BC2406 Analytics I: Visual and Predictive Techniques Semester 1, AY 2020/21

TEAM PROJECT

Prepared By	Chai Jing Yi Verene (U1910479D) Lee Jia Xin (U1810597H) Shi Jiayi Joey (U1922195L) Owen Ong (U1921381G) Shannon Tan Xinyi (U1921019B)
Seminar Group	4
Team	4
Tutorial Instructor	Professor Neumann Chew
Date of Submission	1 November 2020

Table of Contents

1 Business Problem 1

1.1 Concept of Expanding Consumer Base 1

1.2 Importance of Customer Retention in Asset Management Firms 1

1.3 Rethinking the Problem: Targeted Customer Retention 2

2 Analytics Solution 3

2.1 The Opportunity: Use of Customer Retention Analytics to Tackle Retention 3

2.1.1 Difficulties Faced by Companies For Customer Retention 3

2.1.2 Approach 3

2.2 Project Feasibility 3

2.3 Desired Outcomes 4

3 Data Cleaning and Exploration 5

3.1 Data Cleaning Decisions 5

3.1.1 NA Values 5

3.1.2 Erroneous Values 5

3.1.3 Duplicate Rows 5

3.1.4 Redundant Variable 5

3.1.5 Converting variables to the relevant data type 5

3.2 Data Exploration 6

3.2.1 Univariate Graphs 6

3.2.2 Bivariate Graphs 7

4 Implementation & Recommendation 8

4.1 Implementing Predictive Model to Predict Customer's Likelihood of Staying 9

4.1.1 Model Evaluation 10

4.1.2 Implementation of CART and Insights 10

4.2 Implementing CART Model to Determine New Promotional Method for Leaving Customers 12

4.2.1 Model Evaluation 12

4.2.2 Implementation of CART and Insights 12

4.3 Implementing CART Model to Determine Need for Personalisation 14

4.3.1 Model Evaluation 14

4.3.2 Implementation of CART and Insights 14

4.4 Implementation of Solution 15

5 Literature Review and Expert Opinion 17

6 Limitations & Future Directions 18

6.1 Limitations 18

6.2 Future Directions 19
6.2.1 Usage of Sentiment Analysis on Customer Feedback 19
6.2.2 Constant Update of Model with Improvements from Past Predictions 19
6.2.3 Additional Models to Predict Other Factors 20
6.2.4 Extend the Usage of the Model to Customer Acquisition 20
7 Conclusion 20
APPENDIX 21
Appendix A - Data cleaning and exploration 21
Appendix B - Data Visualisation 22
Appendix C - Logistic Regression 32
Appendix D - CART Model 36
Appendix E - Interface 36
REFERENCES 40

Executive Summary

Case Context

White Rock, an asset management company, is looking to automate processes and discover technology to make faster and more informed business decisions. This project aims to help White Rock look at how they can better make decisions for marketing.

Business Problem and Opportunity

In the area of sales and marketing, we have identified retention and expansion of customer base as challenges faced by sales and marketing departments of companies today. Retention, in particular, is the focus of this project due to the emergence of more competitors and changing market trends in recent years. Companies are naturally more concerned with how they can retain the old customers and at the same time, increase their spending with the current market trends.

With the analysis process still requiring a lot of manual intervention, marketing campaigns are less effective in addressing current market trends, which are constantly changing. We found that customer retention issues are highly rampant in the asset management industry due to the rise of the FinTech Industry.

With that, we defined the business problem as the difficulty in retaining old customers in the asset management industry. By tapping on data analytics, we recognise the business opportunity as the ability to retain old customers and also increase their spending by automating the analysis of customers and market data to identify their trends and needs, thus generating a strong marketing function, especially in this technology age where trends are constantly changing.

Approach and Solution

Our proposed solution aims to train three different models by analysing customers' demographics and their receptibility towards past marketing efforts. The models will identify customers who are more likely to leave, their preferred marketing platform and whether personalisation is required for the particular customer. The outcome will enable us to better foresee customer churn as well as to understand the needs of different customer groups. As such, allowing companies to better cater marketing efforts towards each customer's preferences, increasing the likelihood of them staying with the company.

The implementation plan is to provide a simple and user-friendly interface that can display each customer's predicted retention and preferred marketing strategy for ease of further analysis for marketing managers in White Rock.

1 Business Problem

In the asset management industry, firms are responsible for directing clients' wealth or investment portfolio on their behalf. As such, profits of asset management firms are highly dependent on their customer bases. Ensuring that they retain a large consumer base is thus critical for asset management firms, especially with the competitive nature of the industry, where more than 800 organisations are regulated by the Monetary Authority of Singapore (MAS) to conduct asset management functions (Roesti, Wettstein, 2014).

1.1 Concept of Expanding Consumer Base

There are two key areas that firms have to look into when they want to increase their customer base - Customer Acquisition and Customer Retention. Customer acquisition focuses on attracting new customers to the company while customer retention focuses on engaging existing customers.

Comparing these two methods, studies have shown that more focus has been placed on customer acquisition compared to retention, with 44% of companies focusing on customer acquisition and only a mere 18% place their emphasis on customer retention (Barker, 2018). In the context of the business environment today, the existence of variety has made customer retention one of the biggest challenges faced by businesses (McEachern, 2020). Despite efforts by businesses to improve customer retention, emergence of competitors due to fast growth in the region (PwC, n.d.) and changing market trends are making it even more challenging for them to engage existing customers.

1.2 Importance of Customer Retention in Asset Management Firms

Although businesses can essentially expand consumer base through customer acquisition, customer retention remains extremely important due to the benefits that can be reaped from it.

Firstly, compared to acquiring new customers, retaining customers incurs a significantly lower cost, with acquiring new customers costing five times more (Saleh, 2015). Since acquiring new customers drains a significantly higher amount of resources from businesses, it will be important for businesses to ensure that they have a loyal relationship with existing customers so as to prevent incurring extra costs for acquiring new customers. Studies have shown that increasing customer retention by 5% can increase profits by 25-95% (Landis, 2020).

Secondly, the chances of selling to a retained customer is significantly higher than selling to a new customer, where research showed a 60% and 20% chance respectively (Landis, 2020). As such, lacking customer retention can be costly for the company, in terms of having to spend more on acquiring new customers as well as the potential loss of profits that could be generated with customer retention.

On top of that, retained customers are more likely to refer their friends and family, bringing in new customers for the company (Dhami, 2020). This helps companies to acquire new customers, which might otherwise have to be done by spending a large amount of resources on marketing and advertisements.

This highlights the importance of customer retention amongst businesses and how it can be detrimental to their profits should customer retention be overlooked. This is especially so within the asset management industry, where issues relating to customer retention is considered to be highly rampant. Based on a research conducted by KPMG, customer retention is highly dependent on the six pillars of customer experience excellence – personalisation, integrity, expectations, resolution, time and effort, and empathy, due to their close alignment to the basic human psychological drivers (Brown, 2017).

Taking a closer look at the six pillars of customer experience excellence, with the focus placed on personalisation and expectation, personalisation makes use of individualised attention to help drive an emotional connection with its customers. In this sense, customers feel valued, leading to greater retention rates. In addition, expectations relate to the business managing, meeting and exceeding expectations arising from their customers. To do so, businesses need to obtain a deeper understanding of their customer bases to enable them to tailor each experience specifically to the expectations of their customers. Hence, by placing emphasis on the six pillars of customer experience, it would help businesses to frame their approaches and marketing strategies in order to best retain their customers.

1.3 Rethinking the Problem: Targeted Customer Retention

While it is true that retaining customers incurs a lower cost on companies, customer retention may propose an underlying problem for the company- which is the fact that it is difficult for businesses to estimate the cost associated with customer retention.

It should be noted that not all marketing efforts made to boost customer retention rates are effective and will help the company achieve its desired outcome. The fundamental principle behind marketing efforts is the receptiveness of the target audience to such efforts, which will help in the development of a holistic communication strategy (Lorenzon and Pilotti, 2008). With customers who are unreceptive to marketing effort, it will result in the inefficient use of resources as the desired outcomes are not reaped from the inputs. Furthermore, the resources may be diverted away from areas which could possibly bring about more benefit to the company.

With that being said, it is important to identify the receptiveness of customer segments in order to minimise the cost associated with customer retention. Targeting of customers who are likely to leave, and providing them with a marketing style specially tailored to their needs can therefore increase the success of such efforts, thereby reducing the cost associated with customer retention.

2 Analytics Solution

2.1 The Opportunity: Use of Customer Retention Analytics to Tackle Retention

Customer retention analytics has been providing crucial information to businesses. Companies that leverage on these customer data might be able to see profits improve up to as high as 126% (McKinsey, 2014). These data analytics are beneficial to businesses as they are able to generate predictive metrics of their customers' churn, allowing them to foresee which group of customers are more likely to stop supporting their products, thereby coming up with measures beforehand to appeal to these groups of customers. However, based on our research, it has been highlighted that a significant percentage of businesses are either unsatisfied with their current use of customer analytics or lack the proficiency to analyse existing customer data (Forrester, 2017).

2.1.1 Difficulties Faced by Companies For Customer Retention

As mentioned, customer retention remains one of the greatest challenges for businesses.

Firstly, one of the challenges faced by businesses is understanding their customers (King, 2020). With changing market trends and the advancement of technology, companies find it difficult to understand the needs of their customers. However, this issue is pressing as the presence of competitors gives rise to variety for the customers, which makes it harder for companies these days to retain their customers (Berne, 2020). With so many alternatives available, customers can easily turn to other options if they feel unhappy with the company.

Secondly, another challenge faced by businesses is selecting the right tools to serve the customers (King, 2020). With the presence of many different social media platforms available, identifying the most effective platform has been one of the challenges faced by companies while trying to retain customers.

2.1.2 Approach

As such, in this project, we aim to tackle the business problem identified earlier and to better utilise data analytics for customer retention. By analysing customers' demographics and their receptibility towards past marketing efforts, we hope to churn out a predictive model which will allow us to better foresee customer churn as well as to understand the needs of different customer groups. With this, companies can better foresee which groups of customers are more likely to leave and as such come up with more appealing ideas to retain them. Therefore, companies can make use of such information to better allocate marketing funds to appeal to different demographic groups. Our solution would therefore directly tackle the challenges that firms face for customer retention.

2.2 Project Feasibility

In order to tackle the business problem with the help of predictive models, the feasibility of the project has to first be determined through the identification of suitable characteristics.

Firstly, the business problem encompasses a predictive need, where the potential of customers being retained based on their demographics will allow for better segmentation of the customer base and allow for better identification of the ideal promotional method. Furthermore, the prediction of the most suitable marketing channel will enable greater personalisation to cater to each individual customer.

Secondly, knowledge is deemed to be imperfect as customer retention is not an issue which can be determined accurately by businesses alone. More often than less, the success of customer retention largely lies on the end of the customers as they are the ones who determine the factors which would help them decide whether they wish to stay on or leave. Hence, the use of a predictive model will better allow for the identification of patterns and generate insights for the businesses to consider.

Lastly, there is the availability of a training data set. In this case, we have made use of historical data obtained from publicly available datasets to derive associations and patterns to help us generate the predictive model. To predict our output variables, we have made use of a mix of relevant variables which are available in the following .csv files:

- BankChurn.csv : a publicly available dataset on Kaggle
  - Bank Churn for dataset 1
- BankChurn2.csv : BankChurn.csv combined with another publicly available dataset on Kaggle
  - Marketing Campaign for dataset 2

2.3 Desired Outcomes

The focus of our analysis would be to predict the success of customer retention efforts among existing customers and devising a personalised strategy for these customers for improved retention. Through this, we hope to retain customers by converting customers with the greatest potential of leaving to customers who would potentially stay on with the help of marketing efforts. Hence, the desired business outcomes we hope to achieve are as follows.

Firstly, we hope to identify customers who display the highest chance of leaving despite past efforts, with the demographics of the customers taken into account. By developing a predictive model which can help in identifying customer segments which have lower retention rates, it would then enable a more personalised marketing method to be devised, enabling better resource allocation.

Next, we hope to identify the most preferred marketing approach relating to each customer segment and create a personalised strategy to meet their specific needs. Similarly, the development of a predictive model to identify the various approaches would allow for more targeted marketing and better planning for resource allocation.

In addition, we also help to identify customers who are likely to be retained through personalisation. In this case, the predictive model developed will determine if the customers are receptive to the use of personalisation or not, allowing the company to set a future direction for their marketing strategies.

Lastly, we hope to improve current marketing efforts to increase customer retention rates in the company. Through the determination of the more receptive market segments and preferred marketing approach, it would allow for the company to identify areas of emphasis and possible strategies to adopt to effectively retain their existing customers.

3 Data Cleaning and Exploration

This section outlines our data cleaning process, where we retain only variables and data that are deemed to be relevant to our analysis, and correct any irregularities or errors. Subsequently, data exploration was done on the cleaned dataset through visualisation with the use of the ***ggplots2*** package for us to gain a better understanding of our dataset.

3.1 Data Cleaning Decisions

3.1.1 NA Values

NA values were present in both datasets and were removed since they occupy an insignificant amount of rows in the dataset (70 out of 10,070 rows).

3.1.2 Erroneous Values

Some erroneous values were identified in both datasets and were removed:

1. CustomerId : “TESTING”, “TEST”, “TEST123”, “TEST456”
2. Balance: “TESTING”, “TEST”, “TEST123”, “TEST456”

3.1.3 Duplicate Rows

Moreover, the CustomerId column is supposed to be unique to each customer. However, duplicates were found in some rows of the dataset. Thus, these duplicate rows were dropped from the dataset.

3.1.4 Redundant Variable

Variables	Reason
RowNumber	This variable has no value to our analysis as it is simply the count of the rows in the dataset.
Surname	This variable does not help in our analysis as this variable is unable to uniquely identify the customers in the dataset, unlike CustomerId which was retained.

3.1.5 Converting variables to the relevant data type

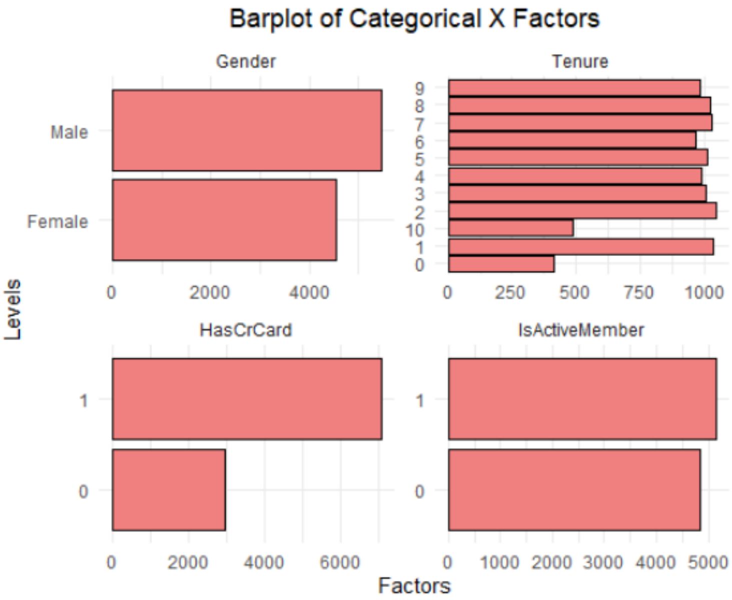
After understanding and looking through the variables in the dataset, the variables are converted into 2 formats, categorical and numerical which can be seen in the Appendix.

3.2 Data Exploration

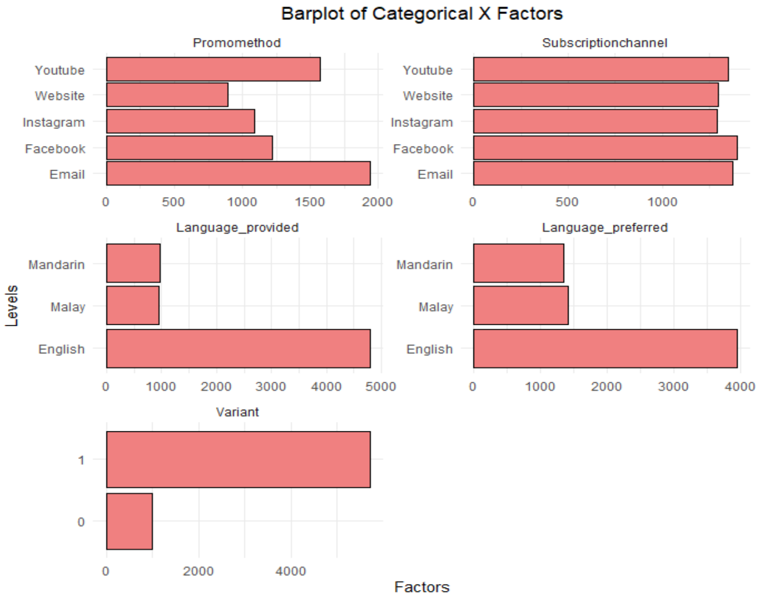
Using ggplot2, we did data exploration on the datasets to better understand the datasets. Selected findings that are interesting will be discussed here and the remaining plots can be found in the Appendix.

3.2.1 Univariate Graphs

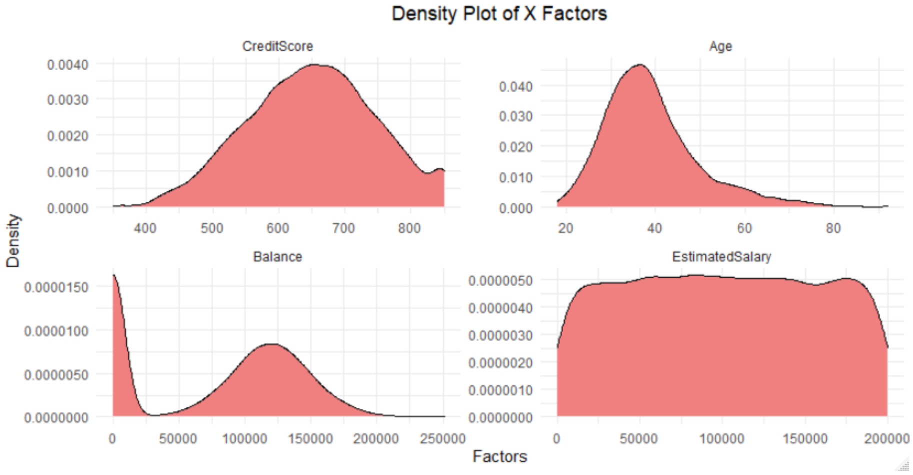
Dataset 1 & 2: ***Categorical Factors*** include *'Gender'*, *'HasCrCard'*, *'IsActiveMember'* and *'Tenure'*. An interesting observation is that there is an almost equal number of active members and non-active members.



Dataset 2: ***Categorical Factors*** includes *'Promomethod', 'Subscriptionchannel', 'Language_provided', 'Language_preferred'* and *'Variant'*. Most people were first introduced to the company by email.

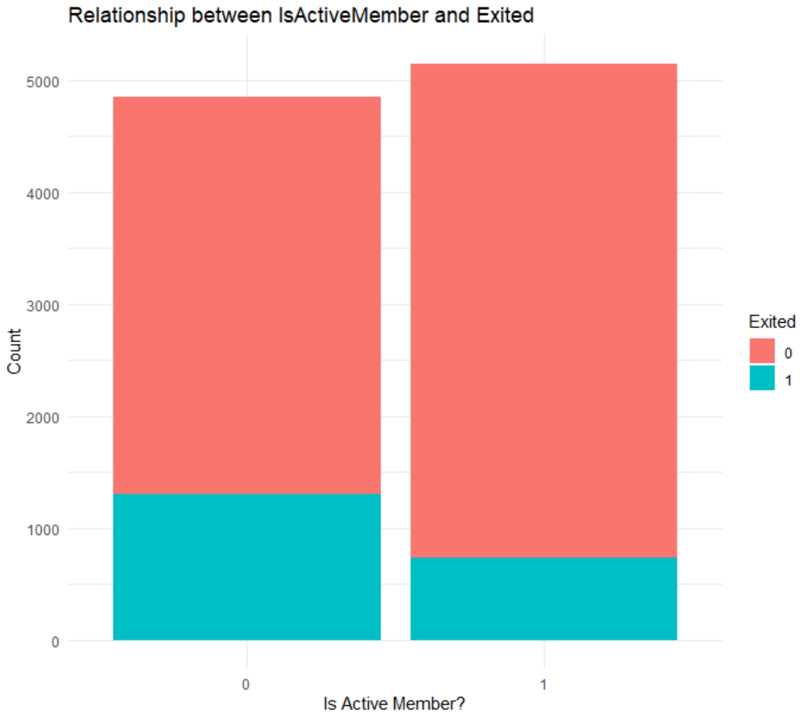


Dataset 1 & 2: ***Numerical Factors*** include *'Age', 'CreditScore', 'Balance'* and *'Estimated Salary'*. Interestingly, for estimated salary, the distribution is quite evenly spread.



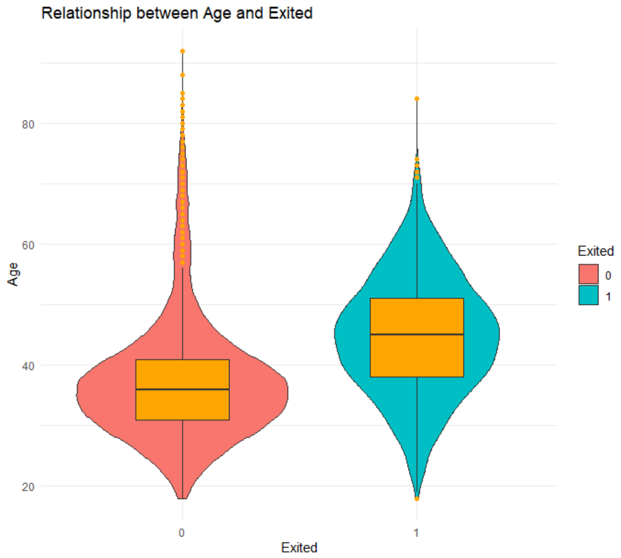
3.2.2 Bivariate Graphs

Dataset 1 & 2: ***Categorical vs Categorical Factors*** include *'HasCrCard vs Exited', 'IsActiveMember vs Exited', 'Gender vs Exited'* and *'Tenure vs Exited'*. Unsurprisingly, a higher number of people that exited were not active members.



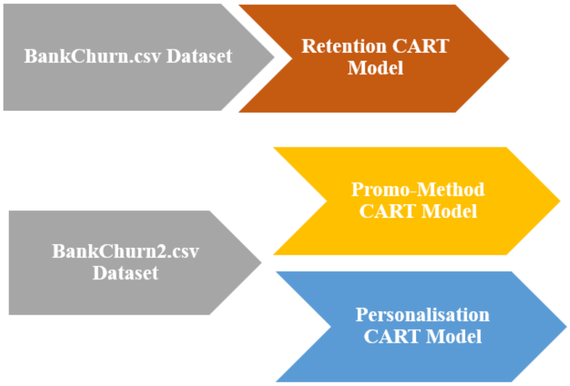
Dataset 2: ***Categorical vs Categorical Factors*** include *'Promomethod vs Exited', 'Subscriptionchnanel vs Exited', 'Language_provided vs Exited', 'Language_preferred vs Exited'* and *'Variant vs Exited'*

Numerical vs Categorical Factors include *'Age vs Exited', 'CreditScore vs Exited', 'Estimated Salary vs Exited'* and *'Balance vs Exited'.* Surprisingly, the difference of the median age between customers who stayed and exited is not very large.



4 Implementation & Recommendation

An overview of our process of training the three models is summarised in the flowchart below.

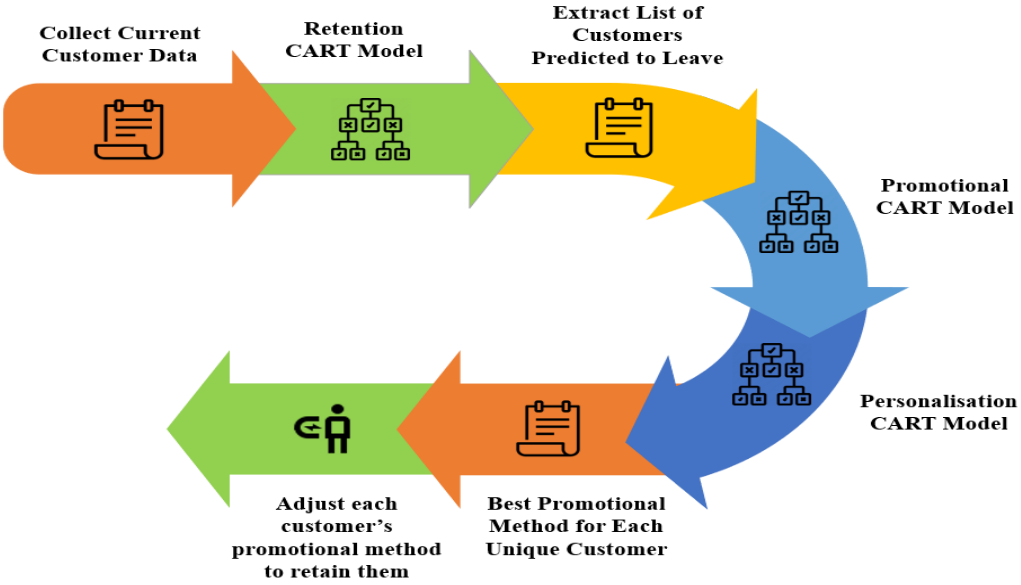


Firstly, a range of data, with demographics and variables that can possibly affect a customer's probability of retaining in the company, will be determined from the given "BankChurn.csv" dataset. This set of data will then be used to train the first predictive model (to be discussed in Section 5.1) that will be used to predict if customers, with similar demographics, will leave or stay with the company.

Next, from our second dataset 'BankChurn2.csv', we extracted out the demographics of customers who are successfully retained and this will be used to train the next two models. The next two models trained will determine the most ideal marketing strategy which will increase the likelihood of the customers staying. With respect to this current project, we will be focusing on predicting the most suitable subscription channel for each customer, as well as whether personalisation will be a factor in affecting their stay as a customer in the company. The models trained will also be analysed further to gain insights for future improvements of the marketing strategies.

Proposed Solution

An overview of our proposed analytical process to enhance the rate of customer retention in White Rock is summarised in the flowchart below.


Collect Current Customer Data
Retention CART Model
Extract List of Customers Predicted to Leave
Promotional CART Model
Personalisation CART Model
Best Promotional Method for Each Unique Customer
Adjust each customer's promotional method to retain them

Upon training the three models, White Rock can start collecting customer data from their current customers with the current necessary variables reflected in each of the datasets. White Rock can then pass these new customer data through the first predictive model, to predict which customers would stay or leave the company. The list of customers predicted to leave can then be passed through the second and third model to have a predicted ideal "SubscriptionChannel" and "Personalisation" tagged to them. The company can then proceed to implement respective measures for each of the customers. The other list with the customers predicted to stay will be kept and monitored to see if they actually stay. These data will also be used to update the trained models to potentially increase its accuracy. (To be discussed in Section 6.2.2)

4.1 Implementing Predictive Model to Predict Customer's Likelihood of Staying

4.1.1 Model Evaluation

Logistic Regression

A logistic regression model was chosen as one of the predictive models to determine the binary variable - customer's likelihood of staying. The best logistic regression model, which consists of solely statistically significant variables that have been determined through the analysis of odds ratio and p-values,was fitted using the *glm()* function in R. Details of the model, such as the model fitting process, can be found in Appendix B.

CART model

A CART model was also considered as one of the predictive models that could predict the binary variable. The R package rpart is used to develop the CART model. An unpruned classification tree is grown based on binary splits which minimises the weighted average Gini impurity of resultant child nodes. The output decision tree is then pruned using minimal cost complexity pruning (one standard deviation rule) to prevent overfitting of the data. Details of the model, such as the model fitting process, can be found in the following sections 4.1.2:

Upon training and testing both predictive models, this is what we have found out in our respective models:

Accuracy	Logistic Regression	CART
Train Set	0.843	0.8604286
Test Set	0.839333	0.8593333

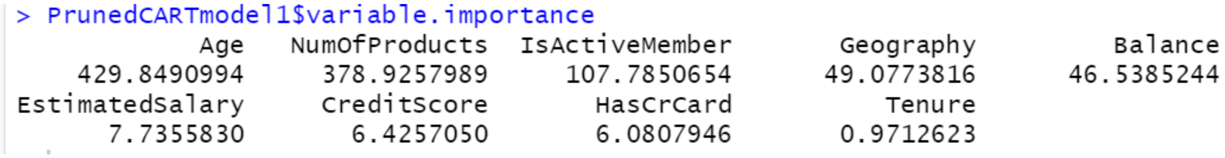
As there is an increase in accuracy for both train and test for the CART model, as well as being able to gain more insights from the CART model, this model will be the main predictive model used for all 3 scenarios. However, the logistic regression model will still be used as a reference to gain additional insights, as well as to proof check the CART model.

4.1.2 Implementation of CART and Insights

(Details regarding the Model will be under Appendix as it is simply too large)

Before training this model, we have firstly removed all of the irrelevant variables (mentioned in Section 3.1.4) The first CART model was then trained based on the remaining variables in the customer dataset with *'Exited'* being the predicted variable.

Subsequently, using the minimal cost complexity approach, the CART model was then pruned to prevent overfitting of the data. With this CART model, we will then be able to pass through future new customer data to predict their likelihood of staying or leaving solely based on these demographics.



```
> PrunedCARTmodel1$variable.importance
            Age   NumOfProducts  IsActiveMember       Geography         Balance 
    429.8490994     378.9257989     107.7850654      49.0773816      46.5385244 
EstimatedSalary     CreditScore       HasCrCard          Tenure 
      7.7355830       6.4257050       6.0807946       0.9712623 
```

Aside from the trained CART model, we can also gain insights and gain indirect benefits into improving our marketing strategies for customer retention.

In this particular model, the 9 variables (*Age, IsActiveMember, CreditScore, Tenure, EstimatedSalary, NumOfProducts, Geography, HasCrCard* as well as *Balance*) are used in the splitting of the optimal CART tree. We have also found out that Gender, although deemed potentially relevant from our data exploration, is not utilised at all in this pruned CART model.
Other than which, in analysing the CART model, emphasis is placed on important stopping rules for deciding when a branch is terminal. Important stopping rules are decision rules that results in confident classifications, and hence terminal nodes with the highest purity. Upon examination of the results, the following stopping rules have been identified to be of the most importance with the chance of misclassification error being the lowest:

For Predicted Exited = 1,
61) Geography=Germany 179   37 1 (0.20670391 0.79329609) *
31) Age>=49.5 376   42 1 (0.11170213 0.88829787) *
241) Balance< 81819.96 134   47 1 (0.35074627 0.64925373) *

To give an example of how the analytical analysis of this pruned tree could be done, White Rock could observe how *'Geography'* = Germany plays an important role in determining cases of non-customer retention, hence this could be one of the reasons and evidence for them to either allocate more resources in bettering the customer retention for customers in Germany. Age, coupled with its variable importance from before, might be an extremely strong key in determining the cases of non-customer retention. Hence, in that case, likewise, White Rock could invent an entirely new marketing strategy in its subscription channel specifically targeting older customers to increase its customer retention rate.

For Predicted *'Exited'* = 0, the rules are mostly the opposites as for *'Exited'* = 1 (Eg. Balance > 81819,96 being highly important in predicting *'Exited'* = 0, as well as *'Geography'* = France, Spain which are the other two classes in our *'Geography'* variable playing a crucial role as well, hence further strengthening the respective variables' importance in determining the cases.

Other potential predictors which are not as crucial in the splitting include *'Tenure', 'EstimatedSalary', 'CreditScore'*; these variables' effect could be unintuitive, or even counterintuitive with our common knowledge and should be approached with professional scepticism or recombined with domain knowledge from relevant experts for reliable interpretation, as they could be indicative of a poor predictive model.

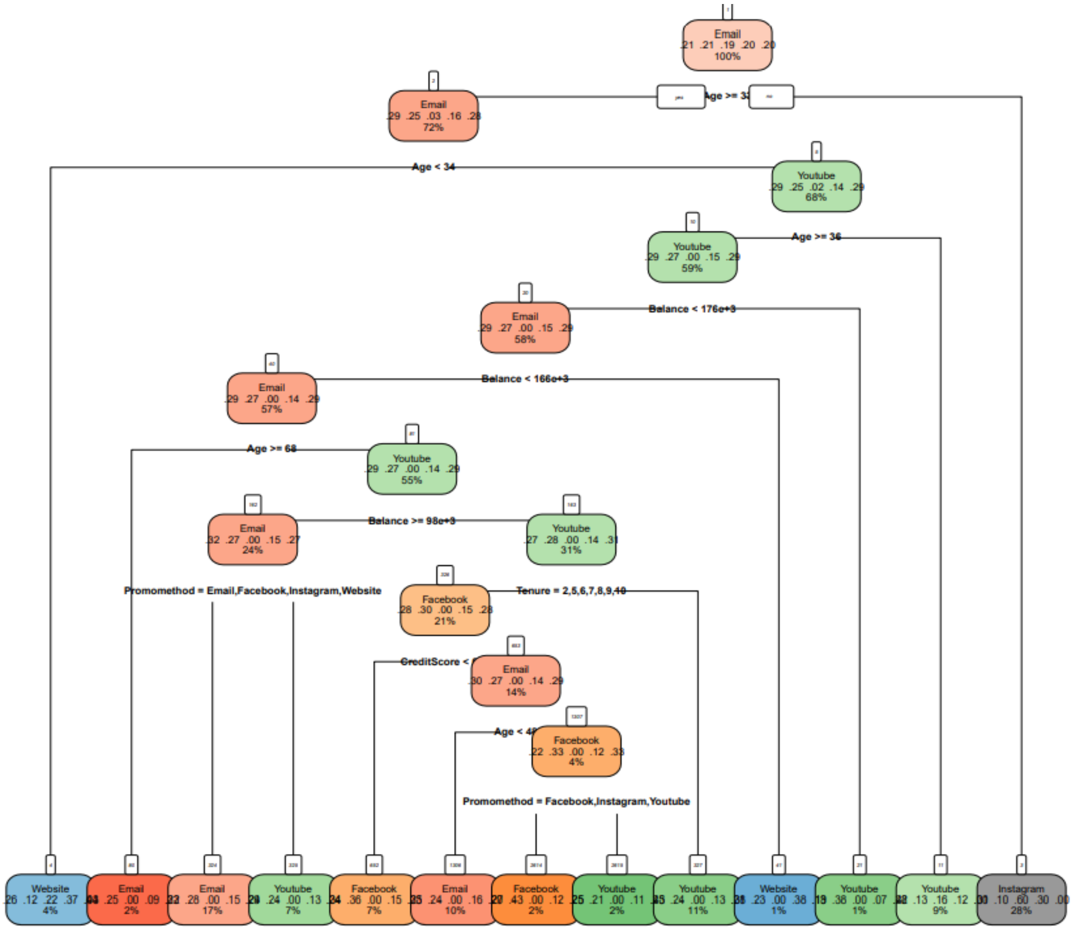
4.2 Implementing CART Model to Determine New Promotional Method for Leaving Customers

4.2.1 Model Evaluation

Upon training and testing both predictive models (similar approach in Section 4.1.1), we obtained an accuracy of
(Details of logistic regression can be found in Appendix B)

Accuracy	Logistic Regression	CART
Train Set	0.3659643	0.421179
Test Set	0.3285625	0.4150591

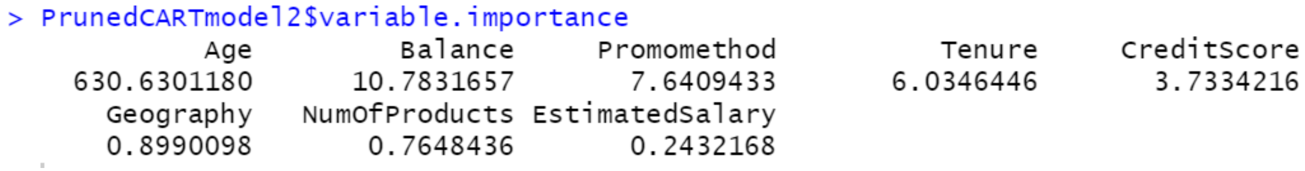
4.2.2 Implementation of CART and Insights


Email .21 .21 .19 .20 .20 100%
Age >= 33
Email .29 .25 .03 .16 .28 72%
Age < 34
Youtube .29 .25 .02 .14 .29 68%
Age >= 36
Youtube .29 .27 .00 .15 .29 59%
Balance < 176e+3
Email .29 .27 .00 .15 .29 58%
Balance < 166e+3
Email .29 .27 .00 .14 .29 57%
Age >= 68
Youtube .29 .27 .00 .14 .29 55%
Balance >= 98e+3
Email .32 .27 .00 .15 .27 24%
Youtube .27 .28 .00 .14 .31 31%
Promomethod = Email,Facebook,Instagram,Website
Facebook .28 .30 .00 .15 .28 21%
Tenure = 2,5,6,7,8,9,10
CreditScore <
Email .30 .27 .00 .14 .29 14%
Age < 48
Facebook .22 .33 .00 .12 .33 4%
Promomethod = Facebook,Instagram,Youtube
Website .26 .12 .22 .37 4%
Email .25 .00 .09 2%
Email .28 .00 .15 17%
Youtube .24 .00 .13 7%
Facebook .36 .00 .15 7%
Email .24 .00 .16 10%
Facebook .43 .00 .12 2%
Youtube .21 .00 .11 2%
Youtube .24 .00 .13 11%
Website .23 .00 .38 1%
Youtube .36 .00 .07 1%
Youtube .13 .16 .12 9%
Instagram .10 .60 .30 .00 28%

The model above was trained using all of the variables in bankchurn2 (the 2nd dataset) excluding the irrelevant variables mentioned in Section 3.1.4, as well as *'SubscriptionChannel'* as that is what we are predicting and *'Variant'*, as this is another variable that we want to predict and hence is not one of the variables used to train this model.

Similarly, using the minimal cost complexity approach, the CART model was then pruned to prevent overfitting of the data. With this CART model, we will then be able to pass through future new customer data whom customers are predicted to leave, in order to predict the most ideal subscription channel to increase their likelihood of retaining in the company.

As this is a multi-class classification tree, its accuracy (0.421179) is relatively lower as compared to a binary class classification tree. However, given the current volume of the dataset and the increased number of classes in the predicted variable, this CART model is still extremely essential and useful as it definitely increases the odds of the customers retaining with the tagged subscription channel as compared to one that is chosen at random.



```
> PrunedCARTmodel2$variable.importance
          Age         Balance     Promomethod          Tenure     CreditScore 
  630.6301180      10.7831657       7.6409433       6.0346446       3.7334216 
    Geography   NumOfProducts EstimatedSalary 
    0.8990098       0.7648436       0.2432168 
```

From the pruned CART model, although there were 8 variables involved in the tree as shown above, only 5 of them are used in the splitting of nodes in the CART model, namely *'Age', 'PromoMethod', 'Balance', 'CreditScore' and 'Tenure'*, hence indicating that, for the customer dataset that was given, these are the main variables in determining the subscription channel that a customer will prefer, which will increase his rates of retaining within the company.

Upon further analysing the splitting rules, we have deduced the following nodes and their respective insights to be particularly interesting and useful:

For predicting SubscriptionMethod = Email,
80) Age>=67.5 108   61 Email (0.44 0.25 0 0.093 0.22) *
For predicting SubscriptionMethod = Instagram,
3) Age< 32.5 1477 588 Instagram (0 0.096 0.6 0.3 0) *

Age proves to be an important role in determining the customer's most ideal subscription channel as shown from these extreme ranges of being higher than 67.5, as well as being below 32.5. This seems logical as it is intuitive to associate the elder customers to be more accustomed to receiving emails while the younger ones to be more active on a social media platform such as Instagram. Such an intuitive observation allows for the users to reliably interpret such information as it proof checks with our current domain knowledge.

For predicting SubscriptionMethod = Youtube,
324) Promomethod=Email,Facebook,Instagram,Website 927 625 Email (0.33 0.28 0 0.15 0.24) *
325) Promomethod=Youtube 350 231 Youtube (0.29 0.24 0 0.13 0.34) *

Promomethod also proves to be an extremely important variable in deciding the ideal subscription channel. As shown on node 325, the subscription channel and the promotion method to acquire that customer should be cohesive. This relationship could be explored further to obtain useful insights in improving the current marketing strategies. (To be discussed in Section 5) As such, the company could consider tying the promotion method and the subscription channel together to increase the likelihood of the customer retaining in the company.

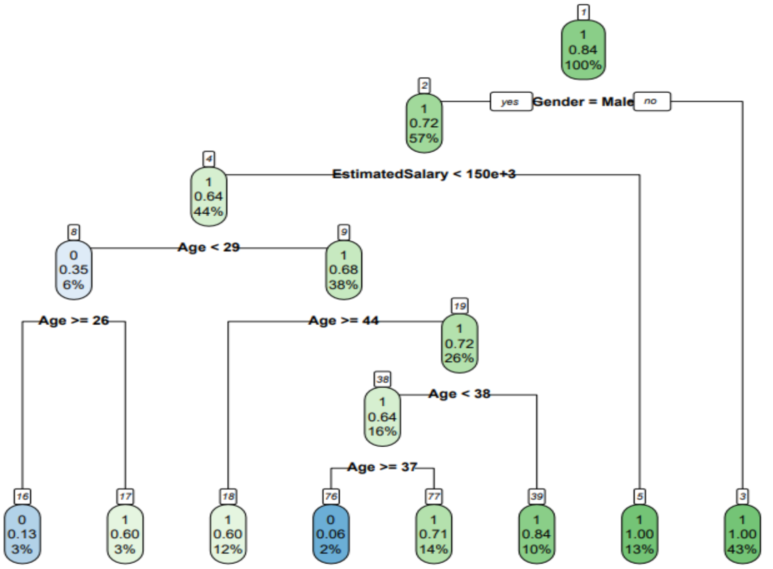
4.3 Implementing CART Model to Determine Need for Personalisation

4.3.1 Model Evaluation

Upon training and testing both models (similar approach in Section 4.1.1), we have obtained an accuracy of
(Details of logistic regression can be found in Appendix B)

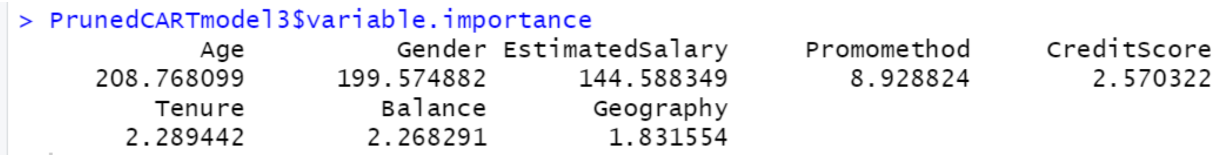
Accuracy	Logistic Regression	CART
Train Set	0.84	0.8797333
Test Set	0.8412204	0.882939

4.3.2 Implementation of CART and Insights



The model above was trained using all of the variables in bankchurn2 (the 2nd dataset) excluding the irrelevant variables mentioned in Section 3.1.4, as well as Variant as that is what we are predicting and SubscriptionChannel, as this is another variable that we want to predict and hence is not one of the variables used to train this model.

Similarly, using the minimal cost complexity approach, the CART model was then pruned to prevent overfitting of the data. With this CART model, we will then be able to pass through future new customer data whom customers are predicted to leave, in order to predict whether personalisation should be utilised in their subscription channel to increase their likelihood of being retained.



```
> PrunedCARTmodel3$variable.importance
            Age          Gender EstimatedSalary     Promomethod     CreditScore
     208.768099      199.574882      144.588349        8.928824        2.570322
         Tenure         Balance       Geography
       2.289442        2.268291        1.831554
```

As shown from this pruned 3rd CART model, the respective important variables are only as follows, *'Age'*, *'Estimated Salary'* and *'Gender'*. The other variables were not considered at all in this construction of the CART model as it was pruned a little earlier as compared to the other CART models, hence indicating that the aforementioned variables might be the ones which are a lot more influential as compared to the others, in determining whether a customer would change his mind if their subscription channel was coupled with *'personalisation'*.

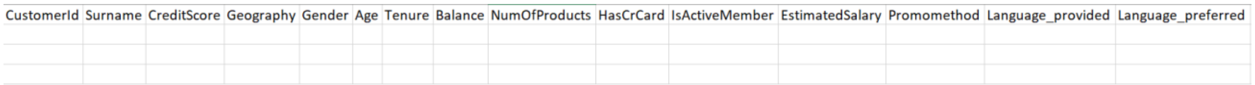
After further analysis of the splitting of the terminal nodes, terminal of the highest purity node came from "Gender = Female" as of node 3:
3) Gender=Female 2298   2 1 (0.000870322 0.999129678) *

This extreme phenomenon from the given dataset highly supports the concept that females do get influenced easily with the additional "personalisation" provided by the company in their subscription channels. This extreme observation is to be approached with professional scepticism or recombined with domain knowledge from relevant experts for reliable interpretation, as they could be indicative of a poor predictive model or a highly biased dataset.

4.4 Implementation of Solution

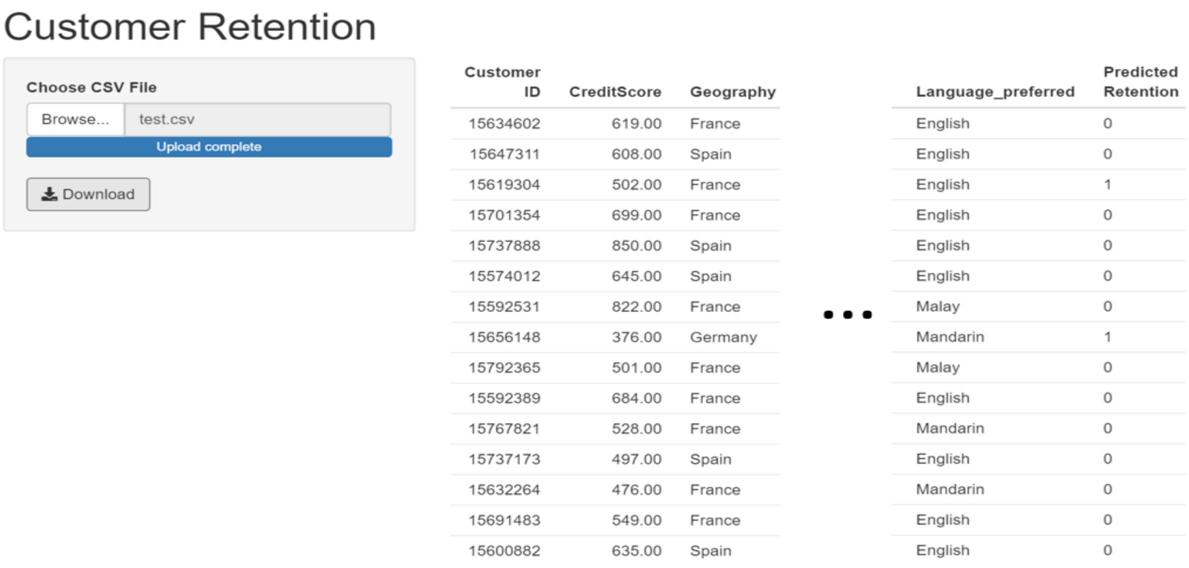
Our analytics model will be implemented through two simple and user-friendly interfaces which can be used by marketing managers to analyse customers retention.

Before the usage of the interfaces, it is suggested that White Rock marketing department collect customer data of the following variables as shown in the picture below. The data should be stored in the form of a .csv file in preparation to be uploaded into the interfaces for analysis.



CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Promomethod	Language_provided	Language_preferred

At each point of application, the user (eg. White Rock marketing department) should upload a .csv file into the first interface that predicts customer retention, this interface is implemented from our first model, which predicts the likelihood of a customer staying in the company. Customers who are predicted to stay will be marked with '0' under *'Predicted Retention'* and customers who are predicted to leave will be marked with a '1' under *'Predicted Retention'*. The user can then download the new table into a .csv file, which will automatically filter out those customers who are predicted to leave.



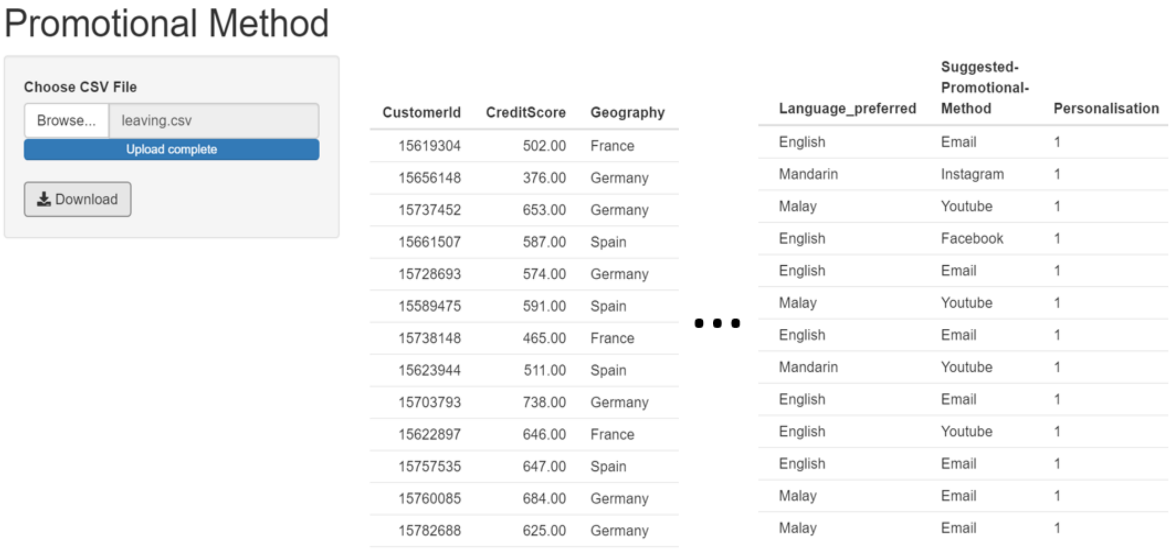
Customer Retention

Choose CSV File

test.csv

Customer ID	CreditScore	Geography		Language_preferred	Predicted Retention
15634602	619.00	France		English	0
15647311	608.00	Spain		English	0
15619304	502.00	France		English	1
15701354	699.00	France		English	0
15737888	850.00	Spain		English	0
15574012	645.00	Spain		English	0
15592531	822.00	France	• • •	Malay	0
15656148	376.00	Germany		Mandarin	1
15792365	501.00	France		Malay	0
15592389	684.00	France		English	0
15767821	528.00	France		Mandarin	0
15737173	497.00	Spain		English	0
15632264	476.00	France		Mandarin	0
15691483	549.00	France		English	0
15600882	635.00	Spain		English	0

With the new .csv file - containing the list of customers who are predicted to leave - downloaded from the first interface, the user can upload this new .csv file into the second interface which is implemented from our second and third CART model to help predict the suggested promotional method and whether personalisation is required.



Promotional Method

Choose CSV File

leaving.csv

CustomerId	CreditScore	Geography
15619304	502.00	France
15656148	376.00	Germany
15737452	653.00	Germany
15661507	587.00	Spain
15728693	574.00	Germany
15589475	591.00	Spain
15738148	465.00	France
15623944	511.00	Spain
15703793	738.00	Germany
15622897	646.00	France
15757535	647.00	Spain
15760085	684.00	Germany
15782688	625.00	Germany

• • •

Language_preferred	Suggested-Promotional-Method	Personalisation
English	Email	1
Mandarin	Instagram	1
Malay	Youtube	1
English	Facebook	1
English	Email	1
Malay	Youtube	1
English	Email	1
Mandarin	Youtube	1
English	Email	1
English	Youtube	1
English	Email	1
Malay	Email	1
Malay	Email	1

We recommend that the marketing department constantly update the .csv files with new changes that they have made from the previous predictions so as to get the latest predictions for each customer. This interface can be reviewed semi-annually according to White Rock's need to get the retention information.

Details of the implementation of the interface can be found in Appendix E.

5 Literature Review and Expert Opinion

To maximise the results we obtained from the predictive models, our outcomes from the models should be complemented with literature reviews and expert opinions. This section aims to briefly discuss some important predictors highlighted by researches about the asset management industry in order to provide substantial insights to our analysis.

Research has shown that asset management companies must be aware of geographical differences between cross-border customers in order to provide a better client experience (Deloitte, 2020). Based on past studies, overall maturity of digital transformation in each region plays an important role in client decision. It is found that continents that are less mature in digital transformation, such as Europe, customers are less price sensitive and place more importance on personal relationships with wealth managers. Also, studies have shown that in continents such as Asia and North America, cultural considerations are key for customers which requires the need for relationship managers that are familiar with local specificities. However, this is not evident in our results as there are only 3 countries from the same continent found in our dataset. Even so, there is still a slight distinction as 'Geography' is a factor used in our models.

Experts have found differences in the nature of customer loyalties between male and female (Melnyk, 2009) and coherent pattern of gender differences in customers' responses to different types of psychological rewards in the context of loyalty programs was observed. Based on studies conducted, women respond more positively than men to loyalty programs that emphasize personalisation. In contrast, men respond more positively to loyalty programs that emphasize status. This is consistent with our results above, where it was observed that females get easily influenced by the "personalisation" factor added in by companies.

According to past studies, individual risk tolerance is found to decrease by 1.7% with every year of increase in age (Sahm, n.d.) and this trend can be attributed to the decreasing amount of time for loss recuperation (Gustafsson and Omark, 2015). Hence, this is consistent with our results as it was observed that customers falling in the higher age group are more likely to leave.

Also, cohesiveness of marketing plans has been proven to bring about higher effectiveness when reaching out to consumers (McCormick, 2020). The more cohesive and consistent the marketing is, the more the customers are able to understand the message behind it, thereby effectively engaging them. Having a coherent subscription and promotional channel creates a more convenient and personalised experience for customers (Dolan, 2020). Looking at success stories such as Venus, where customer conversion and retention rates have increased after increasing cohesiveness of their channels, it is evident that our results are supported.

6 Limitations & Future Directions

6.1 Limitations

Firstly, while the predictive model is able to predict the best promotional method for individual customers based on their demographics, the receptiveness of the customer towards the identified promotional method cannot be determined.

Secondly, since customer trends change over time, even if the model is able to be retrained and can predict effectively from its current trained data, it might not be as accurate for new customer data that comes in.

Thirdly, based on the current model, the variables used are fixed. Hence, this makes the model inflexible as it is unable to accommodate other factors which may have an impact on the results.

Fourthly, it is important to note that the characteristics of customers across various companies may differ. In order to obtain the optimal prediction, the model needs to be trained based on the company's own data so as to factor in the difference in characteristics of customers within the company.

Finally, despite the predictive model providing the best possible prediction, it may not produce the most accurate results as there may be other factors which can possibly have an impact on customer retention. Hence, it can be supplemented by other analytics algorithms to improve the overall accuracy of prediction.

6.2 Future Directions

Due to the presence of potential limitations, the company can make use of other types of data and take into consideration various forms of analysis to improve the strategies which have been implemented. This would therefore help the company identify new methods it can potentially adopt in the future.

6.2.1 Usage of Sentiment Analysis on Customer Feedback

Apart from individual customer's analysis, it is suggested that the marketing department can work together with the customer service department to implement a sentiment analysis algorithm which can help to analyse the general customers' feedback on the company's service and promotional method. The company can feed the feedback to the sentiment analysis algorithm which uses text mining to recognise words such as 'customer service' or 'quality', etc. This helps the company to gain a better understanding of the reasons why customers are staying or leaving without having much biases as they are assessed objectively by the algorithm. In fact, the top reason why customers tend to leave is because they feel that the company does not care about them (MacDonald, 2020). As such, it is important that the company analyses and works on the feedback of customers, showing them that their opinions are valued.

The sentiment analysis will be able to help the company identify emotions conveyed by the customer through the feedback which they have provided. Based on included words and associated sentiments in the users' feedback, the sentiment analysis method would assign a sentiment score to them, finally returning the end results, where a higher score would portray a positive feedback. As implementing sentiment analysis might be tedious and sometimes complicated under instances where users use sarcasm in their comments (Countants, 2020), we can harness the help of emotion detection systems, such as Lexicon, or other more complex machine learning algorithms (MonkeyLearn, n.d.). With that, the company would be better able to identify the direction they can take according to the sentiments of customers.

With better knowledge of what the customer views of the company, the marketing department can make necessary amendments to their promotional method and this can potentially allow us to include more variables and classes in future models in hope of developing a more accurate and useful predictive model that helps to predict customer retention.

6.2.2 Constant Update of Model with Improvements from Past Predictions

To improve the accuracy of the model, it can be constantly updated with the current status of customers who went through marketing changes based on past prediction. This can be done by updating the first model with predictions from new additional data added each time. White Rock can also consider looking into the variable importance of each model as it is updated regularly. By doing so, it would help improve the model as the actual outcomes would be incorporated into the predictive models over time which would help boost the accuracy of future predictions.

6.2.3 Additional Models to Predict Other Factors

One of the current constraints of the models is the inflexibility to account for future factors which may impact the outcome of customer retention. These factors are those that may arise from the implementation of measures from the prediction models. The inability to account for these factors, which might be deemed significant to the company, may affect the prediction outcome in the long run due to volatile customer trends. Hence, by devising additional models for the prediction of additional factors that may arise in the future, it would help the company make more accurate predictions suited to the needs of the company.

6.2.4 Extend the Usage of the Model to Customer Acquisition

While the current focus is to increase the customer retention rate in the company, the ultimate goal in the long run would be to expand its customer base. As mentioned previously in Section 1.1, the expansion of the customer base comprises 2 key areas, customer retention and customer acquisition. By making use of the insights generated from the current model, such as the preferred promotional method, it can then be applied to customer acquisition in the future in order to predict the best promotional method to effectively acquire new customers.

7 Conclusion

Despite the potential cost associated with the use of data analytics, it could potentially generate greater insights which would have been undetermined previously. Moreover, it would provide significant time savings for employees as the manual process of searching and analysing data would be eliminated.

With that being said, our solution was crafted to reap the benefits brought about by data analytics. Through the initial identification of customers who are of higher likelihood of leaving based on their demographics, the company would be able to obtain the best promotional method tied to each of these customers in order to decrease the likelihood of leaving. Furthermore, the need for personalisation for each customer can also be determined to allow the company to better plan their steps forward. Therefore, through the implementation of the data models, it would essentially help in the streamlining of the customer retention process, thereby increasing the effectiveness of customer retention.

APPENDIX

Appendix A - Data cleaning and exploration

<u>Data dictionary</u>

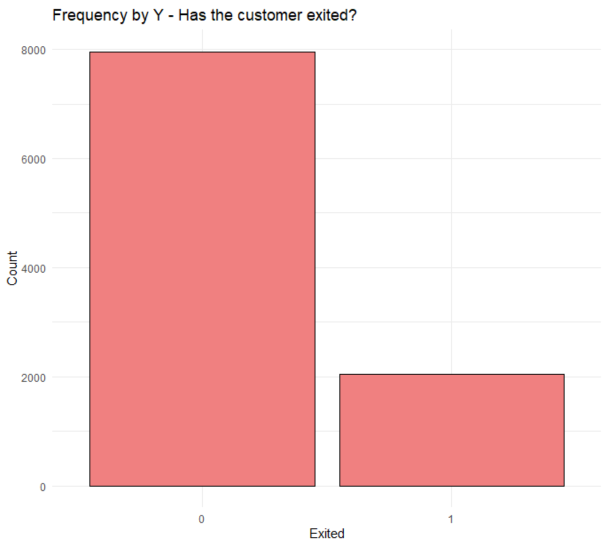
Variables/Factors	Description	Datatype (Categorical/Numerical)
CustomerId *(Dataset 1 & 2)*	Used to identify customers uniquely	Numerical
Credit Score *(Dataset 1 & 2)*	Credit score of customer in the bank	Numerical
Geography *(Dataset 1 & 2)*	The country where customers register their bank account under *(France, Germany, Spain)*	Categorical
Gender *(Dataset 1 & 2)*	Gender of customer *(Male, Female)*	Categorical
Age *(Dataset 1 & 2)*	Age of customer	Numerical
Tenure *(Dataset 1 & 2)*	How long has the customer been with the company *(0-10)*	Categorical
Balance *(Dataset 1 & 2)*	How much the customer have in their account	Numerical
NumOfProducts *(Dataset 1 & 2)*	Number of products that the customer uses *(1, 2, 3, 4)*	Categorical
HasCrCard *(Dataset 1 & 2)*	Whether the customer owns a credit card or not *(0,1)*	Categorical
IsActiveMember *(Dataset 1 & 2)*	Whether the active member is an active member or not *(0,1)*	Categorical
EstimatedSalary *(Dataset 1 & 2)*	Estimated salary of customers	Numerical

Exited *(Dataset 1 & 2)*	Whether or not the customer is retained *(0,1)*	Categorical
Promomethod *(Dataset 2)*	The method first used to attract customers *(Facebook, Instagram, Youtube, Email, Website)*	Categorical
Subscriptionchannel *(Dataset 2)*	The marketing channel used by the company to retain customers after they joined *(Facebook, Instagram, Youtube, Email, Website)*	Categorical
Language_provided *(Dataset 2)*	The language used in the marketing campaigns *(English, Mandarin, Malay)*	Categorical
Langauge_preferred *(Dataset 2)*	The language preferred by the customer to be used in the marketing campaigns *(English, Mandarin, Malay)*	Categorical
Variant *(Dataset 2)*	Whether or not the marketing ads are personalised *(Personalised, Control)* **Control = Generalised*	Categorical

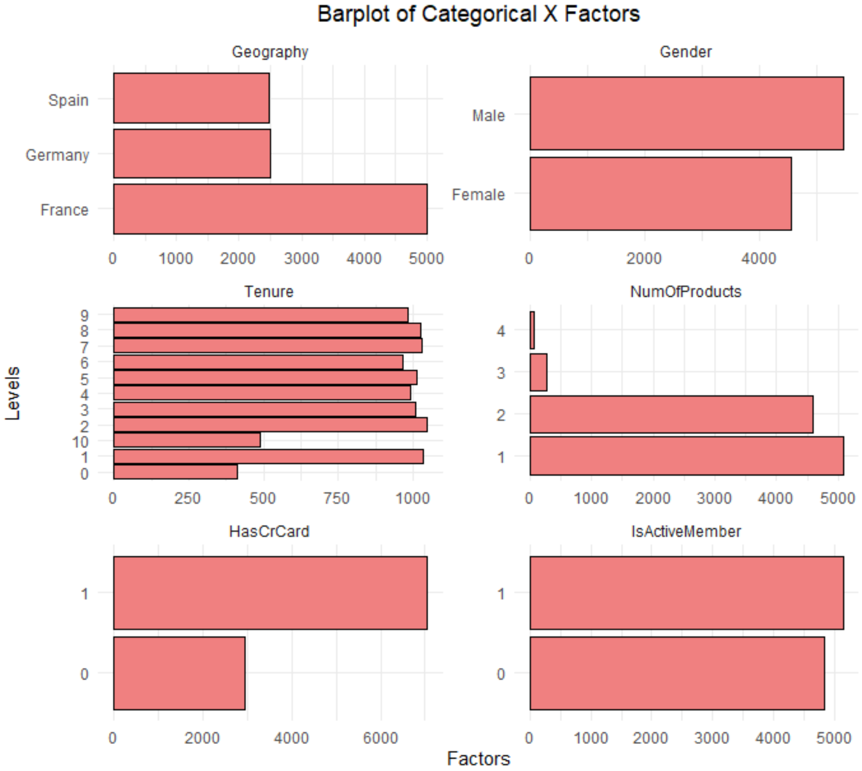
Appendix B - Data Visualisation

<u>Univariate: Categorical</u>

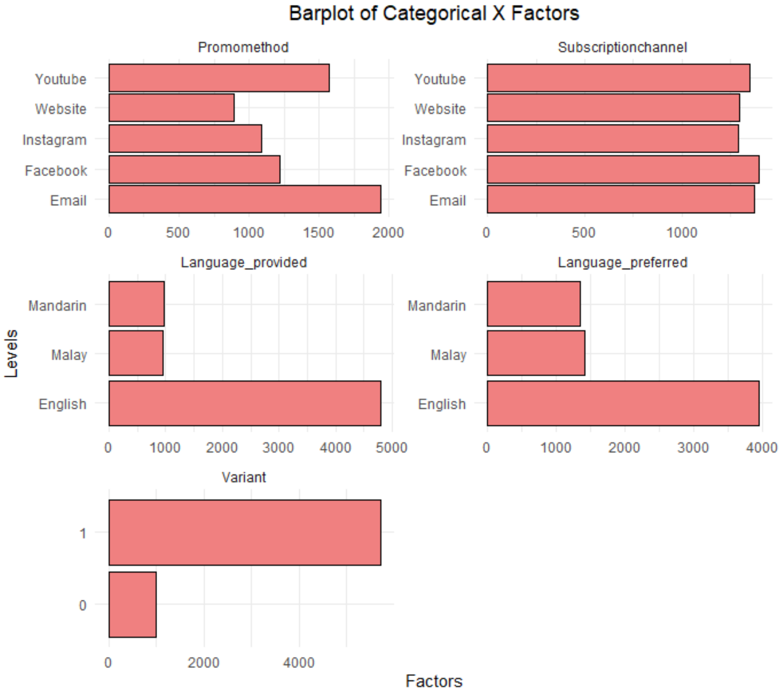
Exited:



HasCrCard, IsActiveMember, Gender, Tenure, Geography, NumOfProducts:

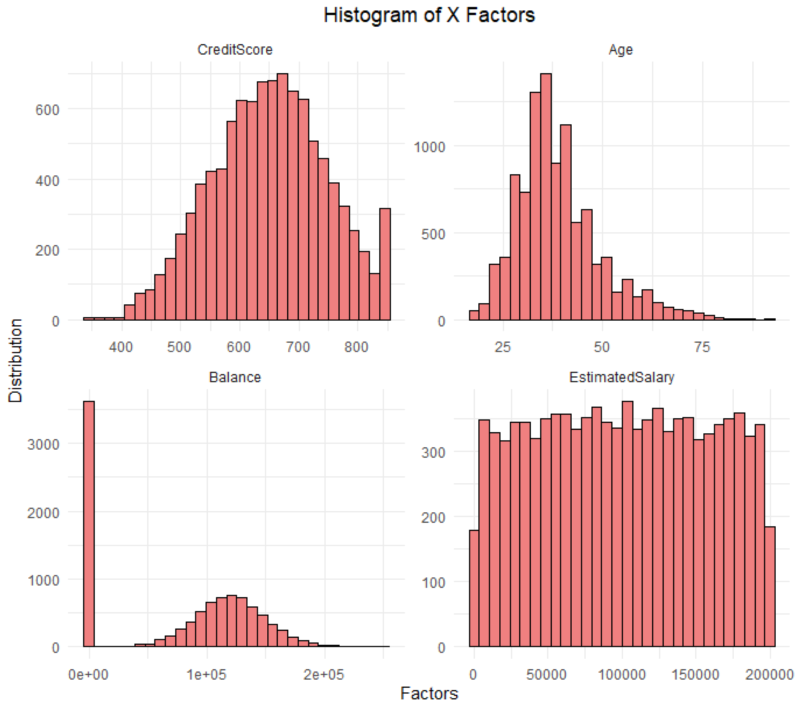


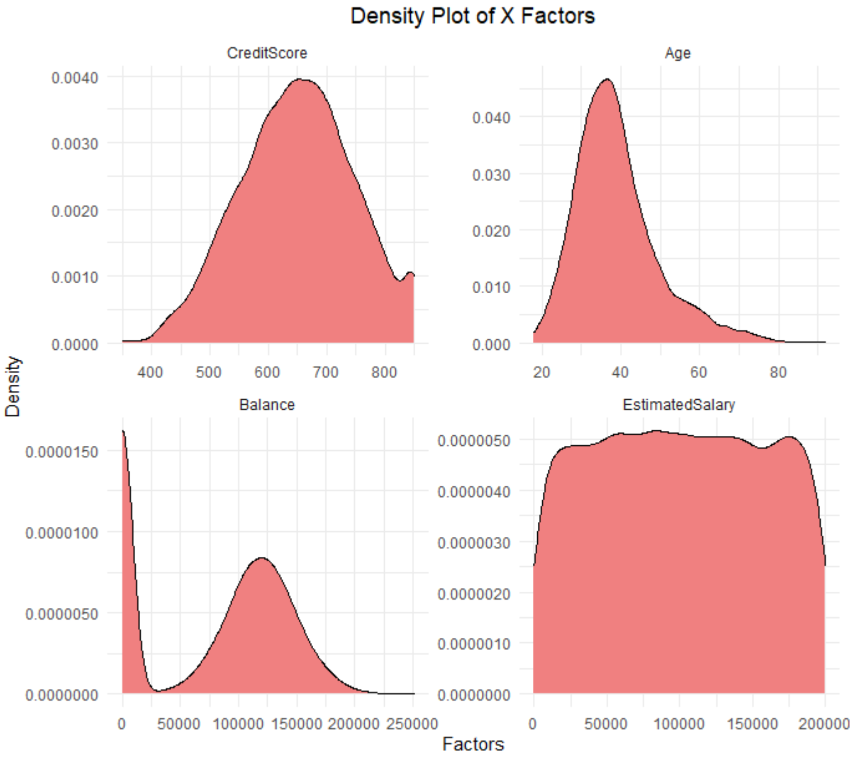
Promomethod, Subscriptionchannel, Language_provided, Language_preferred, Variant:



Univariate: Numerical

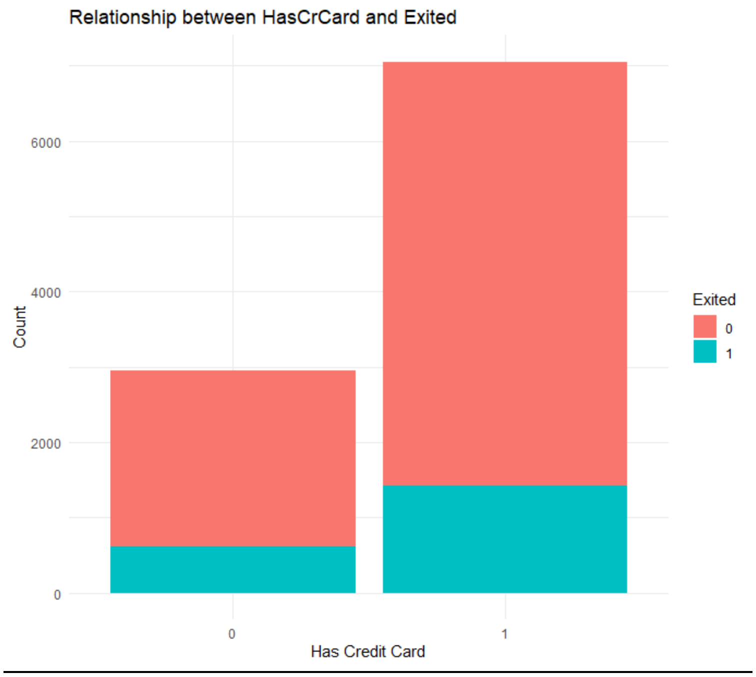
Age, CreditScore, Balance, EstimatedSalary:



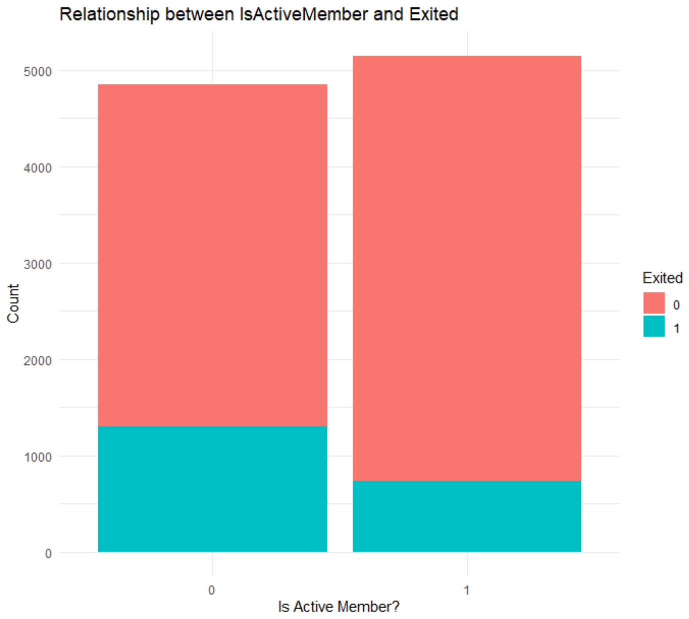


<u>Bivariate: Categorical vs Categorical</u>

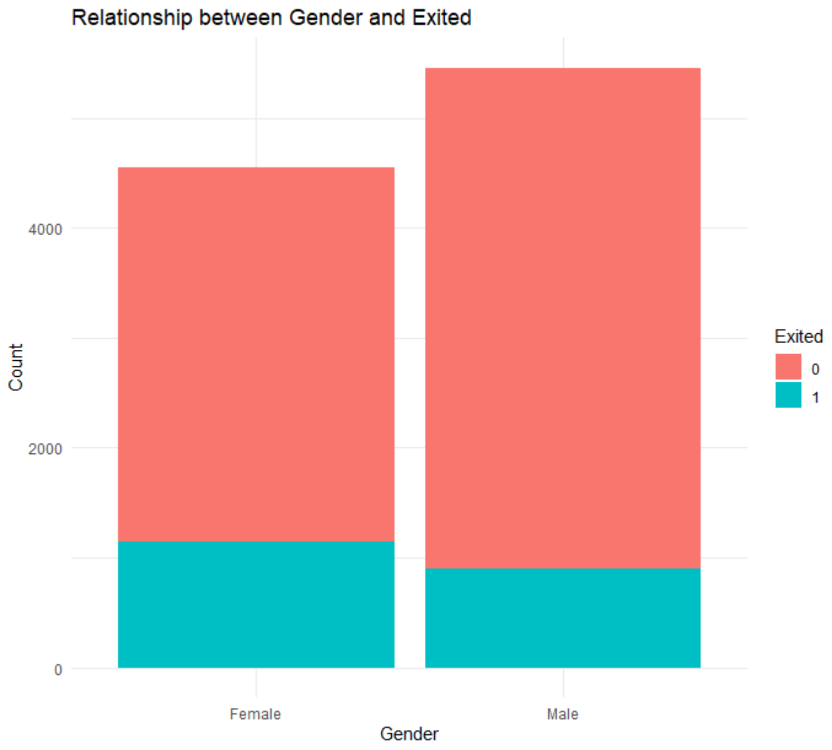
HasCrCard vs Exited:



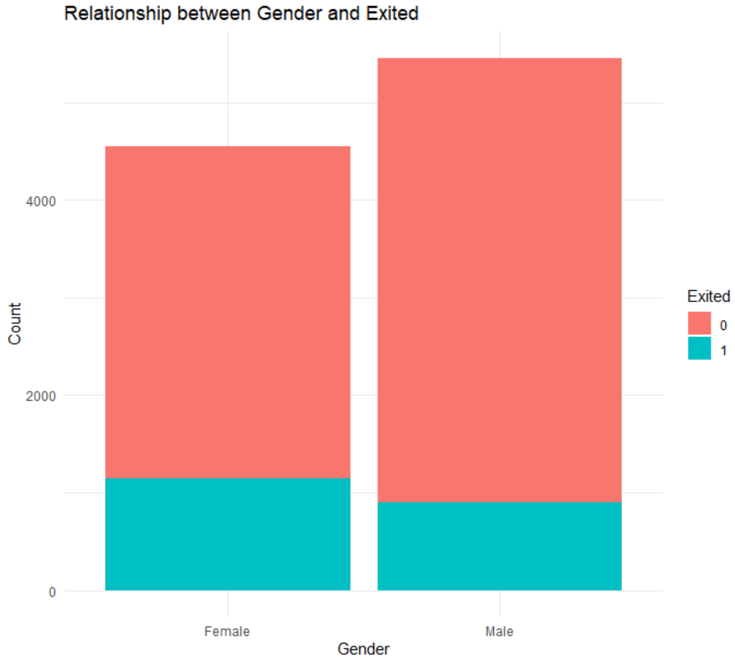
IsActiveMember vs Exited:



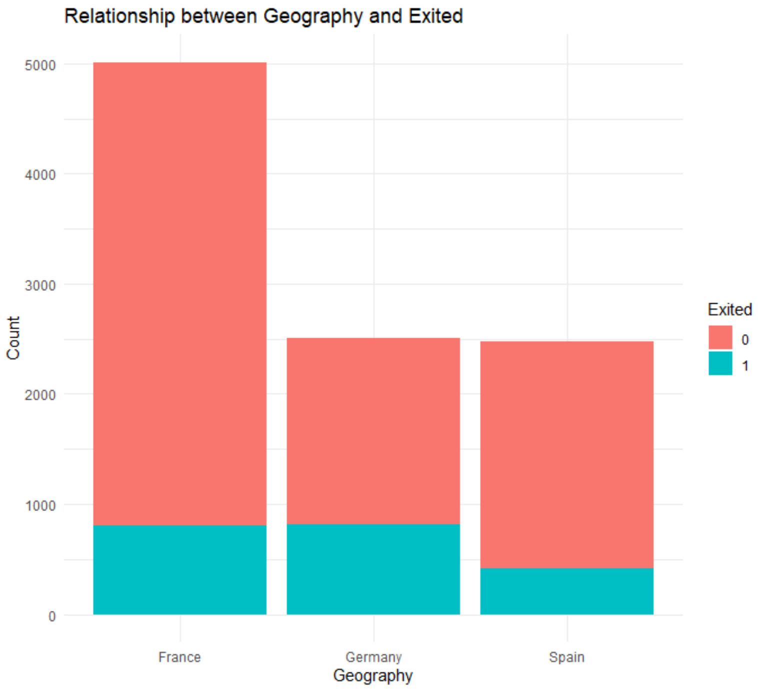
Gender vs Exited:



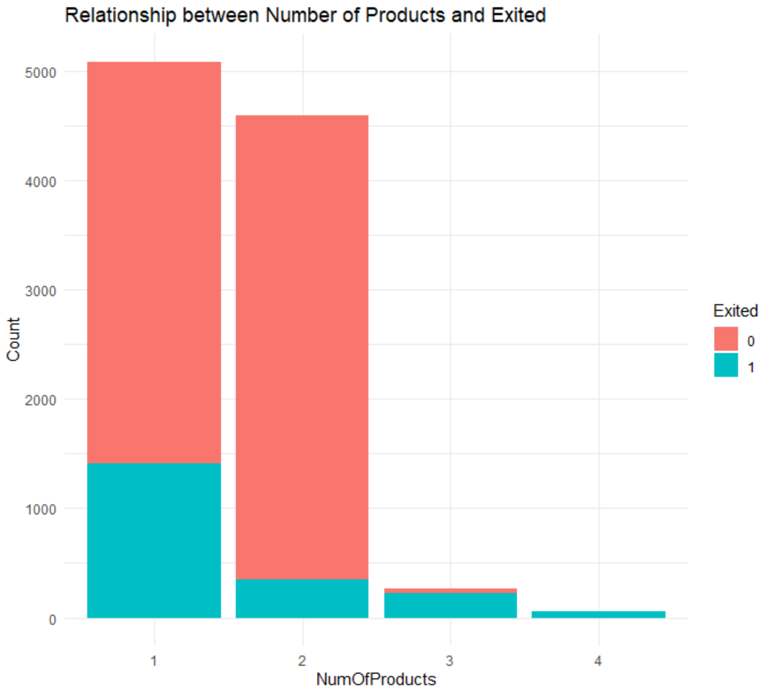
Tenure vs Exited:



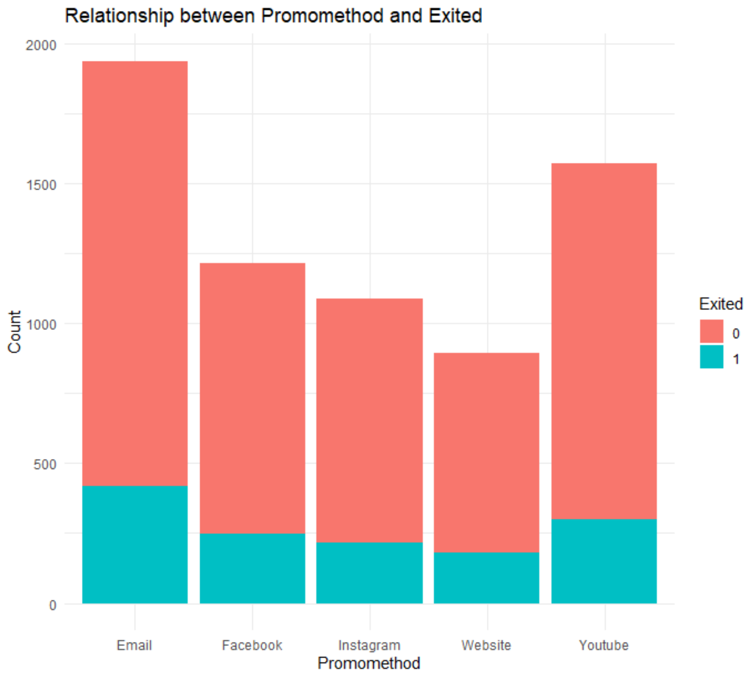
Geography vs Exited:



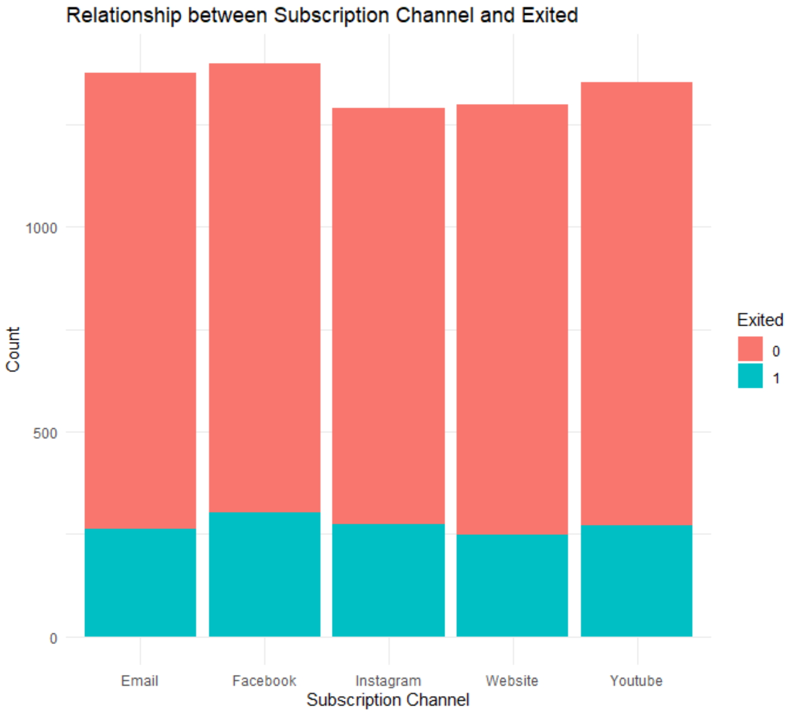
NumOfProducts vs Exited:



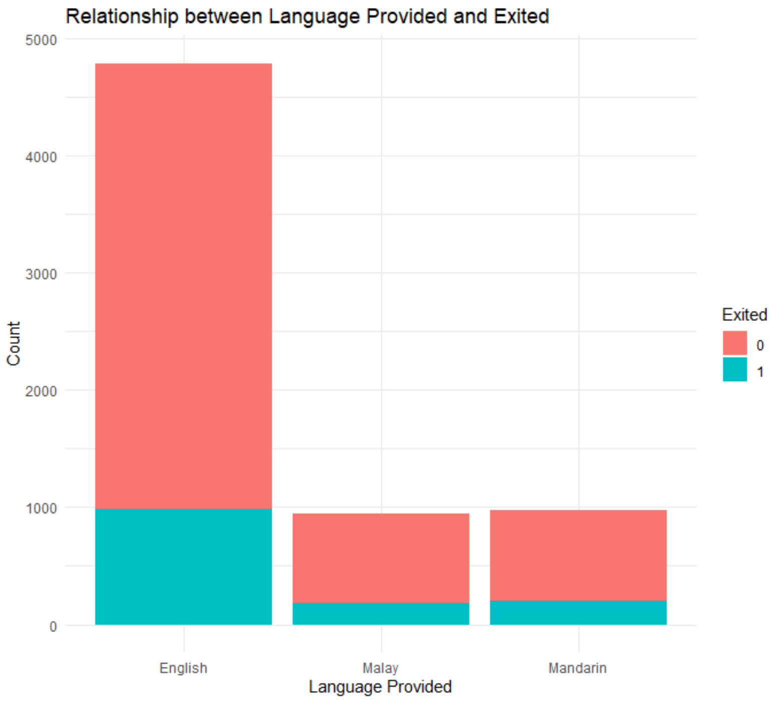
Promomethod vs Exited:



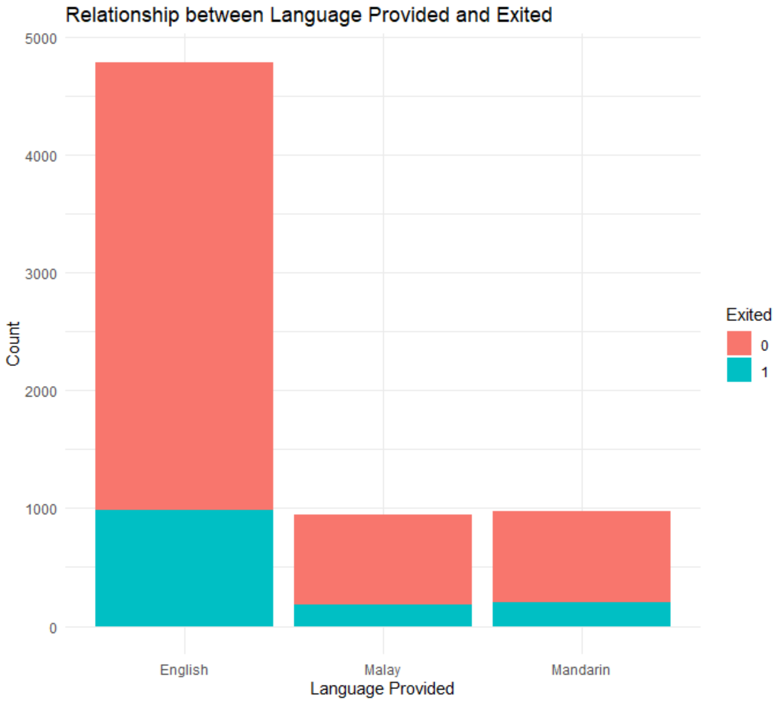
Subscriptionchannel vs Exited:



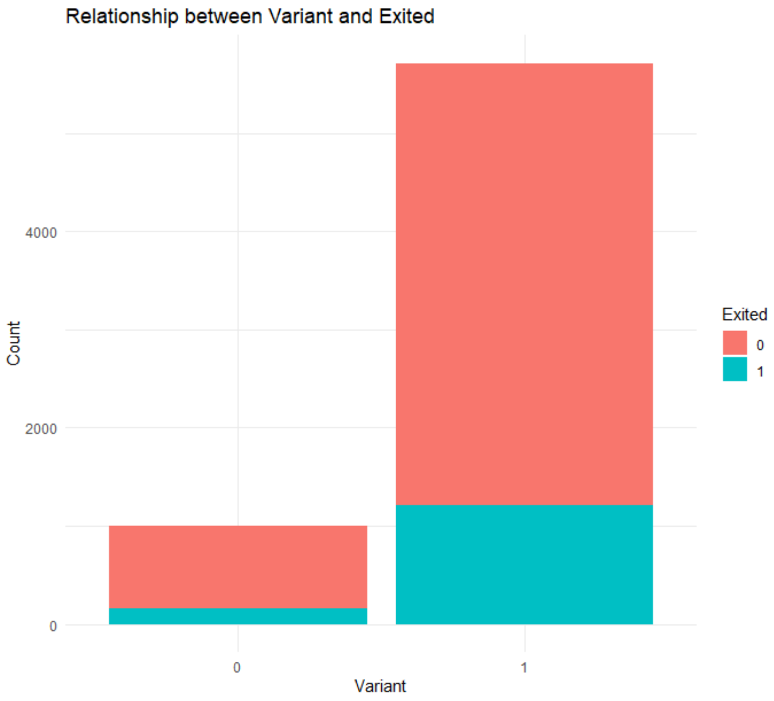
Language_provided vs Exited:



Language_preferred vs Exited:

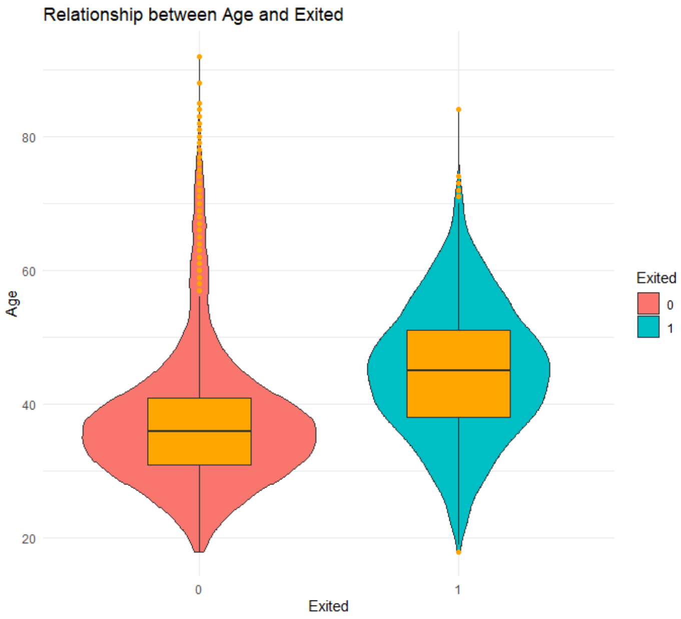


Variant vs Exited:

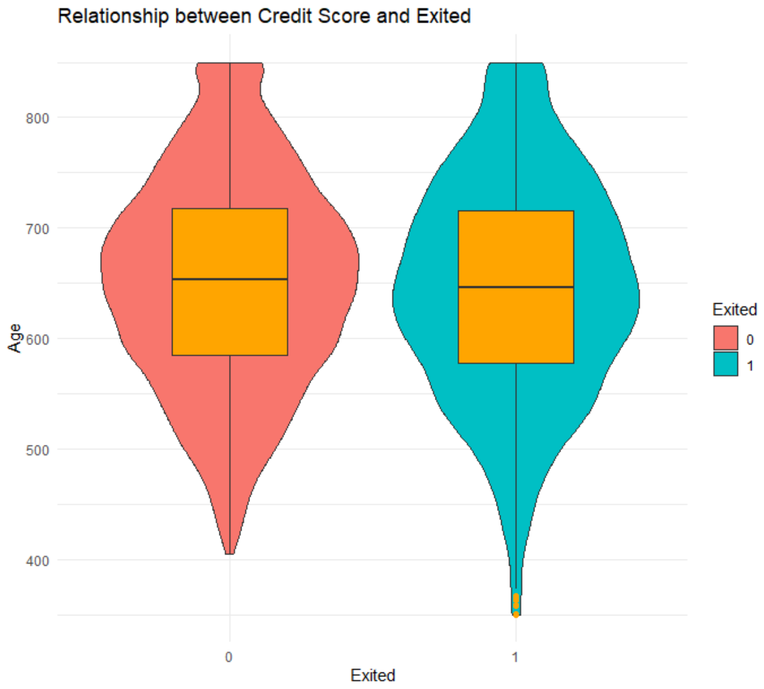


Bivariate: Numerical vs Categorical

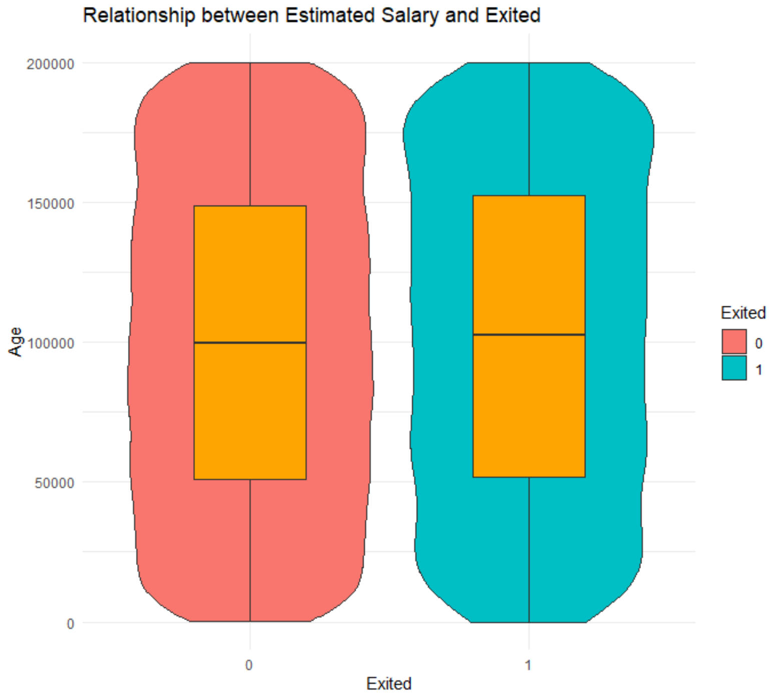
Age vs Exited:



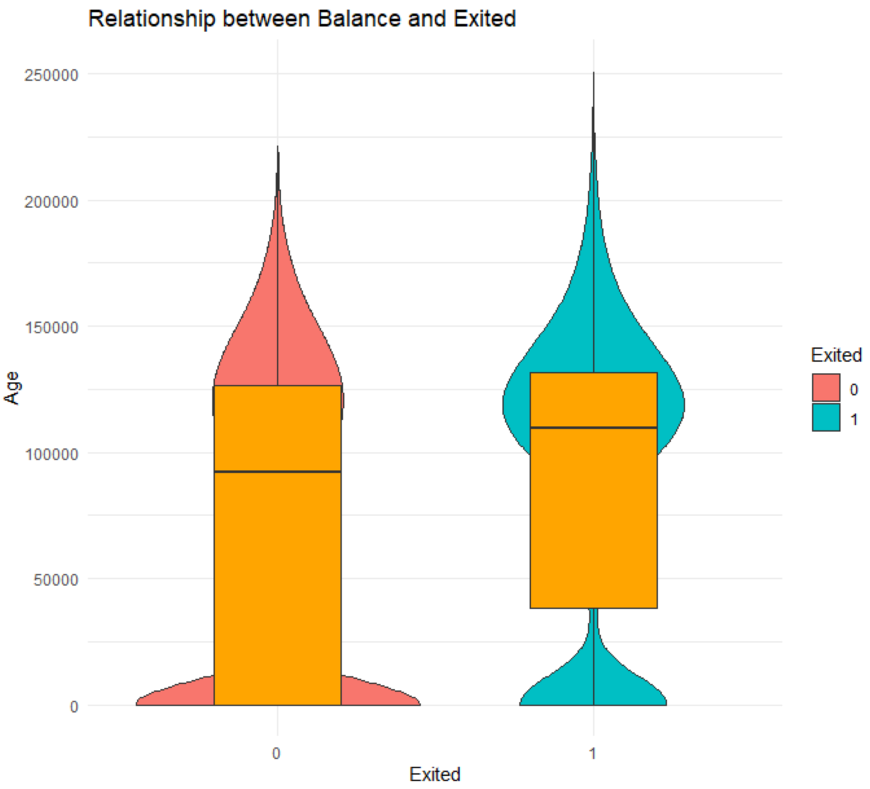
CreditScore vs Exited:



Estimated Salary vs Exited:

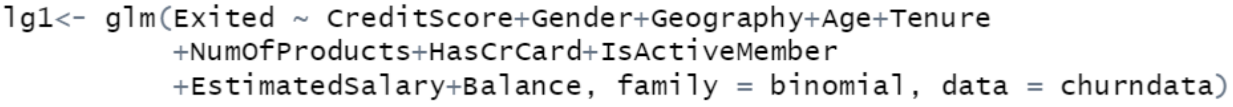


Balance vs Exited:



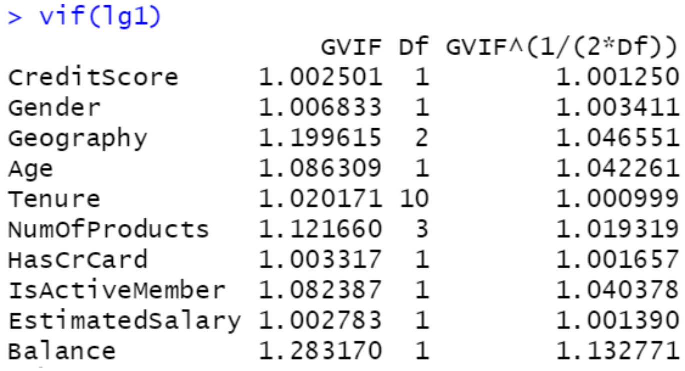
Appendix C - Logistic Regression

For the first model which determines the likelihood of a customer to stay, we built a logistic regression model with the glm() function with *'Exited'* as the output variable. All the other variables except for 'CustomerId' are included in the initial model as explanatory variables.



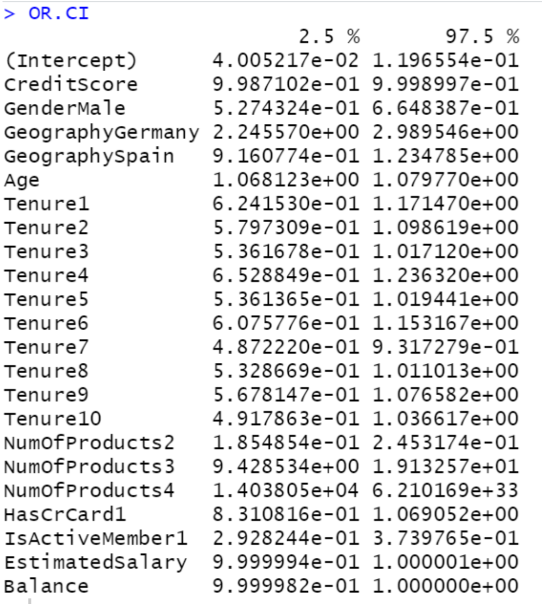
```
lg1<- glm(Exited ~ CreditScore+Gender+Geography+Age+Tenure
          +NumOfProducts+HasCrCard+IsActiveMember
          +EstimatedSalary+Balance, family = binomial, data = churndata)
```

To check for multicollinearity between the explanatory variables, we examined and found that the adjusted GVIF values are all below 2, which means that multicollinearity does not exist between the variables.

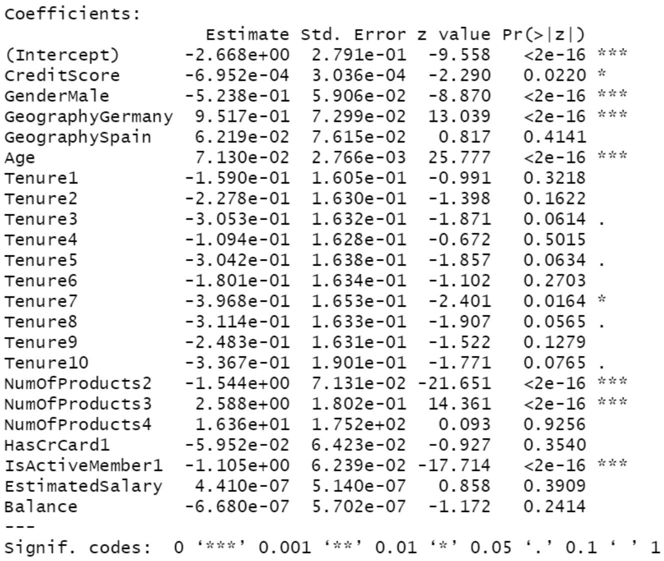


```
> vif(lg1)
                    GVIF Df GVIF^(1/(2*Df))
CreditScore     1.002501  1        1.001250
Gender          1.006833  1        1.003411
Geography       1.199615  2        1.046551
Age             1.086309  1        1.042261
Tenure          1.020171 10        1.000999
NumOfProducts   1.121660  3        1.019319
HasCrCard       1.003317  1        1.001657
IsActiveMember  1.082387  1        1.040378
EstimatedSalary 1.002783  1        1.001390
Balance         1.283170  1        1.132771
```

To sieve out the statistically significant variables, we analysed the p-values and the odds ratio confidence intervals.



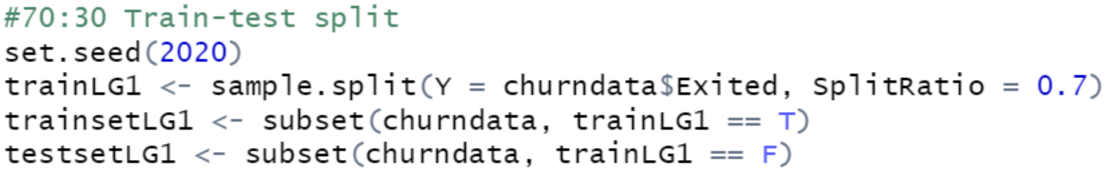
```
> OR.CI
                        2.5 %       97.5 %
(Intercept)     4.005217e-02 1.196554e-01
CreditScore     9.987102e-01 9.998997e-01
GenderMale      5.274324e-01 6.648387e-01
GeographyGermany 2.245570e+00 2.989546e+00
GeographySpain  9.160774e-01 1.234785e+00
Age             1.068123e+00 1.079770e+00
Tenure1         6.241530e-01 1.171470e+00
Tenure2         5.797309e-01 1.098619e+00
Tenure3         5.361678e-01 1.017120e+00
Tenure4         6.528849e-01 1.236320e+00
Tenure5         5.361365e-01 1.019441e+00
Tenure6         6.075776e-01 1.153167e+00
Tenure7         4.872220e-01 9.317279e-01
Tenure8         5.328669e-01 1.011013e+00
Tenure9         5.678147e-01 1.076582e+00
Tenure10        4.917863e-01 1.036617e+00
NumOfProducts2  1.854854e-01 2.453174e-01
NumOfProducts3  9.428534e+00 1.913257e+01
NumOfProducts4  1.403805e+04 6.210169e+33
HasCrCard1      8.310816e-01 1.069052e+00
IsActiveMember1 2.928244e-01 3.739765e-01
EstimatedSalary 9.999994e-01 1.000001e+00
Balance         9.999982e-01 1.000000e+00
```



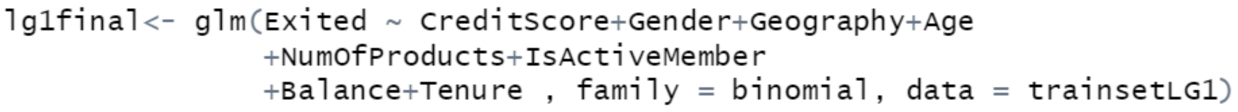
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -2.668e+00  2.791e-01  -9.558   <2e-16 ***
CreditScore     -6.952e-04  3.036e-04  -2.290   0.0220 *  
GenderMale      -5.238e-01  5.906e-02  -8.870   <2e-16 ***
GeographyGermany  9.517e-01  7.299e-02  13.039   <2e-16 ***
GeographySpain   6.219e-02  7.615e-02   0.817   0.4141    
Age              7.130e-02  2.766e-03  25.777   <2e-16 ***
Tenure1         -1.590e-01  1.605e-01  -0.991   0.3218    
Tenure2         -2.278e-01  1.630e-01  -1.398   0.1622    
Tenure3         -3.053e-01  1.632e-01  -1.871   0.0614 .  
Tenure4         -1.094e-01  1.628e-01  -0.672   0.5015    
Tenure5         -3.042e-01  1.638e-01  -1.857   0.0634 .  
Tenure6         -1.801e-01  1.634e-01  -1.102   0.2703    
Tenure7         -3.968e-01  1.653e-01  -2.401   0.0164 *  
Tenure8         -3.114e-01  1.633e-01  -1.907   0.0565 .  
Tenure9         -2.483e-01  1.631e-01  -1.522   0.1279    
Tenure10        -3.367e-01  1.901e-01  -1.771   0.0765 .  
NumOfProducts2  -1.544e+00  7.131e-02 -21.651   <2e-16 ***
NumOfProducts3   2.588e+00  1.802e-01  14.361   <2e-16 ***
NumOfProducts4   1.636e+01  1.752e+02   0.093   0.9256    
HasCrCard1      -5.952e-02  6.423e-02  -0.927   0.3540    
IsActiveMember1 -1.105e+00  6.239e-02 -17.714   <2e-16 ***
EstimatedSalary  4.410e-07  5.140e-07   0.858   0.3909    
Balance         -6.680e-07  5.702e-07  -1.172   0.2414    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For 5% alpha for p-values and confidence interval that excludes 1, only the variables *‘CreditScore’, ‘Gender’, ‘Geography’, ‘Age’, ‘NumOfProducts’, ‘IsActiveMember’, ‘Balance’, ‘Tenure’* are statistically significant. Hence, the final model consists of these 8 variables.

Using the *caTools* package, we performed a 70:30 train-test split on the dataset and only the statistically significant variables deduced from above are included in the new model, *lg1final*, to predict *‘Exited’*.

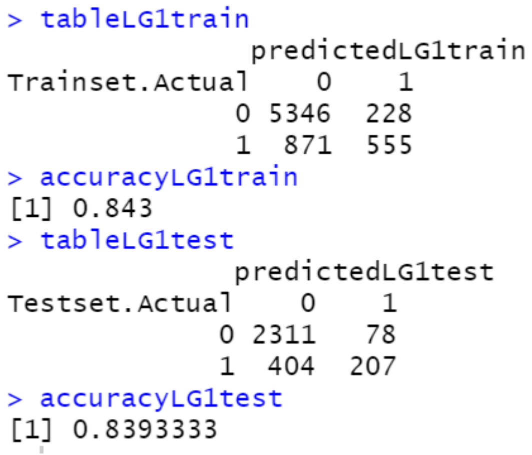


```
#70:30 Train-test split
set.seed(2020)
trainLG1 <- sample.split(Y = churndata$Exited, SplitRatio = 0.7)
trainsetLG1 <- subset(churndata, trainLG1 == T)
testsetLG1 <- subset(churndata, trainLG1 == F)
```



```
lg1final<- glm(Exited ~ CreditScore+Gender+Geography+Age
                +NumOfProducts+IsActiveMember
                +Balance+Tenure , family = binomial, data = trainsetLG1)
```

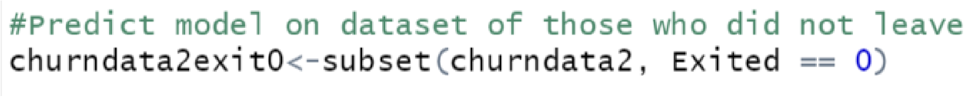
After deciding on a threshold of 0.5 for classification, we predicted the outcome on the train and test set respectively. The following are the confusion matrix and accuracy obtained from our analysis.



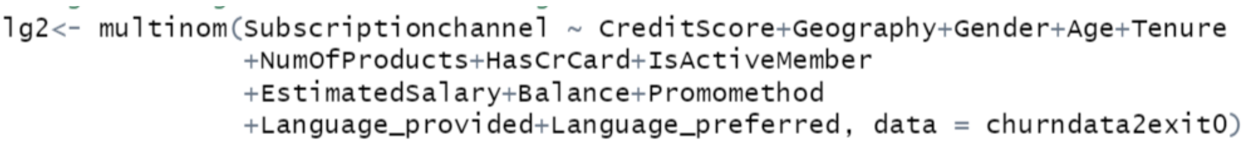
```
> tableLG1train
               predictedLG1train
Trainset.Actual    0    1
              0 5346  228
              1  871  555
> accuracyLG1train
[1] 0.843
> tableLG1test
              predictedLG1test
Testset.Actual    0    1
             0 2311   78
             1  404  207
> accuracyLG1test
[1] 0.8393333
```

For the second model which determines the new promotional method for the customers who are predicted to leave from the first model, we built a multinomial logistic regression model with the *multinom()* function with *'Subscriptionchannel'* as the output variable. Independent variables included are deduced from the variable importance of the first model.

A data table *churndata2exit0* is created to extract the customers whose *'Exited'*=0, which means that they stayed in the company. The model is then trained on this dataset with only customers who stayed with the company, to predict the *'Subscriptionchannel'*.



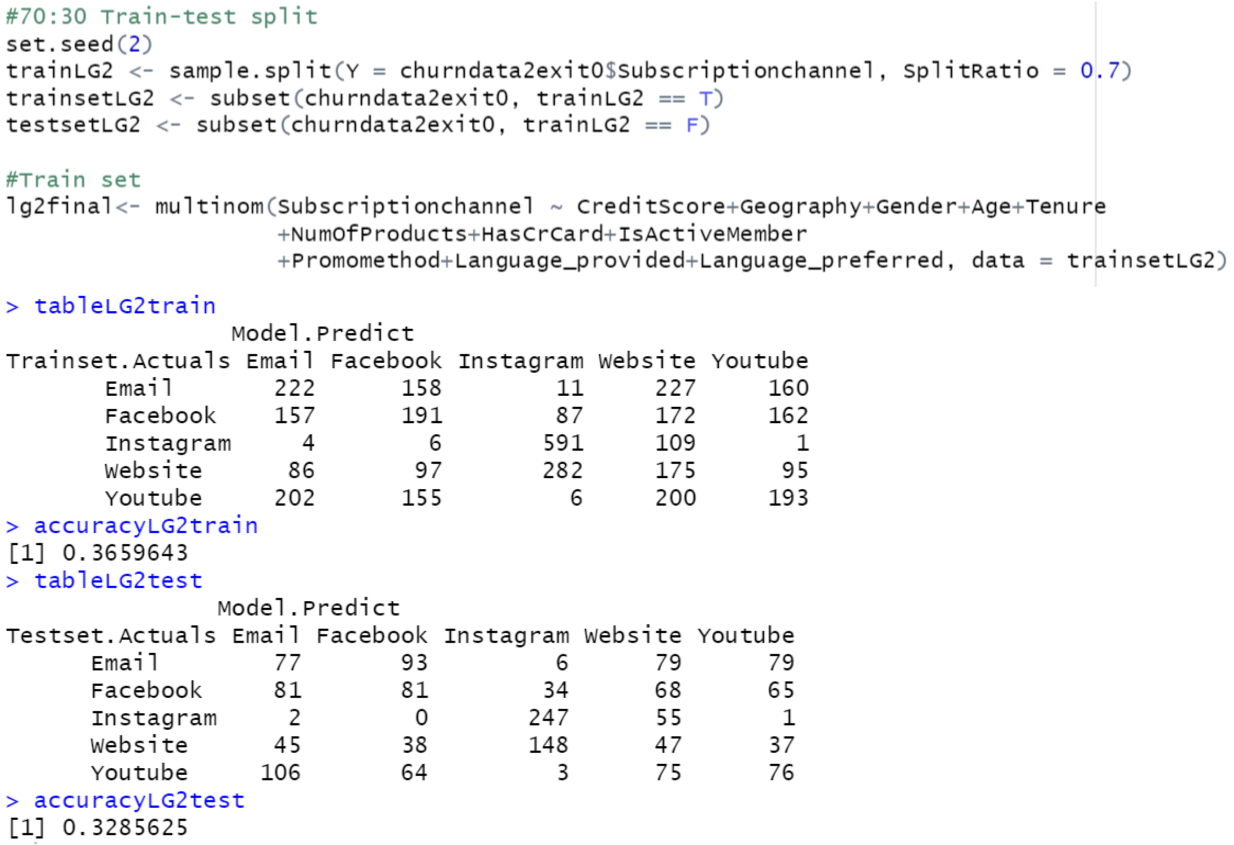
```
#Predict model on dataset of those who did not leave
churndata2exit0<-subset(churndata2, Exited == 0)
```



```
lg2<- multinom(Subscriptionchannel ~ CreditScore+Geography+Gender+Age+Tenure
               +NumOfProducts+HasCrCard+IsActiveMember
               +EstimatedSalary+Balance+Promomethod
               +Language_provided+Language_preferred, data = churndata2exit0)
```

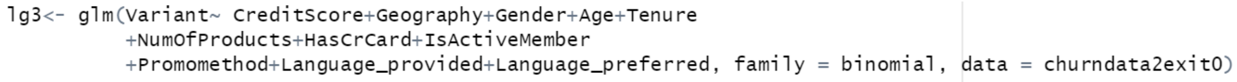
We conducted the z-test to analyse the p-values of the variables included. At 5% alpha, the variables *'CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'NumOfProducts', 'HasCrCard',* *‘IsActiveMember’, ‘Promomethod’, ‘Language_provided’, ‘Language_preferred’.* This is consistent with the result from analysing the odds ratio confidence interval

Keeping only the statistically significant variables in the model, we performed a 70:30 train-test split, similar to the one mentioned previously. The following are the confusion matrix and accuracy obtained.



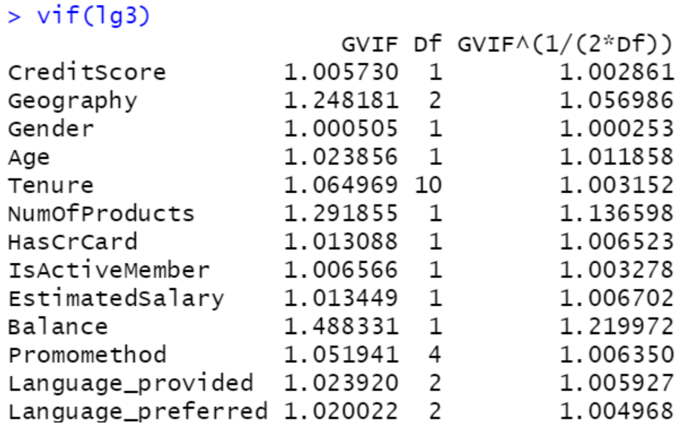
```
#70:30 Train-test split
set.seed(2)
trainLG2 <- sample.split(Y = churndata2exit0$Subscriptionchannel, SplitRatio = 0.7)
trainsetLG2 <- subset(churndata2exit0, trainLG2 == T)
testsetLG2 <- subset(churndata2exit0, trainLG2 == F)

#Train set
lg2final<- multinom(Subscriptionchannel ~ CreditScore+Geography+Gender+Age+Tenure
                    +NumOfProducts+HasCrCard+IsActiveMember
                    +Promomethod+Language_provided+Language_preferred, data = trainsetLG2)

> tableLG2train
                Model.Predict
Trainset.Actuals Email Facebook Instagram Website Youtube
       Email       222      158        11     227     160
       Facebook    157      191        87     172     162
       Instagram     4        6       591     109       1
       Website      86       97       282     175      95
       Youtube     202      155         6     200     193
> accuracyLG2train
[1] 0.3659643
> tableLG2test
               Model.Predict
Testset.Actuals Email Facebook Instagram Website Youtube
      Email        77       93         6      79      79
      Facebook     81       81        34      68      65
      Instagram     2        0       247      55       1
      Website      45       38       148      47      37
      Youtube     106       64         3      75      76
> accuracyLG2test
[1] 0.3285625
```

For the third model which predicts the need for personalisation on the new promotional methods predicted in the previous model, we built a logistic regression model with the glm() function with ‘Personalisation’ as the output variable. Input variables used in this model are the same as those used in the initial model for model 2.

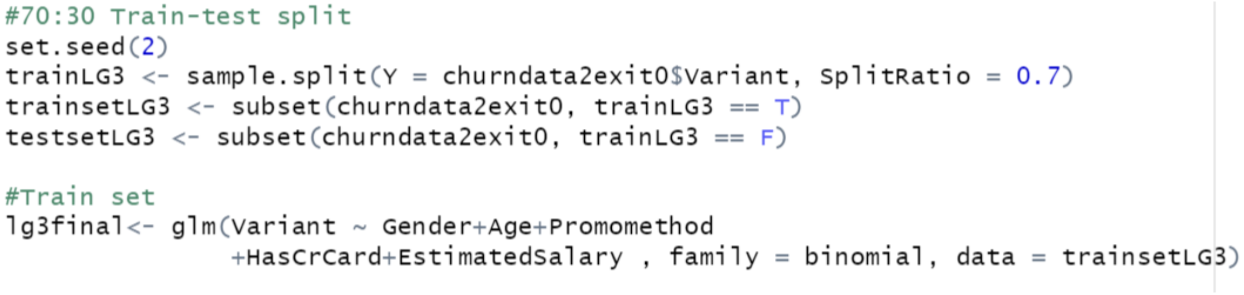


```
lg3<- glm(Variant~ CreditScore+Geography+Gender+Age+Tenure
          +NumOfProducts+HasCrCard+IsActiveMember
          +Promomethod+Language_provided+Language_preferred, family = binomial, data = churndata2exit0)
```

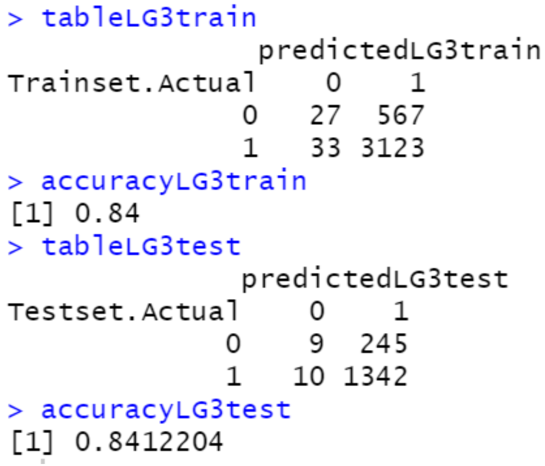
Following the same procedures in the first model, we found that there is no multicollinearity issues between the explanatory variables, we then analysed the statistical significance of the variables and performed a 70:30 train-test split on the new model. The following are the confusion matrix and accuracy obtained.



```
> vif(lg3)
                      GVIF Df GVIF^(1/(2*Df))
CreditScore       1.005730  1        1.002861
Geography         1.248181  2        1.056986
Gender            1.000505  1        1.000253
Age               1.023856  1        1.011858
Tenure            1.064969 10        1.003152
NumOfProducts     1.291855  1        1.136598
HasCrCard         1.013088  1        1.006523
IsActiveMember    1.006566  1        1.003278
EstimatedSalary   1.013449  1        1.006702
Balance           1.488331  1        1.219972
Promomethod       1.051941  4        1.006350
Language_provided 1.023920  2        1.005927
Language_preferred 1.020022  2        1.004968
```



```
#70:30 Train-test split
set.seed(2)
trainLG3 <- sample.split(Y = churndata2exit0$Variant, SplitRatio = 0.7)
trainsetLG3 <- subset(churndata2exit0, trainLG3 == T)
testsetLG3 <- subset(churndata2exit0, trainLG3 == F)

#Train set
lg3final<- glm(Variant ~ Gender+Age+Promomethod
               +HasCrCard+EstimatedSalary , family = binomial, data = trainsetLG3)
```

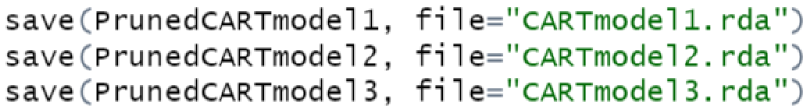


```
> tableLG3train
                predictedLG3train
Trainset.Actual    0    1
              0   27  567
              1   33 3123
> accuracyLG3train
[1] 0.84
> tableLG3test
               predictedLG3test
Testset.Actual    0    1
             0    9  245
             1   10 1342
> accuracyLG3test
[1] 0.8412204
```

Appendix D - CART Model

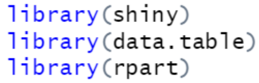
Appendix E - Interface

For the ease of accessing the predictive models each time, each of our CART models is saved as a .rda file.

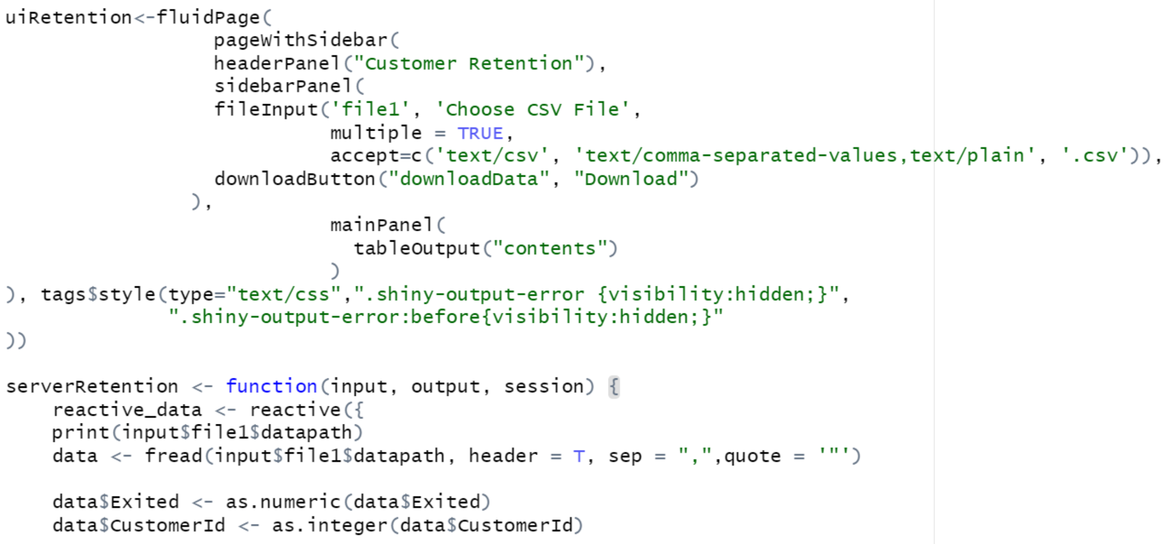


```
save(PrunedCARTmodel1, file="CARTmodel1.rda")
save(PrunedCARTmodel2, file="CARTmodel2.rda")
save(PrunedCARTmodel3, file="CARTmodel3.rda")
```

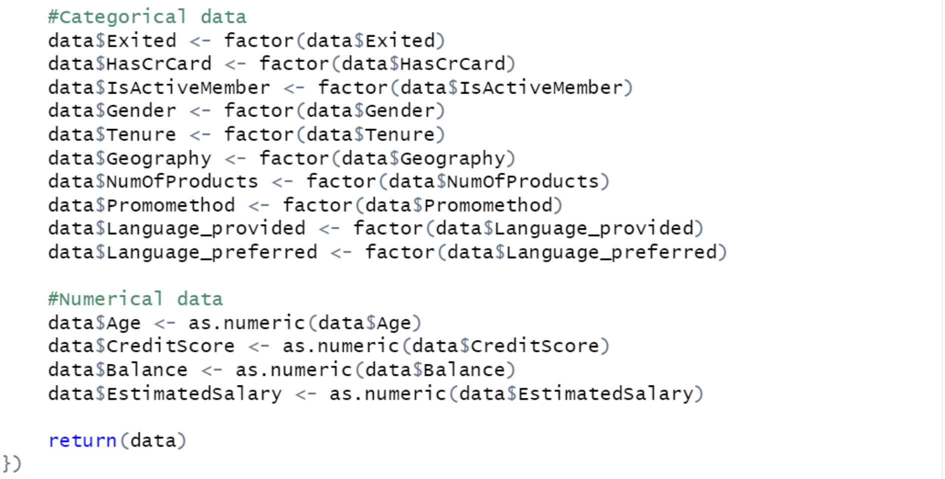
The interfaces are implemented with R Shiny with *library(shiny)*. Full code of the implementation are as follows:



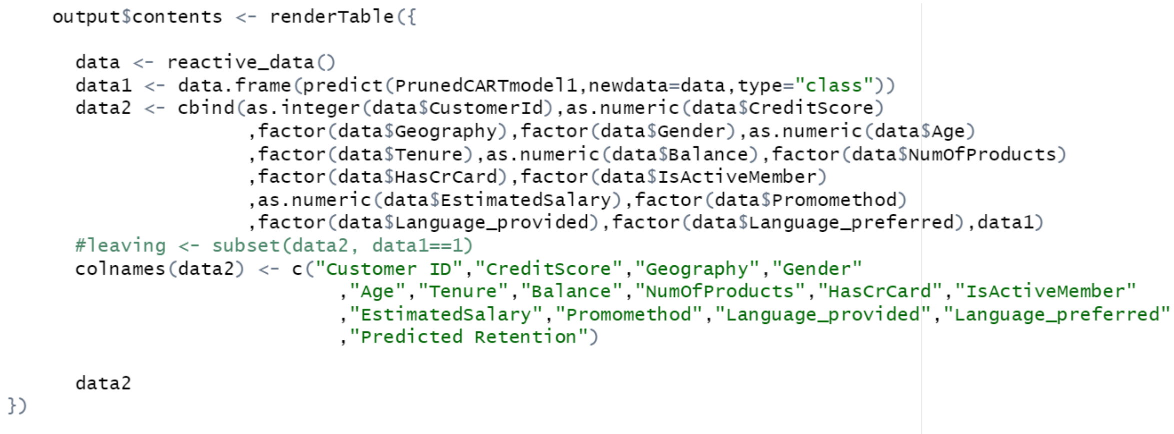
```
library(shiny)
library(data.table)
library(rpart)
```



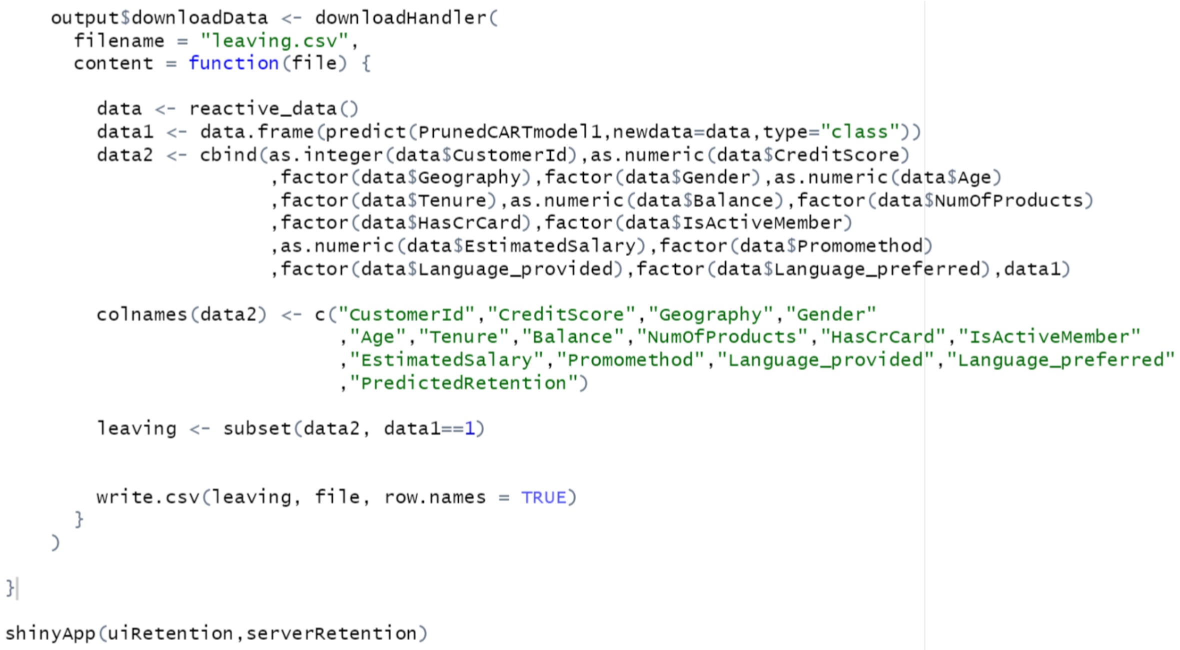
```
uiRetention<-fluidPage(
                  pageWithSidebar(
                  headerPanel("Customer Retention"),
                  sidebarPanel(
                  fileInput('file1', 'Choose CSV File',
                            multiple = TRUE,
                            accept=c('text/csv', 'text/comma-separated-values,text/plain', '.csv')),
                  downloadButton("downloadData", "Download")
                ),
                            mainPanel(
                              tableOutput("contents")
                            )
), tags$style(type="text/css",".shiny-output-error {visibility:hidden;}",
              ".shiny-output-error:before{visibility:hidden;}"
))

serverRetention <- function(input, output, session) {
    reactive_data <- reactive({
    print(input$file1$datapath)
    data <- fread(input$file1$datapath, header = T, sep = ",",quote = '"')

    data$Exited <- as.numeric(data$Exited)
    data$CustomerId <- as.integer(data$CustomerId)
```



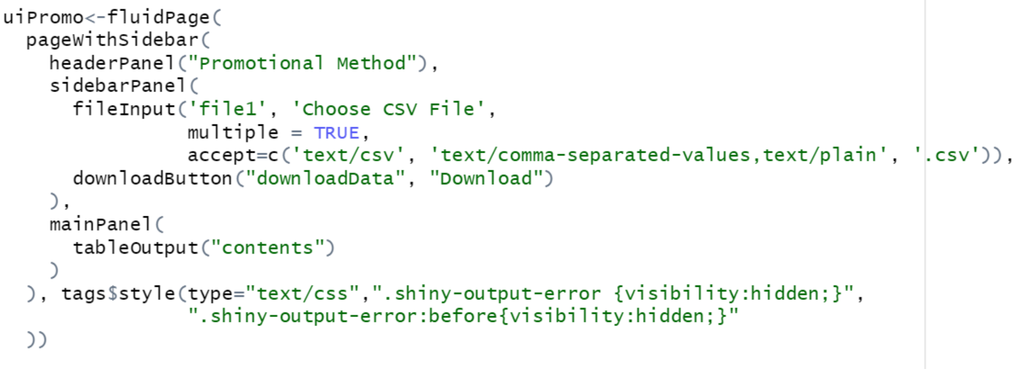
```
    #Categorical data
    data$Exited <- factor(data$Exited)
    data$HasCrCard <- factor(data$HasCrCard)
    data$IsActiveMember <- factor(data$IsActiveMember)
    data$Gender <- factor(data$Gender)
    data$Tenure <- factor(data$Tenure)
    data$Geography <- factor(data$Geography)
    data$NumOfProducts <- factor(data$NumOfProducts)
    data$Promomethod <- factor(data$Promomethod)
    data$Language_provided <- factor(data$Language_provided)
    data$Language_preferred <- factor(data$Language_preferred)

    #Numerical data
    data$Age <- as.numeric(data$Age)
    data$CreditScore <- as.numeric(data$CreditScore)
    data$Balance <- as.numeric(data$Balance)
    data$EstimatedSalary <- as.numeric(data$EstimatedSalary)

    return(data)
})
```



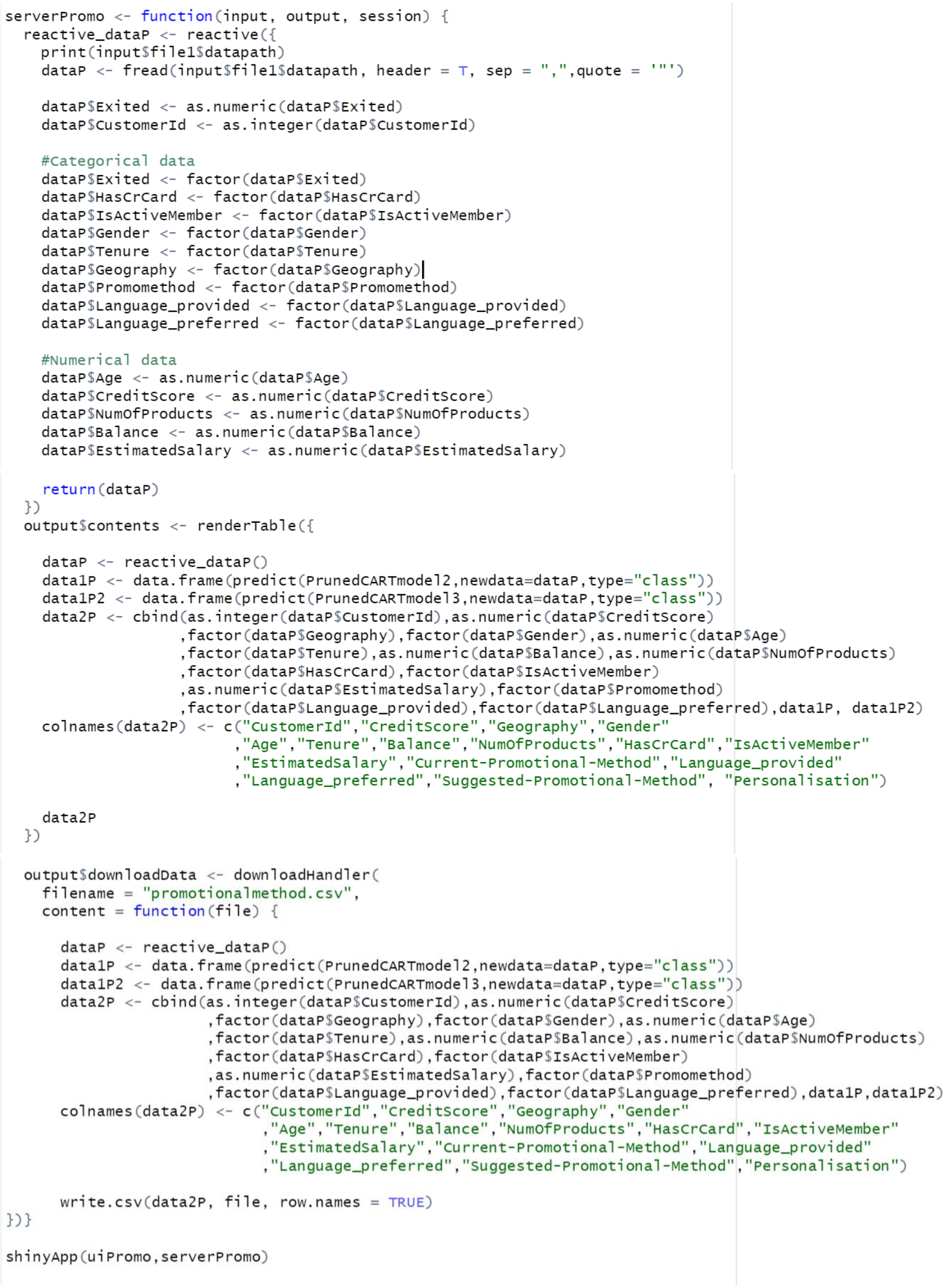
```r
    output$contents <- renderTable({

      data <- reactive_data()
      data1 <- data.frame(predict(PrunedCARTmodel1,newdata=data,type="class"))
      data2 <- cbind(as.integer(data$CustomerId),as.numeric(data$CreditScore)
                     ,factor(data$Geography),factor(data$Gender),as.numeric(data$Age)
                     ,factor(data$Tenure),as.numeric(data$Balance),factor(data$NumOfProducts)
                     ,factor(data$HasCrCard),factor(data$IsActiveMember)
                     ,as.numeric(data$EstimatedSalary),factor(data$Promomethod)
                     ,factor(data$Language_provided),factor(data$Language_preferred),data1)
      #leaving <- subset(data2, data1==1)
      colnames(data2) <- c("Customer ID","CreditScore","Geography","Gender"
                             ,"Age","Tenure","Balance","NumOfProducts","HasCrCard","IsActiveMember"
                             ,"EstimatedSalary","Promomethod","Language_provided","Language_preferred"
                             ,"Predicted Retention")

      data2
})
```



```r
    output$downloadData <- downloadHandler(
      filename = "leaving.csv",
      content = function(file) {

        data <- reactive_data()
        data1 <- data.frame(predict(PrunedCARTmodel1,newdata=data,type="class"))
        data2 <- cbind(as.integer(data$CustomerId),as.numeric(data$CreditScore)
                       ,factor(data$Geography),factor(data$Gender),as.numeric(data$Age)
                       ,factor(data$Tenure),as.numeric(data$Balance),factor(data$NumOfProducts)
                       ,factor(data$HasCrCard),factor(data$IsActiveMember)
                       ,as.numeric(data$EstimatedSalary),factor(data$Promomethod)
                       ,factor(data$Language_provided),factor(data$Language_preferred),data1)

        colnames(data2) <- c("CustomerId","CreditScore","Geography","Gender"
                             ,"Age","Tenure","Balance","NumOfProducts","HasCrCard","IsActiveMember"
                             ,"EstimatedSalary","Promomethod","Language_provided","Language_preferred"
                             ,"PredictedRetention")

        leaving <- subset(data2, data1==1)


        write.csv(leaving, file, row.names = TRUE)
      }
    )

}

shinyApp(uiRetention,serverRetention)
```



```r
uiPromo<-fluidPage(
  pageWithSidebar(
    headerPanel("Promotional Method"),
    sidebarPanel(
      fileInput('file1', 'Choose CSV File',
                multiple = TRUE,
                accept=c('text/csv', 'text/comma-separated-values,text/plain', '.csv')),
      downloadButton("downloadData", "Download")
    ),
    mainPanel(
      tableOutput("contents")
    )
  ), tags$style(type="text/css",".shiny-output-error {visibility:hidden;}",
                ".shiny-output-error:before{visibility:hidden;}"
  ))
```



```
serverPromo <- function(input, output, session) {
  reactive_dataP <- reactive({
    print(input$file1$datapath)
    dataP <- fread(input$file1$datapath, header = T, sep = ",",quote = '"')

    dataP$Exited <- as.numeric(dataP$Exited)
    dataP$CustomerId <- as.integer(dataP$CustomerId)

    #Categorical data
    dataP$Exited <- factor(dataP$Exited)
    dataP$HasCrCard <- factor(dataP$HasCrCard)
    dataP$IsActiveMember <- factor(dataP$IsActiveMember)
    dataP$Gender <- factor(dataP$Gender)
    dataP$Tenure <- factor(dataP$Tenure)
    dataP$Geography <- factor(dataP$Geography)
    dataP$Promomethod <- factor(dataP$Promomethod)
    dataP$Language_provided <- factor(dataP$Language_provided)
    dataP$Language_preferred <- factor(dataP$Language_preferred)

    #Numerical data
    dataP$Age <- as.numeric(dataP$Age)
    dataP$CreditScore <- as.numeric(dataP$CreditScore)
    dataP$NumOfProducts <- as.numeric(dataP$NumOfProducts)
    dataP$Balance <- as.numeric(dataP$Balance)
    dataP$EstimatedSalary <- as.numeric(dataP$EstimatedSalary)

    return(dataP)
  })
  output$contents <- renderTable({

    dataP <- reactive_dataP()
    data1P <- data.frame(predict(PrunedCARTmodel2,newdata=dataP,type="class"))
    data1P2 <- data.frame(predict(PrunedCARTmodel3,newdata=dataP,type="class"))
    data2P <- cbind(as.integer(dataP$CustomerId),as.numeric(dataP$CreditScore)
                   ,factor(dataP$Geography),factor(dataP$Gender),as.numeric(dataP$Age)
                   ,factor(dataP$Tenure),as.numeric(dataP$Balance),as.numeric(dataP$NumOfProducts)
                   ,factor(dataP$HasCrCard),factor(dataP$IsActiveMember)
                   ,as.numeric(dataP$EstimatedSalary),factor(dataP$Promomethod)
                   ,factor(dataP$Language_provided),factor(dataP$Language_preferred),data1P, data1P2)
    colnames(data2P) <- c("CustomerId","CreditScore","Geography","Gender"
                         ,"Age","Tenure","Balance","NumOfProducts","HasCrCard","IsActiveMember"
                         ,"EstimatedSalary","Current-Promotional-Method","Language_provided"
                         ,"Language_preferred","Suggested-Promotional-Method", "Personalisation")

    data2P
  })

  output$downloadData <- downloadHandler(
    filename = "promotionalmethod.csv",
    content = function(file) {

      dataP <- reactive_dataP()
      data1P <- data.frame(predict(PrunedCARTmodel2,newdata=dataP,type="class"))
      data1P2 <- data.frame(predict(PrunedCARTmodel3,newdata=dataP,type="class"))
      data2P <- cbind(as.integer(dataP$CustomerId),as.numeric(dataP$CreditScore)
                      ,factor(dataP$Geography),factor(dataP$Gender),as.numeric(dataP$Age)
                      ,factor(dataP$Tenure),as.numeric(dataP$Balance),as.numeric(dataP$NumOfProducts)
                      ,factor(dataP$HasCrCard),factor(dataP$IsActiveMember)
                      ,as.numeric(dataP$EstimatedSalary),factor(dataP$Promomethod)
                      ,factor(dataP$Language_provided),factor(dataP$Language_preferred),data1P,data1P2)
      colnames(data2P) <- c("CustomerId","CreditScore","Geography","Gender"
                            ,"Age","Tenure","Balance","NumOfProducts","HasCrCard","IsActiveMember"
                            ,"EstimatedSalary","Current-Promotional-Method","Language_provided"
                            ,"Language_preferred","Suggested-Promotional-Method","Personalisation")

      write.csv(data2P, file, row.names = TRUE)
})}

shinyApp(uiPromo,serverPromo)
```

REFERENCES

Roesti, Y., & Wettstein, S. (2015). The Singapore Asset Management Industry Building a Strong Foundation for Future Growth. Retrieved from https://www.synpulse.com/_Resources/Persistent/499ccf97ee46e3f2ef895e72e00c37a9eb46a1ba/White-Paper_Singapore-Asset-Management-Industry_EN.pdf

Grant, J. (n.d.). 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

McEachern, A. (2020, July 09). Customer Retention 101: Grow Your Business by Selling More to Current Customers. Retrieved from https://www.shopify.com/blog/customer-retention-strategies

Saleh, K. (2019, November 11). Customer Acquisition Vs.Retention Costs – Statistics And Trends. Retrieved from https://www.invespcro.com/blog/customer-acquisition-retention/

Landis, T. (2020, May 6). Customer Retention Marketing vs. Customer Acquisition Marketing. Retrieved from https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/

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

Dhami, S. (2020, October 06). Customer Loyalty: Innovative Customer Retention Strategies. Retrieved from https://www.ringcentral.co.uk/blog/customer-loyalty-innovative-customer-retention-strategies/

Patel, N. (2019, March 21). What is Customer Retention, Importance, Examples & Techniques. Retrieved from https://www.crazyegg.com/blog/customer-retention/

Barker, S. (2020, March 02). 9 Barriers in Customer Retention and How to Break Them. Retrieved from https://thenextscoop.com/customer-retention-barriers/

KPMG. (2017, June). Spotlight on the Asset Management Industry. Retrieved from https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/07/spotlight-on-the-asset-management-industry.pdf

CaseyQuirk. (2019). How technology will redefine relationships with asset management clients. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/how-technology-will-redefine-asset-management-relationships.pdf

Miller, G. (2019). 21 Surprising Customer Retention Statistics for 2018. Retrieved from https://www.annexcloud.com/blog/21-surprising-customer-retention-statistics-2018/

Lorenzon, A., & Pilotti, L. (2008, May 5). Consumer receptiveness in the development of a holistic communication strategy: Trust, advocacy and brand ecology. Innovative Marketing. Retrieved from https://businessperspectives.org/images/pdf/applications/publishing/templates/article/assets/2084/im_en_2008_1_Lorenzon.pdf

PwC. (n.d.) Asset Management 2020 - A Brave New World. Retrieved from https://www.pwc.com/gx/en/asset-management/publications/pdfs/pwc-asset-management-2020-a-brave-new-world-final.pdf

MonkeyLearn. (n.d.). Sentiment Analysis: A Definitive Guide. Retrieved from https://monkeylearn.com/sentiment-analysis/

Marx, A. (2020, July 21). Customer retention analytics: 5 strategies to reduce churn. Retrieved from https://getthematic.com/insights/5-ways-data-and-text-analytics-improve-customer-retention/

MacDonald, S. (2020, October 12). 7 Award Winning Customer Service Email Templates. Retrieved from https://www.superoffice.com/blog/customer-service-email-templates-for-business/

Hayes, B. (2017, October 29). Archive: Analytics. Retrieved from http://businessoverbroadway.com/category/analytics-2/page/3/

Countants. (2020, January 08). How Does Customer Feedback Sentiment Analysis Work in Search Marketing? Retrieved from https://medium.com/datadriveninvestor/how-does-customer-feedback-sentiment-analysis-work-in-search-

Berné, Carmen & Mugica, Jose & Yagüe, María. (2001). The effect of variety-seeking on customer retention in services. Journal of Retailing and Consumer Services. Retrieved from https://www.researchgate.net/publication/222773169_The_effect_of_variety-seeking_on_customer_retention_in_services

King, S. (2020, May 19). 5 customer retention challenges facing every business marketer. Retrieved from https://www.mention-me.com/blog/customer-retention-challenges-facing-every-business-marketer

MJ. Barone, R., RF. Baumeister, K., RF. Baumeister, K., Buss, D., DM. Buss, D., P. Chandon, V., . . . CKB. Yim, D. (2012, January 21). Make me special: Gender differences in consumers' responses to loyalty programs. Retrieved from https://link.springer.com/article/10.1007/s11002-011-9160-3

Sahm, CR. (n.d.) How Much Does Risk Tolerance Change? Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276321/

Gustafsson, C & Omark, L. (2015). Financial literacy's effect on financial risk tolerance. Retrieved from https://www.diva-portal.org/smash/get/diva2:826787/FULLTEXT01.pdf

2020 investment management industry outlook | Deloitte Insights. (2019). Retrieved 2020, from https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/financial-services/ch-deloitte-wealth-management-private-banking-efma.pdf

McCormick, K. (2020, July 29). 12 Ways to Effectively Promote a New Product or Service. Retrieved from https://www.wordstream.com/blog/ws/2020/07/29/how-to-promote-a-product