability of predictive models and artificial intelligence, there now exists an opportunity to enhance many aspects of the business. Client retention has always been a critical aspect of business sustainability and profitability. However, it has only increased in complexity with the abundance of data. Therefore, businesses that choose not to implement data analytics to aid in it’s client retention strategy risks drops in profitability.

The Client Retention Priority Model represents an opportunity to White Rock. The model helps White Rock to mine the value of the internal data that WR has been collecting to better allocate resources to improve client retention and increase revenue. We believe that if given a real client dataset, the Client Retention Priority Model will be able to achieve the above mentioned and make it available for implementation.

# 

# References

*2019 Global Wealth Management*. (2019). From EY: https://assets.ey.com/content/dam/ey-sites/ey-com/en\_gl/topics/wealth-and-asset-management/wealth-asset-management-pdfs/ey-global-wealth-management-research-report-2019.pdf

*Banking Customer Experience Program*. (2019, July 20). From qualtrics: https://www.qualtrics.com/customer-experience/banking-report/

Birkin, A. W. (2019, April 16). *When wealth management clients want to switch, will you be their first choice?*

*Customer Experience: Increase customer loyalty and retention with smart CX strategy*. (2020, October 27). From gartner: https://www.gartner.com/en/marketing/insights/customer-experience

Global, E. (2019 , April 30). *EY research challenges value of financial advice to understand why 33% of wealth management clients plan to switch providers*. From EY: https://www.ey.com/en\_gl/news/2019/04/ey-research-challenges-value-of-financial-advice-to-understand-why-33-of-wealth-management-clients-plan-to-switch-providers

Jones, M. (n.d.). *The cost of customer retention*. From coxblue: https://www.coxblue.com/the-cost-of-customer-retention/

Jovancic, N. (2019, March 14). *4 Types of Bias in Research and How to Make Your Surveys Bias-Free*. From lead quizzes: https://www.leadquizzes.com/blog/types-of-bias-in-research/

Malik, T. (2018, December 5). *HOW BIG DATA CAN CONTRIBUTE IN REDUCING BANKS’ CHURN RATE?* From invivoo: https://blog.invivoo.com/how-big-data-can-contribute-in-reducing-banks-churn-rate/

Mansfield, M. (2016, October 25). *CUSTOMER RETENTION STATISTICS – The Ultimate Collection for Small Business*. From small business trends: https://smallbiztrends.com/2016/10/customer-retention-statistics.html

Mazareanu, E. (2020, June 18). *Customer service: Churn rate by industry U.S. 2018.* From Statista: https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/

Nanayakkara, N., Birkin, A., & Wightman, M. (2019, April 2016). *When wealth management clients want to switch, will you be their first choice?* From EY: https://www.ey.com/en\_sg/wealth-asset-management/when-wealth-management-clients-want-to-switch-will-you-be-their-first-choice

Sethuraman, S. (2019 , September 25). *Why Customers Leave & What Can Banks Do?* From Tiger Analytics.: https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/

Vincent, G., Goldstein, J., & Cunniff, S. (n.d.). *10 Disruptive trends in*. From Deloitte: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/strategy/us-cons-disruptors-in-wealth-mgmt-final.pdf

Waite, S., Massa, A., & Cannon, C. (2019, August 8). *Asset Managers With $74 Trillion on Brink of Historic Shakeout*. From Bloomberg: https://www.bloomberg.com/graphics/2019-asset-management-in-decline/

# Appendix A: Datasets and Data Dictionary

Dataset #1: Bank Churn

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Data Type** | **Variable Type** |
| RowNumber | Row numbers from 1 to 10,000 | Numeric | Categorical |
| CustomerId | Unique Customer ID generated by the bank to identify clients | Numeric | Categorical |
| Surname | Client’s last name | Character | Categorial |
| CreditScore | Client’s credit score with the bank | Numeric | Continuous |
| Geography | Identifies the country which the client holds an investment in | Character | Categorical |
| Gender | Gender of Client | Character | Categorical |
| Age | Client’s age in completed years | Integer | Continuous |
| Tenure | Number of years the client has been with the bank | Integer | Continuous |
| Balance | Current bank balance of clients. For clients who have exited, the amount reflected relates to the last balance before exit. | Integer | Continuous |
| NumOfProducts | Number of products which the client is currently using | Integer | Continuous |
| HasCrCard | Indicates whether a customer holds a credit card with the bank.  1: Has Credit Card  0: No Credit Card | Logical | Categorical |
| IsActiveMember | Indicates whether the client is an active member with the financial institution (i.e. if they actively engage with other departments such as insurance, bonds etc)  1: Active Member  0: Not Active Member | Logical | Categorial |
| EstimatedSalary | Estimated annual salary, in absolute value, of the client in dollars. | Integer | Continuous |
| Exited | Indicates whether a client has closed their account with the financial institution.  1: Account Closed  0: Account not Closed | Logical | Categorical |

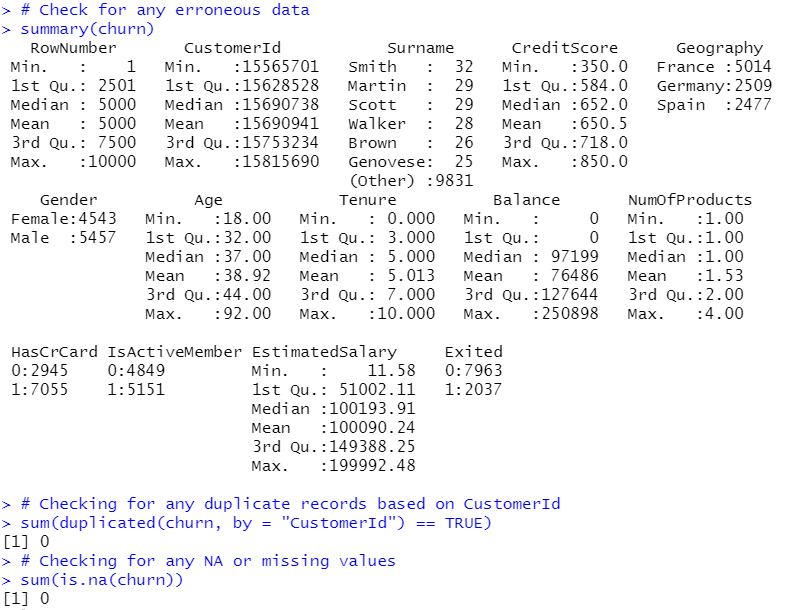
Dataset #2: Personal Loan Classification

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Data Type** | **Variable Type** |
| ID | Customer ID | Numeric | Categorical |
| Age | Client’s age in completed years | Integer | Continuous |
| Experience | Number of years of professional experience | Integer | Continuous |
| Income | Annual income of clients in dollars, in thousands. | Integer | Categorical |
| ZIPCode | Home Address ZIP code. | Numeric | Categorical |
| Family | Number of family members of the client | Numeric | Continuous |
| CCAvg | Average spending on credit cards per month, in thousands | Integer | Continuous |
| Education | Education Level of Client.  1: Undergrad; 2: Graduate;  3: Advanced/Professional | Numeric | Categorical |
| Mortgage | Value of house mortgage, if any, in thousands | Integer | Continuous |
| Personal Loan | Did this customer accept the personal loan offered in the last campaign?  1: Personal Loan accepted  0: Personal Loan not accepted | Logical | Categorical |
| Securities Account | Does the customer have a securities account with the bank?  1: Has Securities Account  0: No Securities Account | Logical | Categorical |
| CD Account | Does the customer have a certificate of deposit (CD) account with the bank?  1: Has CD account  0: No CD account | Logical | Categorical |
| Online | Does the customer use internet banking facilities?  1: Uses internet banking  0: Does not use internet banking | Logical | Categorical |
| CreditCard | Does the customer use a credit card issued by UniversalBank?  1: Uses credit card  0: Does not use credit card | Logical | Categorical |

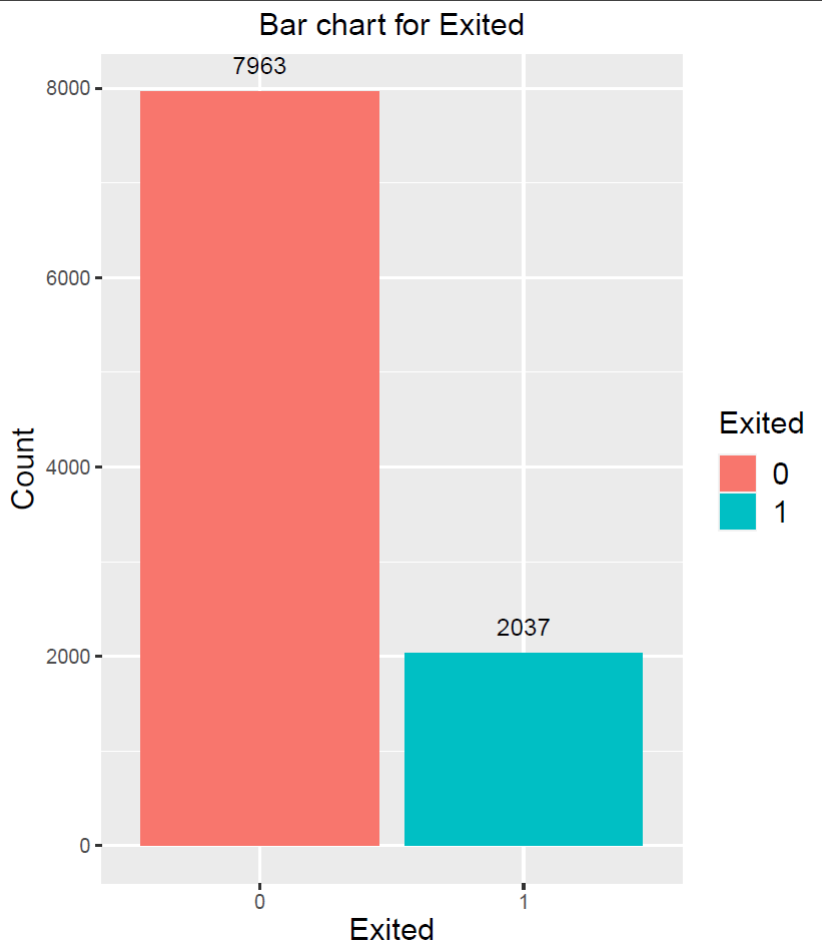
# Appendix B: Data Exploration for Churn Dataset

Since there is no erroneous, duplicated or missing/NA data values, we proceed to the data exploration for the dataset

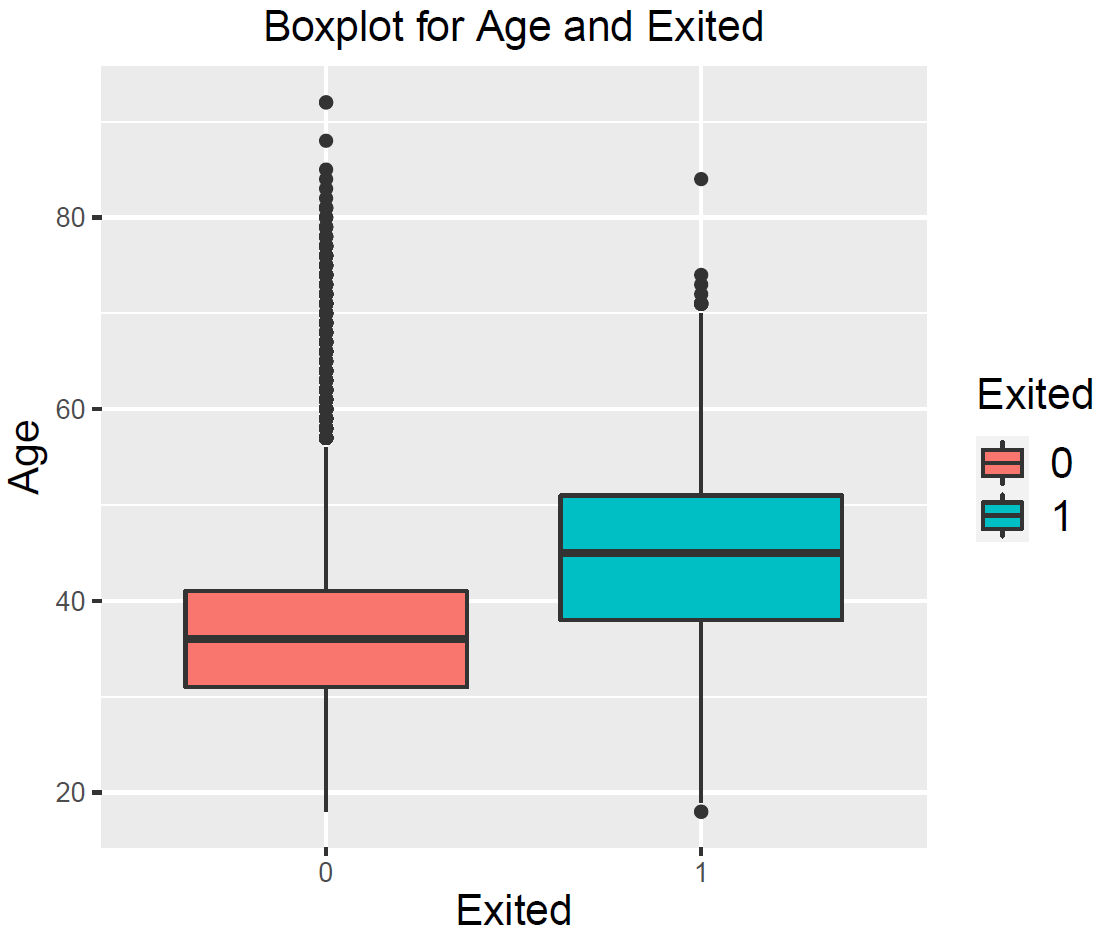
**Appendix B-1 : Checking whether data cleaning is required**



**Appendix B-2 : Distribution of Exited customers**



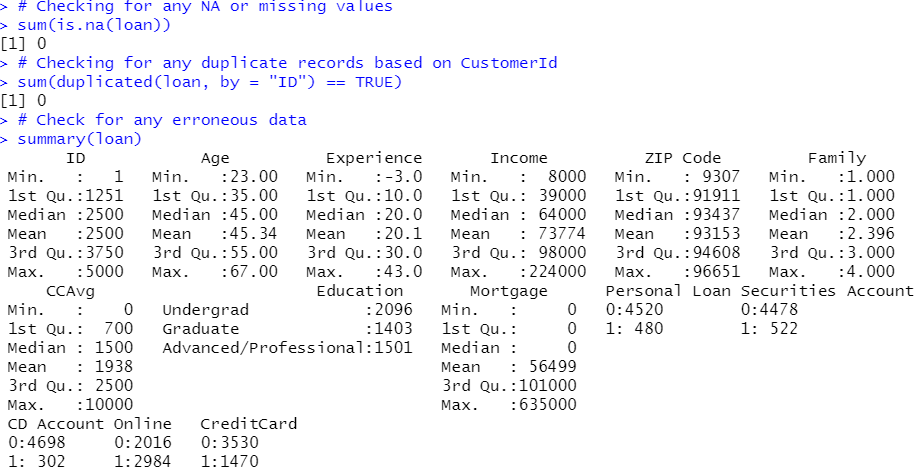
**Appendix B-3 : Boxplot of Age and Exited**



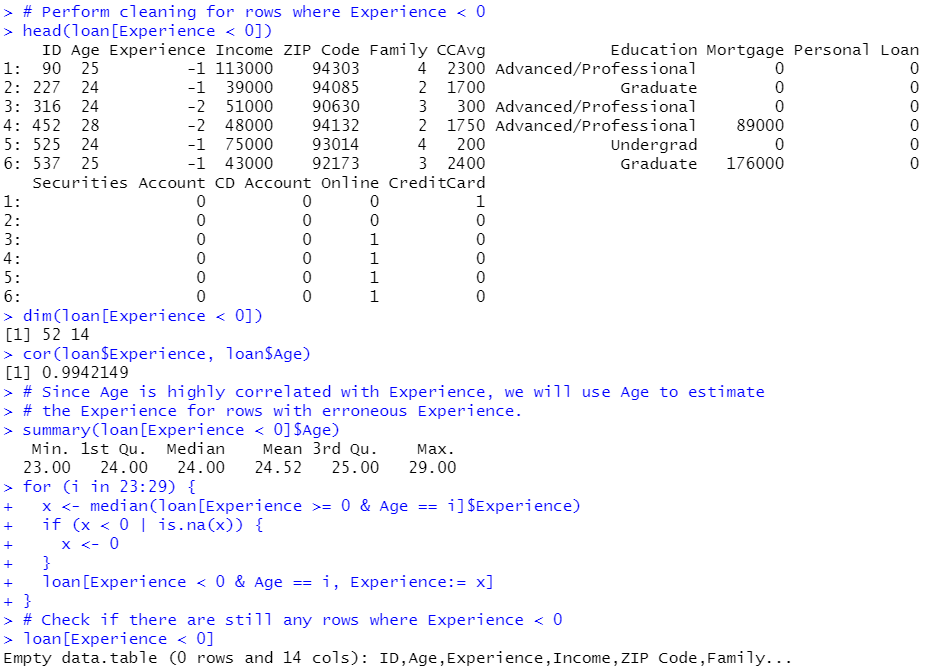
# 

# Appendix C: Data Cleaning for Personal Loan Dataset

**Appendix C-1 : Checking for missing, duplicated or erroneous data. Found that there is data where Experience < 0**

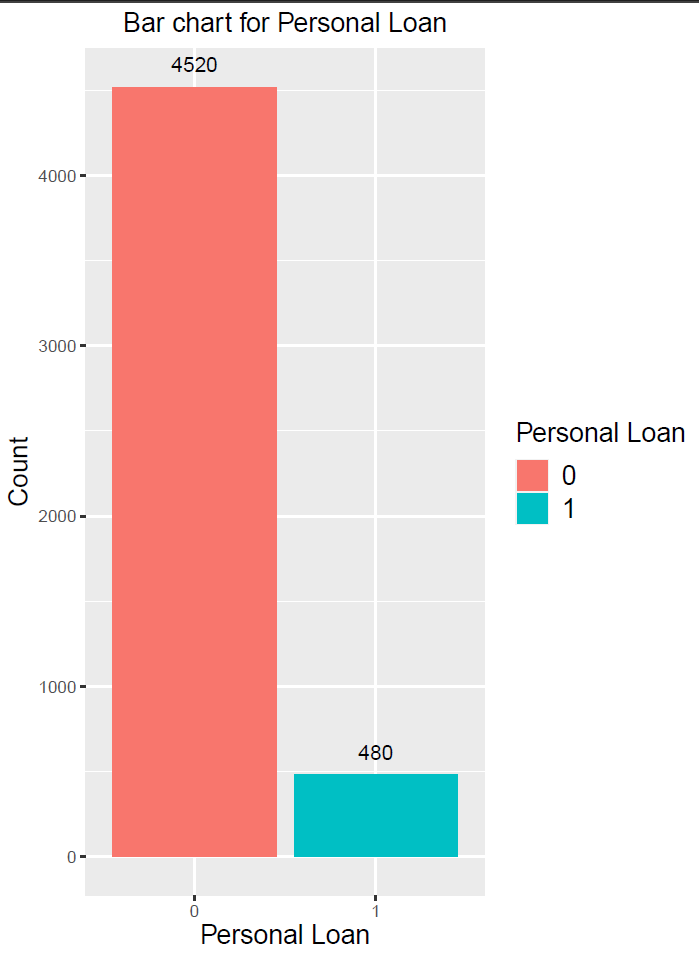


**Appendix C-2 : Perform cleaning for rows where Experience < 0, using age to estimate experience**

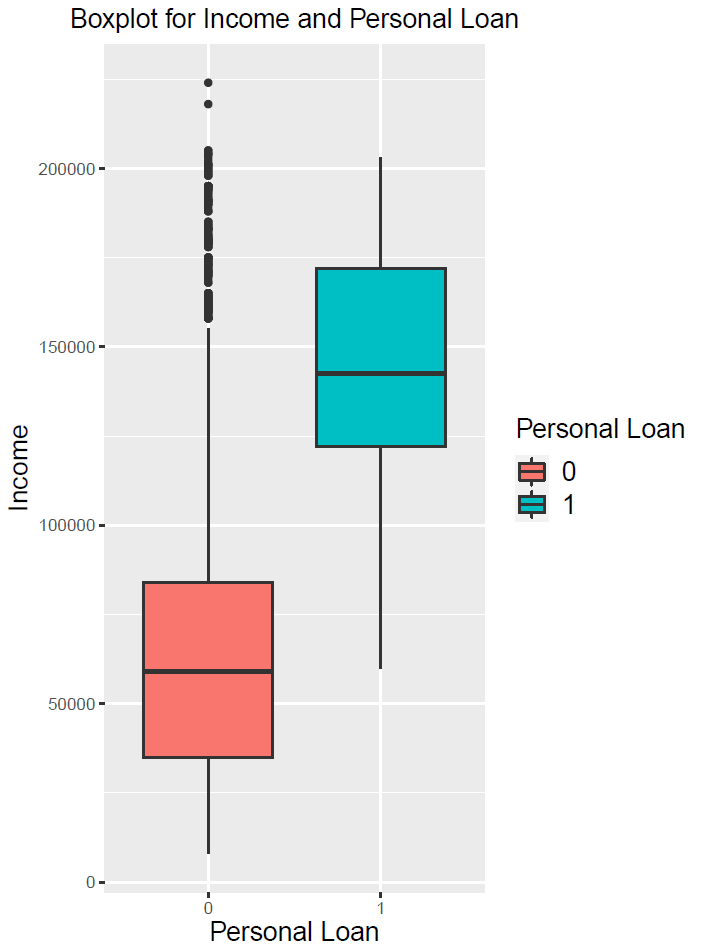


# Appendix D: Data Exploration for Personal Loan Dataset

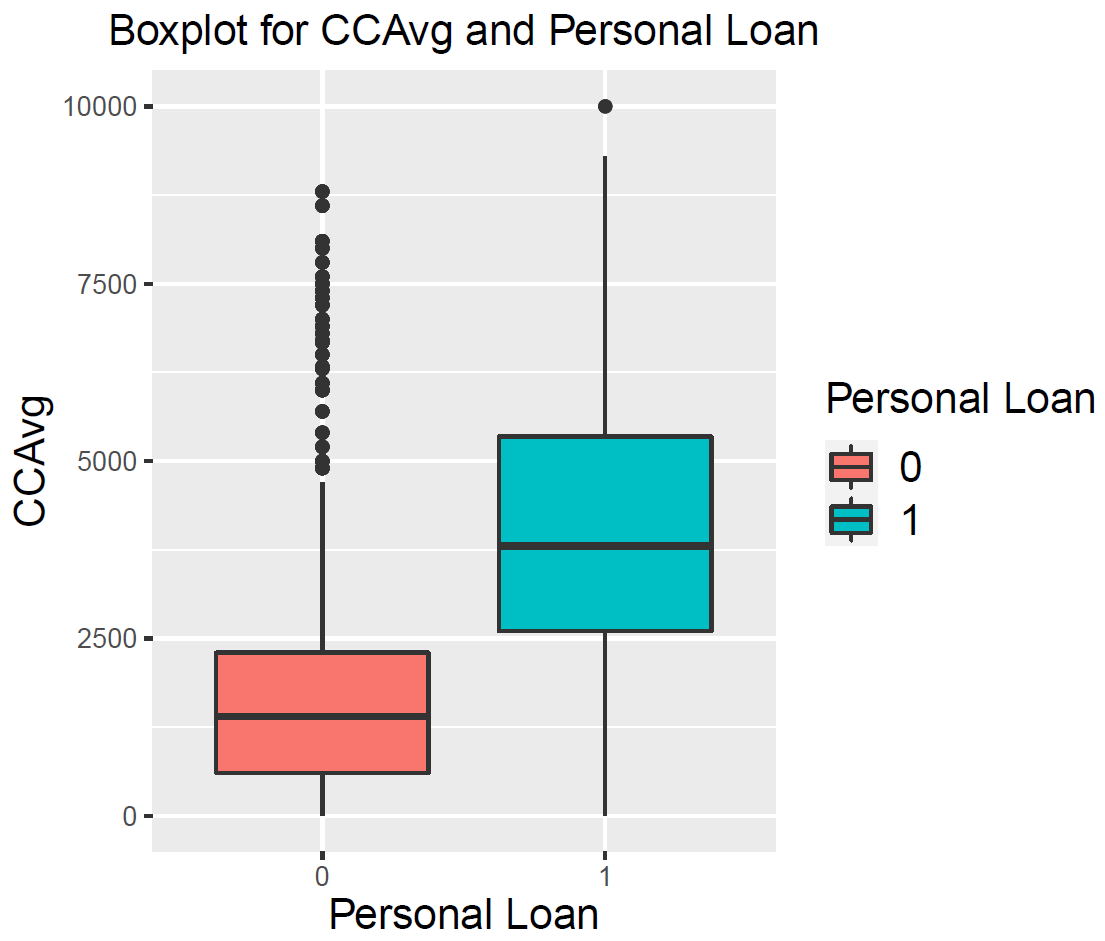
**Appendix D-1 : Check the distribution of customers who accepted the Personal Loan**



**Appendix D-2 : Boxplot of Income and Personal Loan**

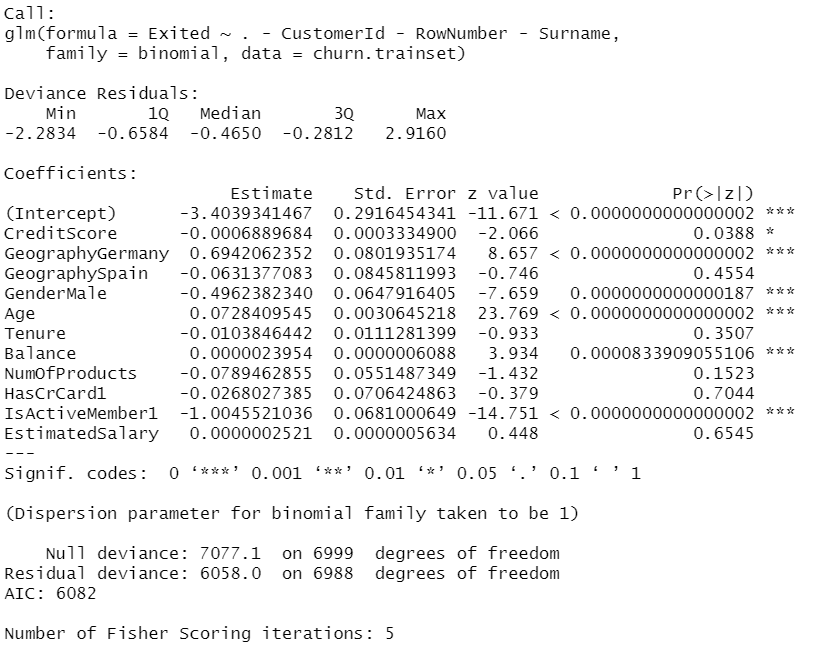
****

**Appendix D-3 : Boxplot of CCAvg and Personal Loan**

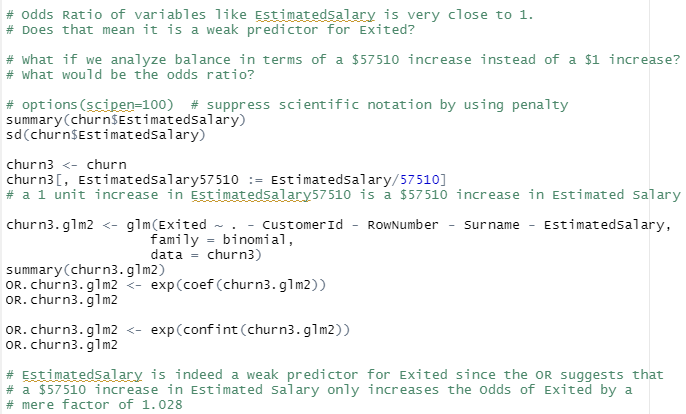
****

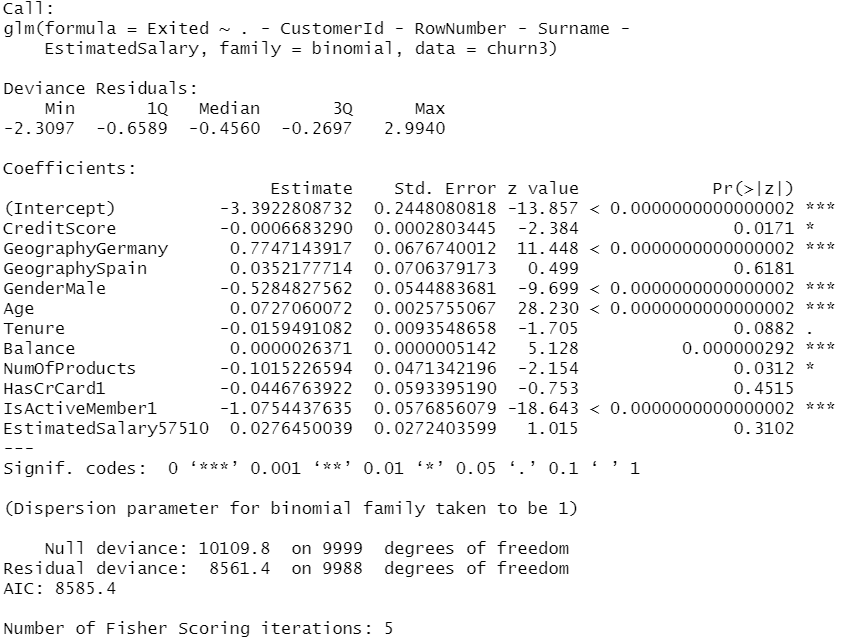
# Appendix E: Churn Logistic Model

**Appendix E-1.1:Churn Logistic Model Summary**

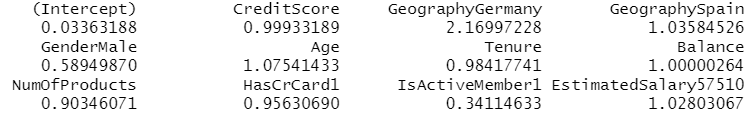


**Appendix E-1.2: Used a new variable to represent the unit changes of estimated salary by 1 standard deviation.**

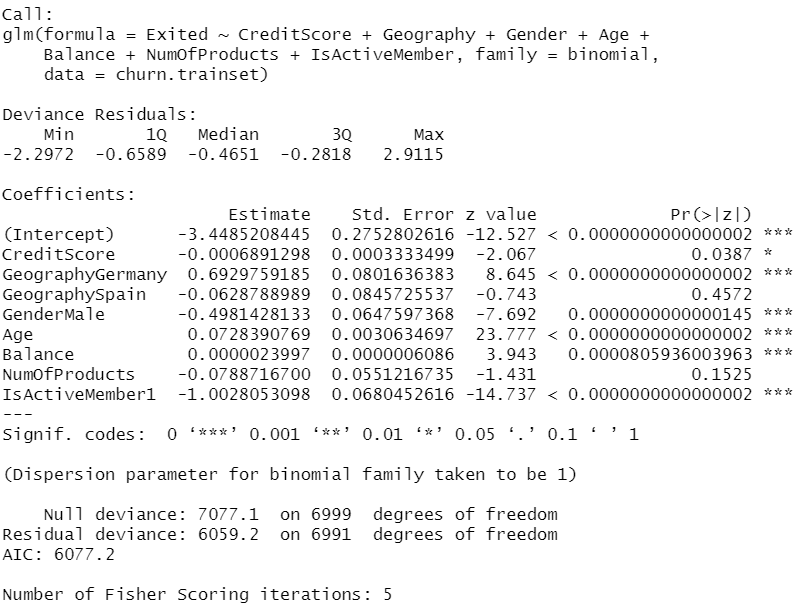
****



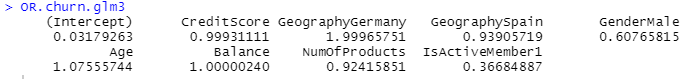
**Appendix E-1.3: :Churn Logistic Model Odds Ratio**



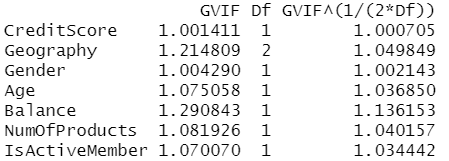
**Appendix E-1.4: Finalised Churn Logistic Model Summary**



**Appendix E-1.5: Finalised Churn Logistic Odds Ratio**

****

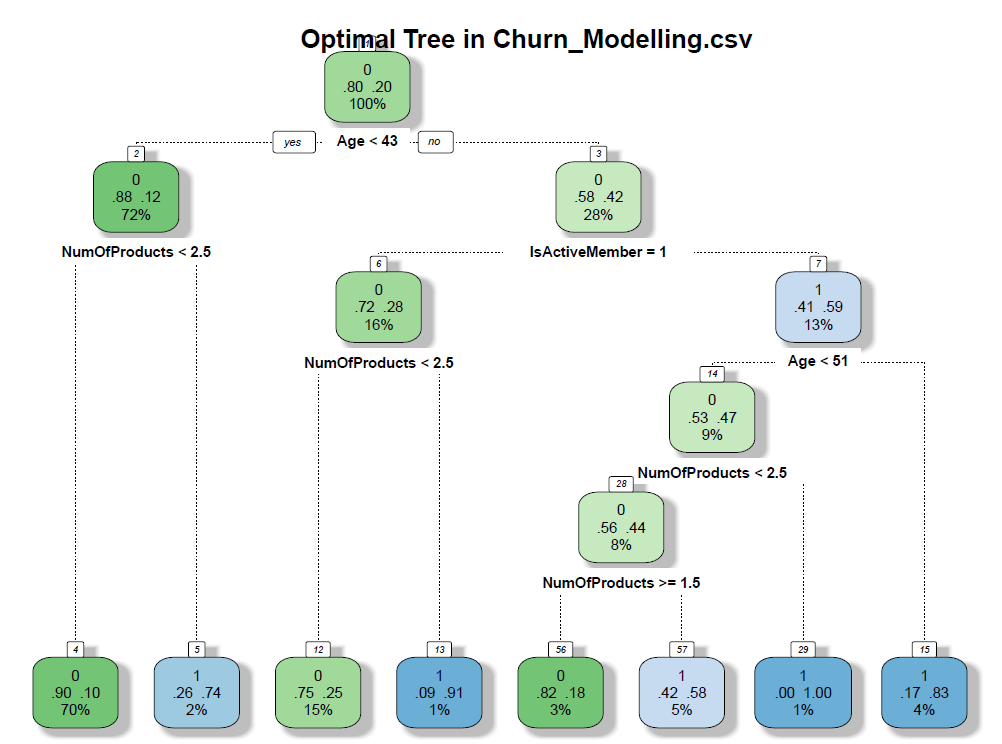
**Appendix E-1.6: Multicollinearity Check**



**Appendix E-2: Confusion Matrix Statistics for Churn Logistic Model**

|  |  |  |
| --- | --- | --- |
| Conditional Probability | Probability of outcome | Type |
| Client predicted not to churn and indeed does not churns  (predicted =0, actual =0) | [2306 / (83 + 2306 )] \* 100 = 96.53% | True Negative Rate |
| Client predicted not to churn but churns  (predicted =0, actual =1) | [469 /(142 + 469)] \* 100 = 76.76% | False Negative Rate |
| Client predicted to churn and indeed churns  (predicted =1, actual =1) | [142/(142 + 469)] \* 100 = 23.24% | True Positive Rate |
| Client predicted to churn but does not churn (predicted =1, actual =0) | [83 /(83 + 2306 )] \* 100 = 3.47% | False Positive |

**Appendix E-3: Optimal Churn CART Model**



**Appendix E-4: Confusion Matrix Statistics for Churn CART Model**

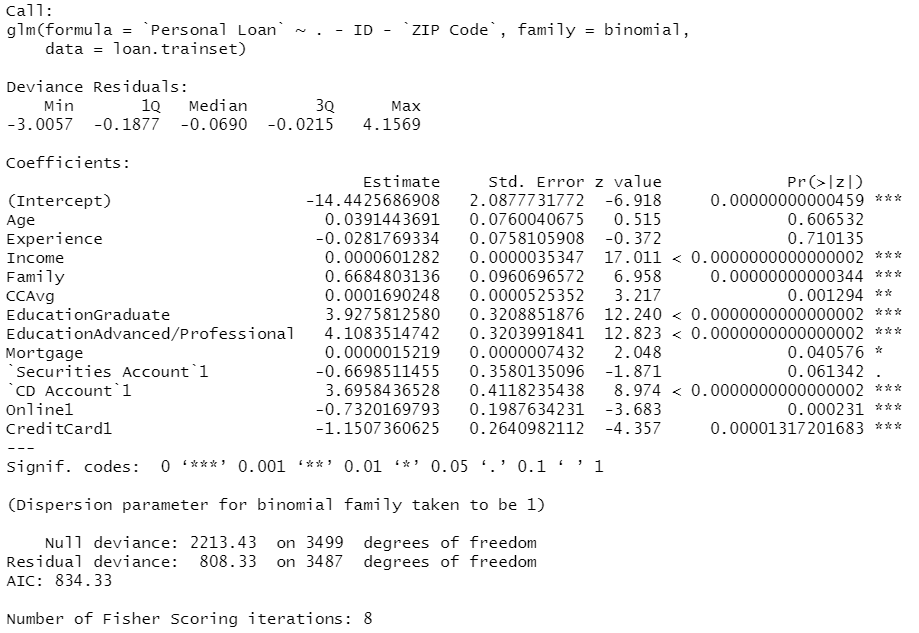
|  |  |  |
| --- | --- | --- |
| Conditional Probability | Probability of outcome | Type |
| Client predicted not to churn and indeed does not churns  (predicted =0, actual =0) | [2299 / (90+ 2299 )] \* 100 = 96.23% | True Negative Rate |
| Client predicted not to churn but churns  (predicted =0, actual =1) | [330 /(281 + 330)] \* 100 = 54.01% | False Negative Rate |
| Client predicted to churn and indeed churns  (predicted =1, actual =1) | [281/(281 + 330)] \* 100 = 45.99% | True Positive Rate |
| Client predicted to churn but does not churn (predicted =1, actual =0) | [90 /(90+ 2299 )] \* 100 = 3.77% | False Positive Rate |

**Appendix E-5: Variable Importance for Churn CART Model**

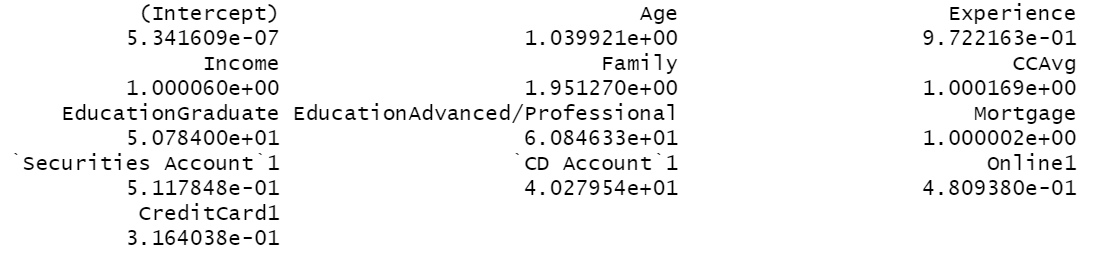
****

# Appendix F :Product Acceptance Logistic Model

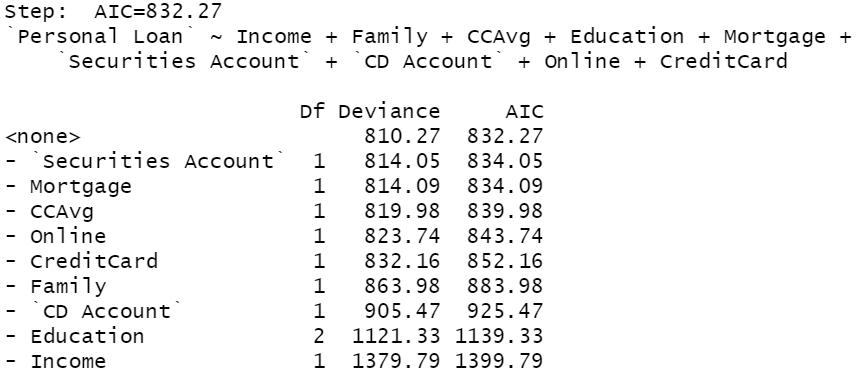
**Appendix F-1: Product Acceptance Logistic Model Summary**



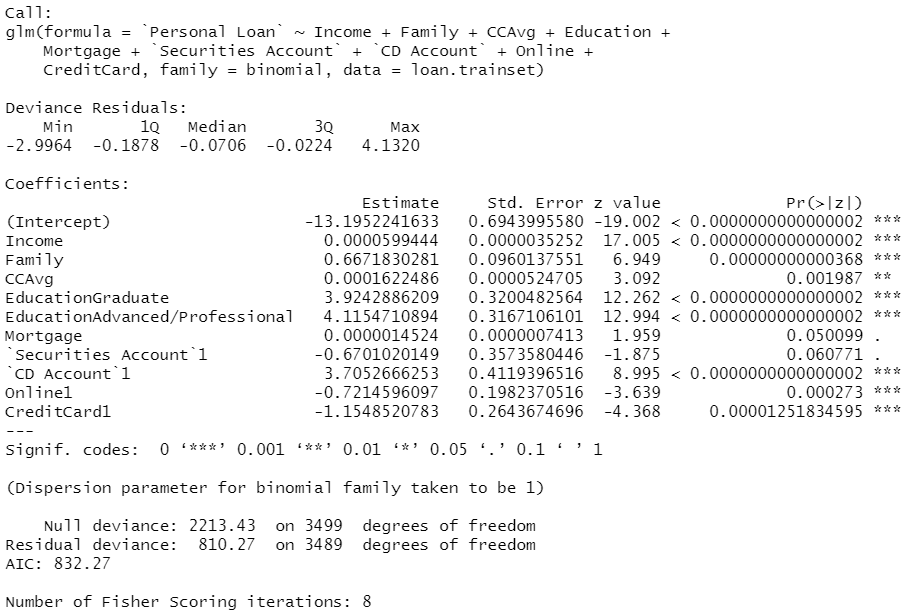
**Appendix F-1.2: Product Acceptance Logistic Model Odds Ratio**



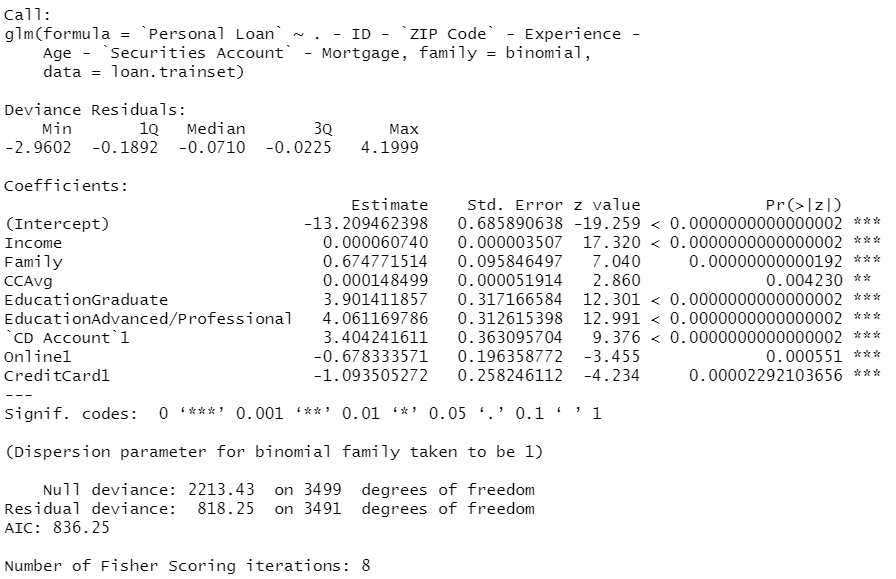
**Appendix F-1.3: Backward Elimination Method**



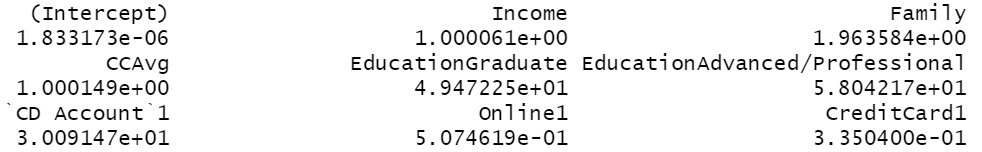
**Appendix F-1.4: Product Acceptance Logistic Model Summary (After Backward Elimination)**



**Appendix F-1.5: Finalised Product Acceptance Logistic Model Summary**

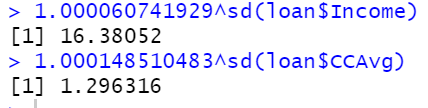


**Appendix F-1.6: Finalised Product Acceptance Logistic Model Odds Ratio**

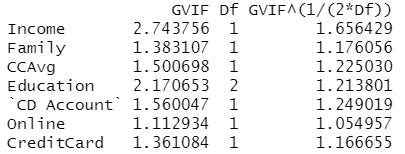


**Appendix F-1.7: Check Whether Income and CCAvg are Weak Predictors**

An increase in income by 1 s.d. increases the odds of personal loan acceptance by a factor of 16. Therefore, Income is not a weak predictor. Likewise for CCAvg, the increase in CCAvg by 1 s.d. increases the odds of personal loan acceptance by a factor of 1.29. Therefore, CCAvg is not a weak predictor



**Appendix F-1.8: Multicollinearity Check**

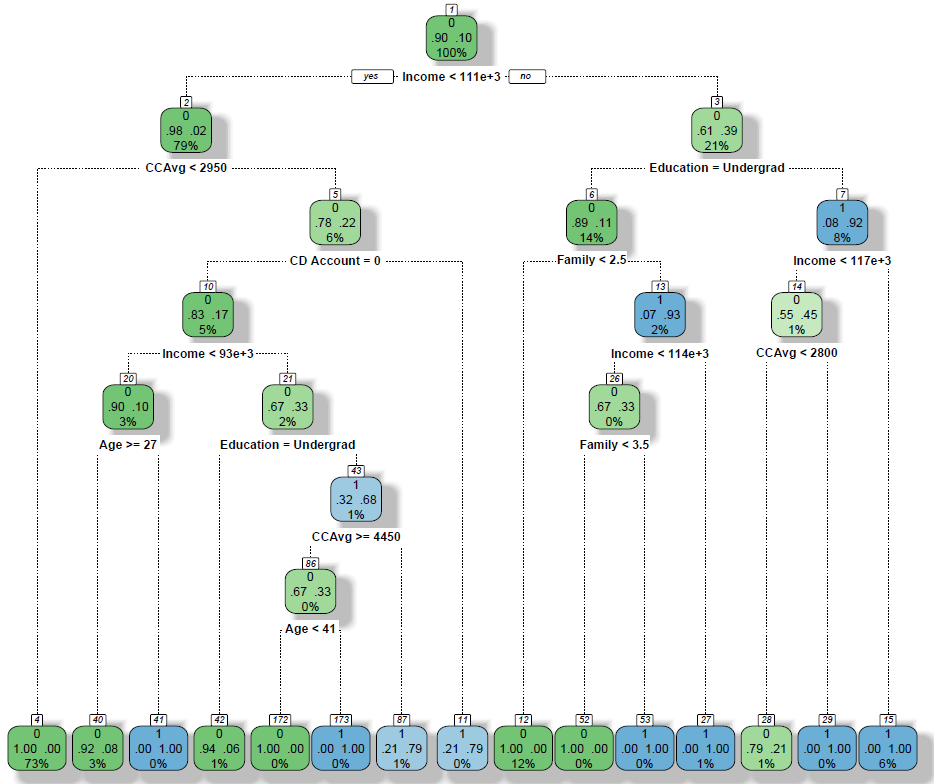


**Appendix F-2: Confusion Matrix for Product Acceptance Logistic Model**

|  |  |  |
| --- | --- | --- |
| Conditional Probability | Probability of outcome | Type |
| Client predicted not to purchase and indeed does not purchase  (predicted =0, actual =0) | [1340/ (16+ 1340)] \* 100 = 98.82% | True Negative Rate |
| Client predicted not to purchase but purchase  (predicted =0, actual =1) | [51/(93 + 51)] \* 100 = 35.42% | False Negative Rate |
| Client predicted to purchase and indeed purchase  (predicted =1, actual =1) | [93/(93 + 51)] \* 100 = 64.58% | True Positive Rate |
| Client predicted to purchase but does not purchase (predicted =1, actual =0) | [16 /(16+ 1340)] \* 100 = 1.18% | False Positive Rate |

# 

# Appendix F-3: Optimal Product Acceptance CART Model



# Appendix F-4: Variable Importance for Product Acceptance Model



**Appendix F-5: Confusion Matrix Statistics for Product Acceptance CART Model**

|  |  |  |
| --- | --- | --- |
| Conditional Probability | Probability of outcome | Type |
| Client predicted not to purchase and indeed does not purchase  (predicted =0, actual =0) | [1347/ (9 + 1347)] \* 100 = 99.33% | True Negative Rate |
| Client predicted not to purchase but purchase  (predicted =0, actual =1) | [16/(128 + 16)] \* 100 = 11.11% | False Negative Rate |
| Client predicted to purchase and indeed purchase  (predicted =1, actual =1) | [128/(128 + 16)] \* 100 = 88.88% | True Positive Rate |
| Client predicted to purchase but does not purchase (predicted =1, actual =0) | [9 /(9+ 1347)] \* 100 = 0.66% | False Positive Rate |

# 