NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | Nanyang Business School

RE6013:

Business Analytics and Applied Machine Learning

AY20/21 Semester 2

COVER PAGE

Team 10:

Ang Wan Qi (U1721634L)

Justin Peh (U1620687C)

Loh Xing Bao Colin (U1721073H)

Ma Jiebing (U1722200B)

Wei Zi Yun Mark (U1721491K)

Tutor: Mr Neumann Chew

Date: 4 April 2021

Table of Contents

Executive Summary

This project investigates the factors driving customers to churn from WhiteRock's retail clients from the dataset that provides customers' social and financial data. Our analysis revealed that one of the key factors driving churn is the age of the customer and the number of products the customer is using. Specifically, our investigations suggest that there is a higher churn rate among customers who are older and/or customers who use more than 2 bank products. We believe that these two factors are correlated, in that the 3rd and 4th banking products are those targeted at older clients. The suspected reason for churn is due to these products not meeting the financial needs of their intended senior-aged market. As such, clients who first joined the bank in their 30s and 40s are also likely to be those whose main intentions are to take up the 3rd and 4th product. But due to the lack of satisfaction with these products, they are very likely to churn as well. We produced a CART-based model that can predict which clients are likely to churn with 85.7% accuracy. The model suggested that the age, the age at the start of tenure and the number of banking products by the customer has much more importance than all other variables in explaining churn. As such, to lower churn rate, we recommend the bank targets its marketing towards younger prospects, whose needs they can meet better. Alternatively, the bank can improve its financial services meant for older clients.

1 Background

1.1 Customer Churn

Customer churn is defined as the loss of existing customers. It is an important metric in the business operations of White Rock and its subsidiaries as Customer Retention Costs (CRC) is typically lower than Customer Acquisition Costs (CAC). Reducing customer churn will thus lower overall business costs while maintaining the same volume of Assets Under Management (AUM).

1.2 Dataset

The dataset consists of socio-economic data of our customers obtained during the last financial year. For more details on the dataset, please refer to the Appendix 6.1 or the data dictionary document provided separately.

2 Data Cleaning and Preparation

2.1 Data Preparation

We dropped `RowNumber, CustomerId` and `Surname` because these variables will clearly not be helpful for data analytics. We also apply the `factor()` function to the following variables: `Geography, Gender, HasCrCard, IsActiveMember` and `Exited` because these variables are categorical.

3 Data Exploration

3.1 General comments on value distribution of various columns

1. Most customers in the dataset were retained (almost 80%).
2. Customers are generally high net worth individuals (greater than USD100,000 estimated salary) and salary distribution is largely similar across most age groups, less those above 75 years of age as shown in **Figure 1**.

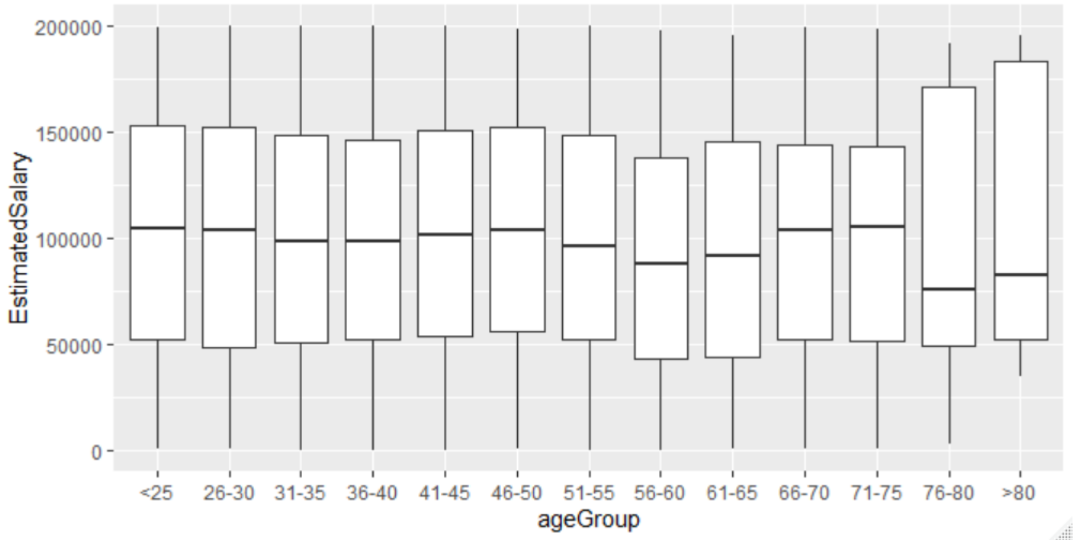


Figure 1: Box plot of EstimatedSalary against the various ageGroup

3. Most customers are based in France
4. Most customers use either 1 or 2 products (97% of the dataset), with the remaining using either 3 or 4 products.
5. The number of customers for each tenure duration is (for the most part) evenly distributed as shown in **Figure 2**.

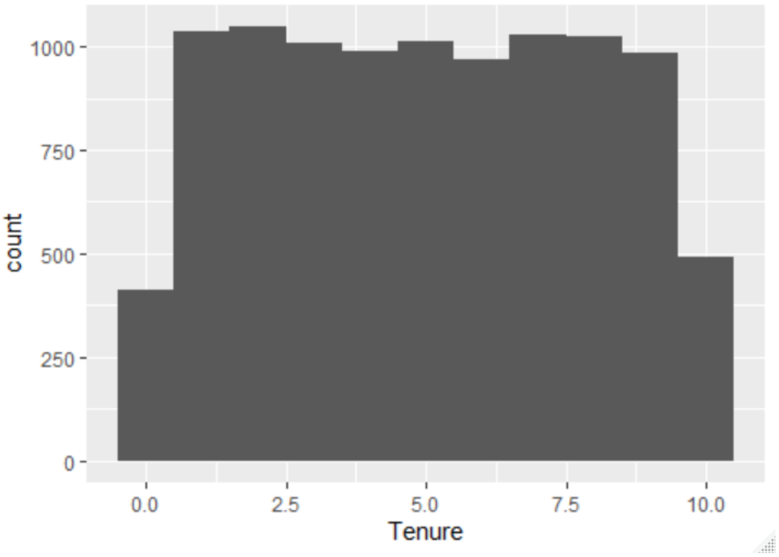


Figure 2: Number of customers against tenure

3.2 Correlation between continuous variables

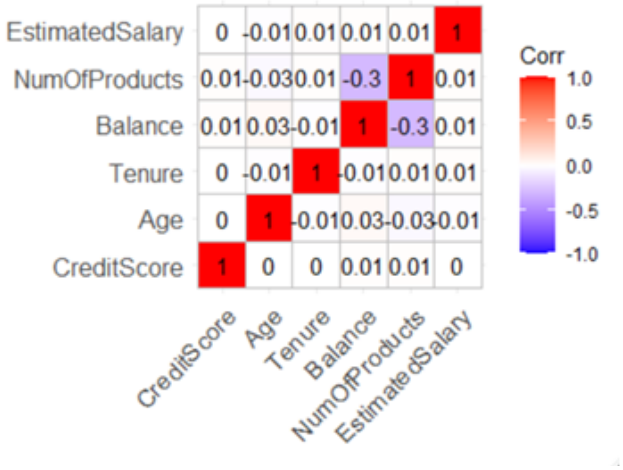


Figure 3: Correlation between continuous variables

We believed it was possible that this bank may judge a customer's creditworthiness based on `EstimatedSalary, NumOfProducts, Balance, Tenure and Age`. To check against this, we used the correlation matrix illustrated in **Figure 3**. The matrix showed that `CreditScore` has little relationship with the other continuous variables. Hence, for the purpose of this project, we can assume these variables are independent of each other.

3.3 Checking for possible gender inequality

Since it is reasonable to assume that financial metrics (i.e., `EstimatedSalary` and `Balance`) may influence churn decisions, we checked for any correlation between `Gender` and either of these variables, to account for possible gender inequality. The presence of gender inequality would increase importance of `Gender` in explaining churn.

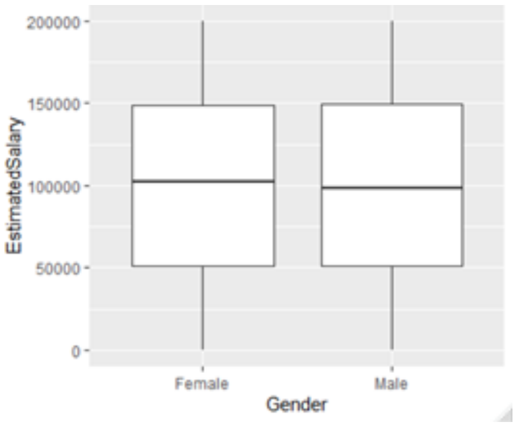


Figure 5: EstimatedSalary against Gender

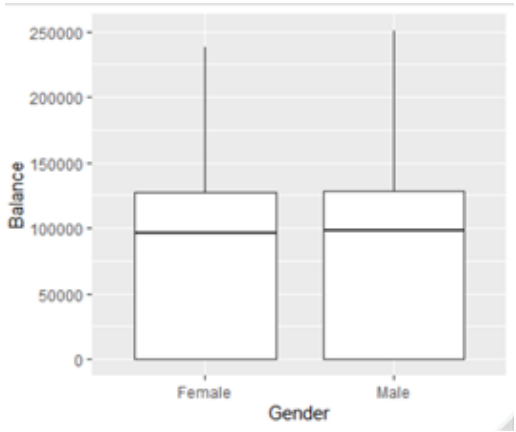


Figure 4: Balance against Gender

Based on these two box plots in **Figure 5** and **Figure 4**, the data seems to suggest that financial metrics are similar across both genders. Since the financial profiles across both genders are similar, we should expect their CreditScore to be similar as well if there is no gender discrimination of any sort.

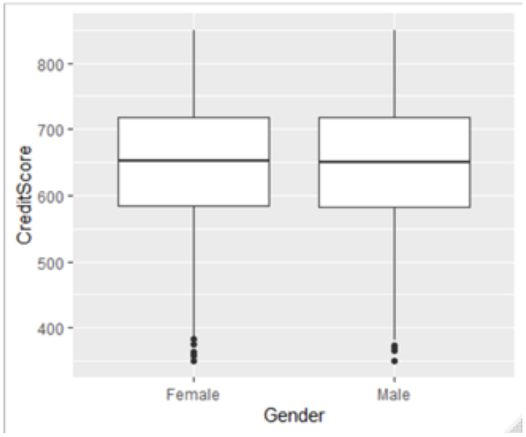


Figure 6: CreditScore against Gender

Given that CreditScore distribution is similar across males and females illustrated in **Figure 6**, we can therefore assume that Gender is independent of EstimatedSalary, Balance and CreditScore.

3.4 Exploring customer distribution across Geography

Looking at **Figure 7**, it appears that although most of the customers in the dataset are registered in France, Germany-based customers show an unusually higher proportion to churn compared to the others. We investigated this anomaly further as it may provide clues as to why customers churn.

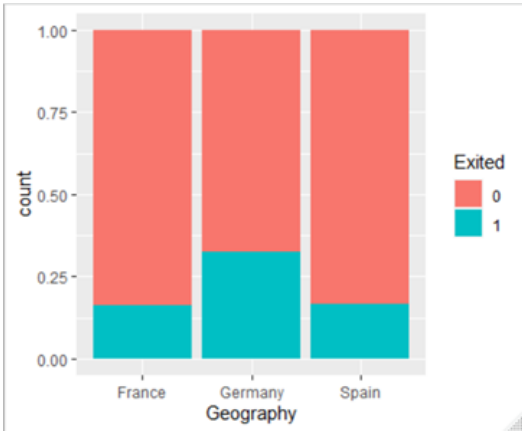


Figure 7: Number of customers in each location that stayed and left

We investigated if it could be due to Germany having an especially higher (or lower) income distribution. However, **Figure 8** seems to show that income level is about the same across all 3 countries.

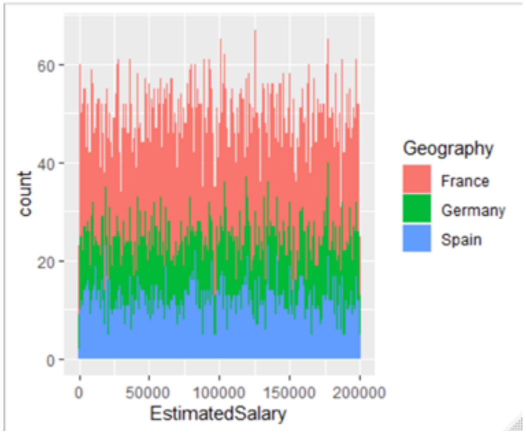
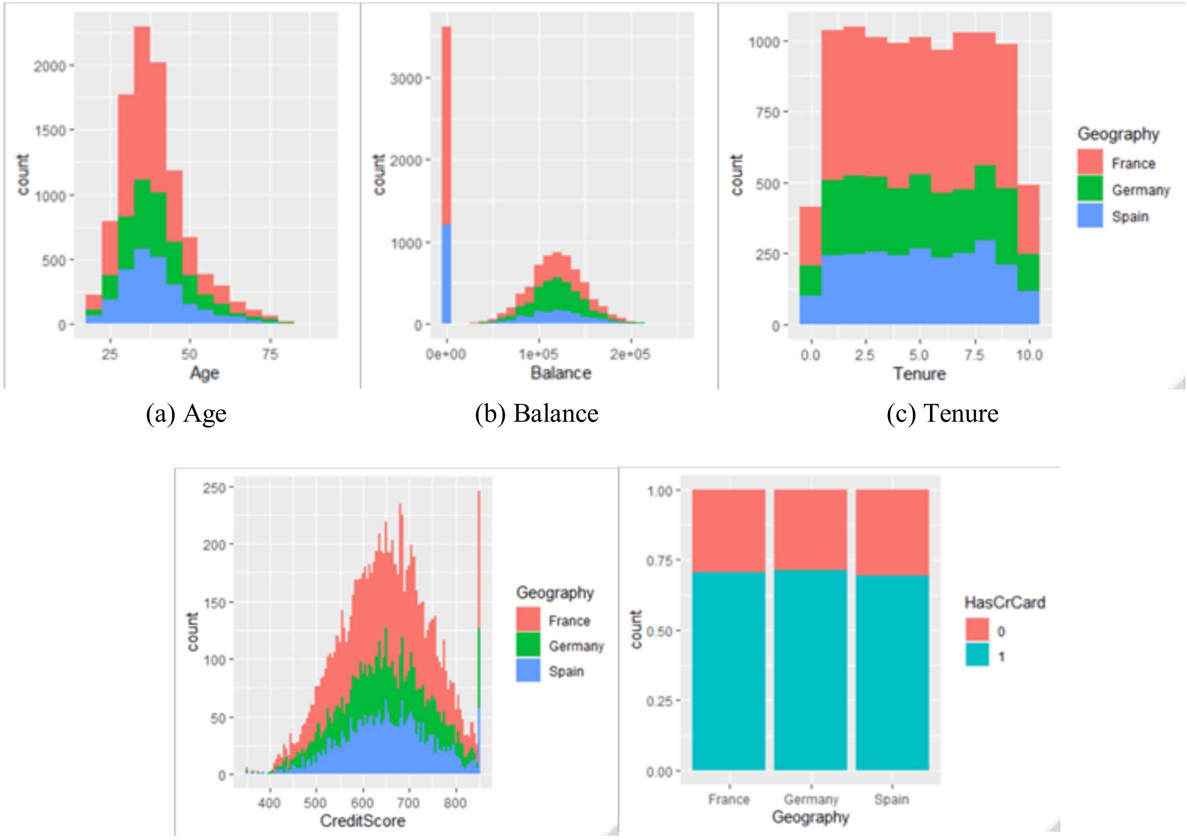


Figure 8: Count of customers against EstimatedSalary in each location

We also then checked if Germany-based customers stood out in terms of other variables that might naturally be different across different countries.



(a) Age

(b) Balance

(c) Tenure

(d) CreditScore

(e) HasCrCard

Figure 9: Distribution of various values across the 3 countries

The diagrams in **Figure 9** show the distribution of values across the 3 countries for the following variables: `Age`, `Balance`, `Tenure`, `CreditScore` and `HasCrCard`. These diagrams show that there is no distinguishable difference in terms of these variables. However, it seems that Germany-based customers tend to hold more products (3 or 4) compared to those in other countries as shown in **Figure 10**. This prompted us to investigate if the number of products had some relationship with customer churn.

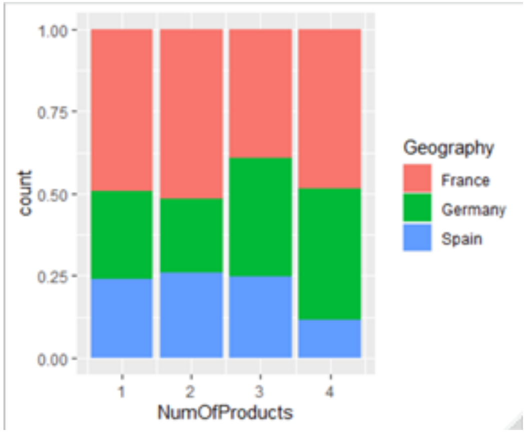


Figure 10: Number of products each customer holds in across 3 countries

3.5 Investigating Product Uptake VS Customer Churn

In this section, we aim to further investigate the relationship between product uptake rate, customer features, and churn rate to determine if there exists any links that may explain causality between our product offerings and customer churn.

We first define the following variable, `NumOfProducts`, which describes the number of our financial products that each customer is subscribed to. We first take a look at the number of customers with each `NumOfProducts`, where `NumOfProducts` = {1, 2, 3, 4}.

Number of Products Used (`NumofProducts`)	Customer Count (Total = 10000)
1	5048
2	4590
3	266
4	60

Figure 11: Number of customers with the respective number of products

The metrics in **Figure 11** can also be represented in graphical format as shown in **Figure 12**:

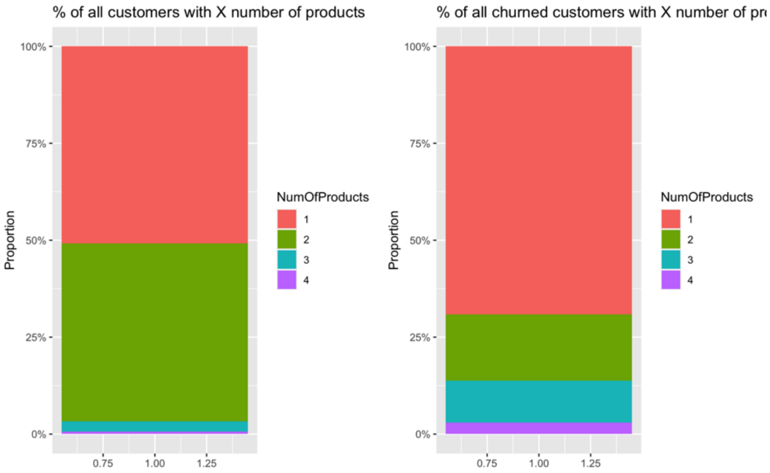


Figure 12: Percentage of customers with X products

As we can see from **Figure 12** , the majority of customers (>95%) have either 1 or 2 products. Conversely, those with 3 or 4 products are in the minority. However, customers with 3 or 4 products represent an outsized proportion of churned customers. This is important and will be explored later in the report. We can also see that those subscribed to 2 of our financial products were far less likely to churn than those that subscribed to 1, 3 or 4, and notably, customers with 4 products see a near 100% churn rate as illustrated in **Figure 13**.

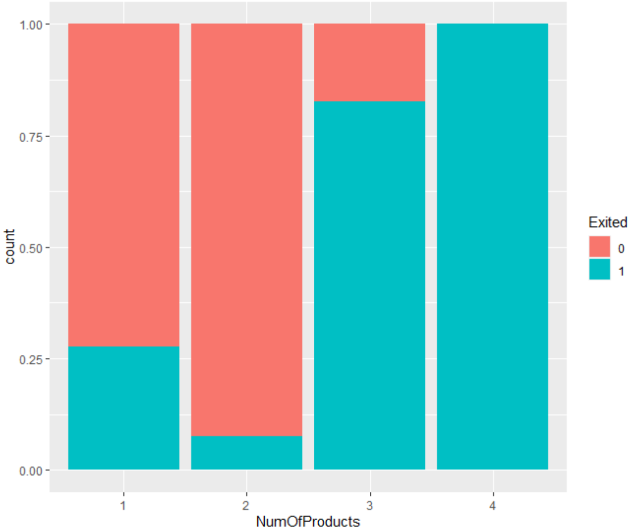


Figure 13: Proportion of customers who exited/stayed with X products

Here, we assume that customers with the same number of products are mostly using the same set of products. Specifically:

1. Most customers with 1 product are using the same product, which we shall call Product A.
2. Most customers with 2 products are using the same 2 products, which we shall call Products A and B.
3. Most customers with 3 or more products are using Product A and B, as well as C (and D) as their third (and fourth) product.

3.5.1 Multiple-Product Customers

Conventional wisdom would indicate that customers who use a greater number of products will be more engaged with the company and will hence exhibit lower churn rates. However, through our previous findings, we saw that churn rate is substantially *higher* for customers who have used 3 or 4 products from the bank. This section will hence look to determine potential reasons as to why this might be the case. To do this, we extract the subset of customers with >2 products as `hasManyProd`. From there, we compare certain descriptive characteristics of this group of customers with those who only use/used <2 products:

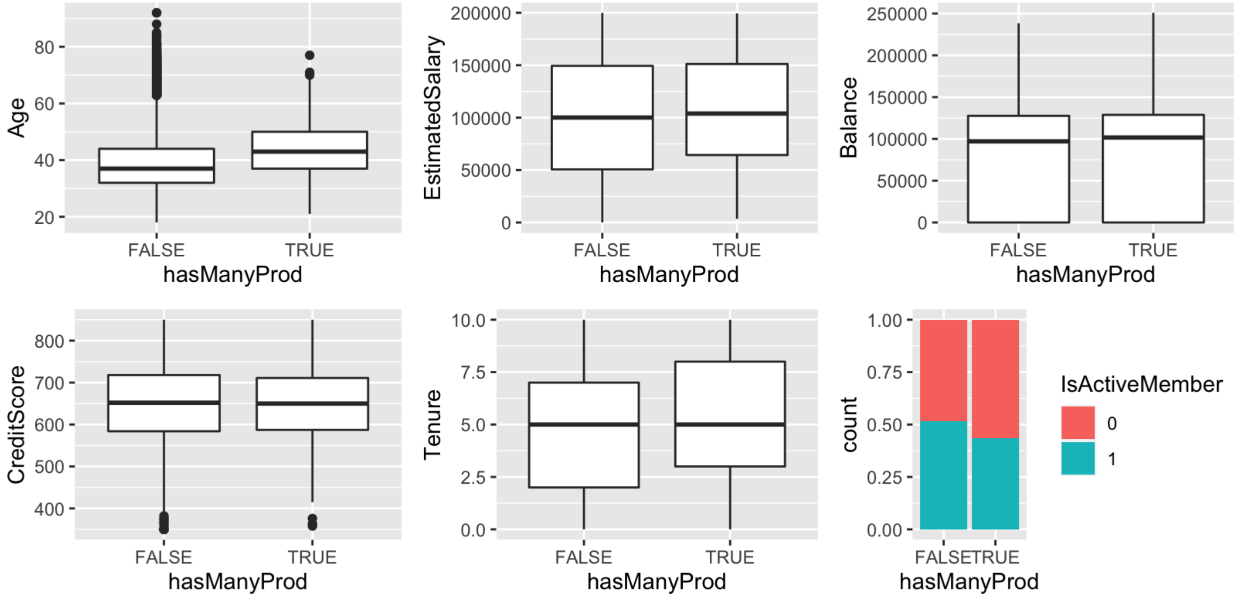


Figure 14: Box plots comparing hasManyProd with different variable

Henceforth, we will refer to customers with 3 or more products as MPC (many-product-customers), and customers with less than 3 products as FPC (few-product-customers).

From the graphs in **Figure 14**, we can infer the following:

1. MPC tended to be generally older than FPC
2. MPC generally exhibited similar salaries to FPC, although the salary floor for MPC is higher
3. MPC and FPC generally have very similar credit scores and account balances
4. MPC generally have longer tenures with the bank than FPC
5. MPC contain a slightly *lower* proportion of active members compared to FPC

Although churners also tended to have lower account balances than stayers, churn decisions are more likely to affect decisions about account balances than vice versa. Hence, we only consider Age (Point 1 above) to be significant.

We guessed that most customers churn after taking up C and D because they found them unsatisfactory for their financial needs. Even if this inference was true, they do not explain the majority of churners (who only have 1 or 2 products).

Further, we also guessed that customers who churned with just 1 product have similar financial needs as the customers who churn with 3 or 4 products. Specifically, it was possible that customers who churned on 1 product were those who were using Product C (rather than Product A). To test this hypothesis, we created a new variable: `ageAtStartOfTenure` (`Age - Tenure`). This variable represented the age of the customer when they started using their first product offered by the bank. We compare the age distribution between stayers and churners, all of whom have only 1 product.

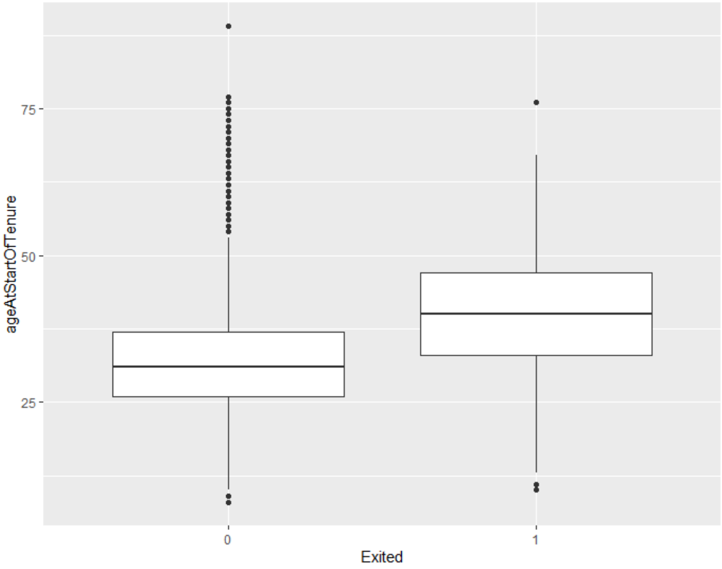


Figure 15: AgeAtStartOfTenure against Exited

From the box plot in **Figure 15**, it can be seen that those who churned tended to start their tenure later in life than those who stayed. In fact, their age distribution appears similar to customers who subscribed to 3 or more products.

Given the similarity in age, we hypothesize that customers who own one product and only recently started their tenure in their late 30s to 40s are those who are actively comparing between banks to find the most competitive rates and packages for products similar to Product C. This may explain their higher flight risk, especially when even long-time customers churned as well upon using Product C.

3.5.2 Possible Theories

Given that most customers have 1 or 2 products, our hypothesis is that products A and B are likely savings accounts and bank loans, which most customers use and where customers' expectations are usually low. It thus forms a stable customer base where most customer needs are met and less likely to churn. As these are basic banking services that are largely undifferentiated products across the industry, these products could easily meet clients' expectations, and not induce clients to churn.

Products C and D tend to be the 3rd and 4th products adopted by older customers with longer tenures. Due to the low uptake rate but outsized churn rate for these products, we can infer that these products are more niche than their more popular counterparts. Specifically, we guess that Product C and D correspond to various types of retirement plans. However, prospective customers may have much higher expectations of these specialised products, perhaps due to them being more advanced in years. Hence, the bank may find it harder to satisfy customers as easily as it did for Product A and B, which are more generic. This, in turn, leads to lower adoption rates, explaining the substantially higher churn rates among product users who need such products.

3.6 Impact of Customer's Financial Habits on Churn Rate

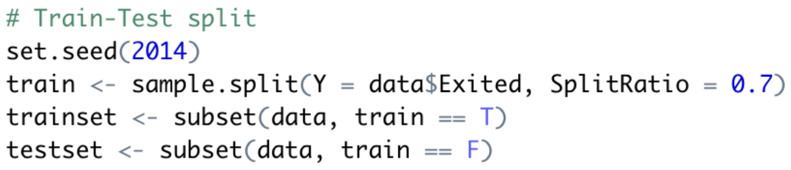
We investigated their financial habits through the creation of two new variables: `depositRate` (`Balance` divided by `Tenure`) as well as their savings-to-salary ratio (`depositRate` divided by `EstimatedSalary^2`), as an analog for their spending habits. We did not find any conclusive relationship between a customer's financial habits and their propensity to churn. Hence, it does not warrant further investigation.

4 Churn Prediction

After an in-depth exploration of the dataset, we proceed to build models to predict the possibility of a customer churning. For this project, we make use of 2 different models namely: Logistics Regression and CART. The process and results are detailed in the sections that follow.

4.1 Preparation of Data

Based on the above data exploration, we decided to make use of `AgeAtStartOfTenure` in addition to the original variables available in the dataset. The list of `x` variables used in the prediction model is as follows: `CreditScore`, `Geography`, `Gender`, `Age`, `Tenure`, `Balance`, `NumOfProducts`, `HasCrCard`, `IsActiveMember`, `EstimatedSalary`, `AgeAtStartOfTenure`. There is a total of 10,000 rows and a 70:30 train-test split to obtain a trainset of 7,000 rows and a testset of 3,000 rows as shown in **Figure 16**. The `trainset` will be used to train the model to predict the `y` variable `Exited` and the `testset` will be used to evaluate the performance of the trained model on unseen data. In addition, we also converted `NumOfProducts` into a categorical variable, since customers subscribing to different numbers of products exhibit very different characteristics.

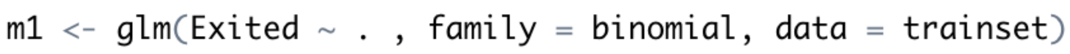


```
# Train-Test split
set.seed(2014)
train <- sample.split(Y = data$Exited, SplitRatio = 0.7)
trainset <- subset(data, train == T)
testset <- subset(data, train == F)
```

Figure 16: R code for train-test split

4.2 Logistic Regression

To predict churn with logistic regression, we first create a model using the `glm` function in R, `'binomial'` for the `family` parameter and `trainset` as the data as shown in **Figure 17**.



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

Figure 17: R code to generate logistics regression model

4.2.1 Results for Logistics Regression

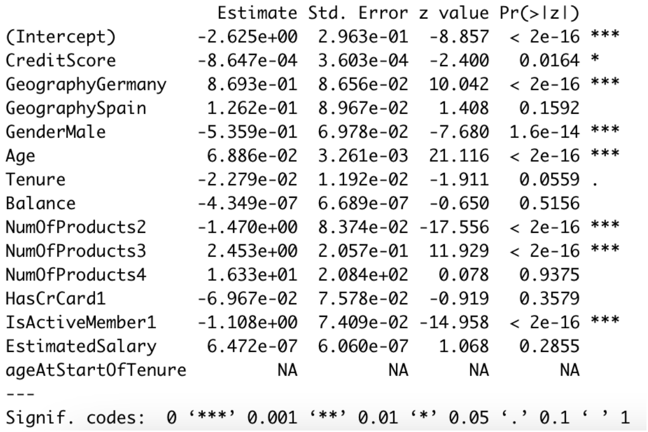
This section presents the results and performance obtained from the logistic regression model.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4515	-0.5979	-0.3724	-0.185	3.2498

Table 1: Deviance residuals of logistics regression model

Coefficients:



```
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -2.625e+00  2.963e-01  -8.857  < 2e-16 ***
CreditScore         -8.647e-04  3.603e-04  -2.400   0.0164 *  
GeographyGermany     8.693e-01  8.656e-02  10.042  < 2e-16 ***
GeographySpain       1.262e-01  8.967e-02   1.408   0.1592    
GenderMale          -5.359e-01  6.978e-02  -7.680  1.6e-14 ***
Age                  6.886e-02  3.261e-03  21.116  < 2e-16 ***
Tenure              -2.279e-02  1.192e-02  -1.911   0.0559 .  
Balance             -4.349e-07  6.689e-07  -0.650   0.5156    
NumOfProducts2      -1.470e+00  8.374e-02 -17.556  < 2e-16 ***
NumOfProducts3       2.453e+00  2.057e-01  11.929  < 2e-16 ***
NumOfProducts4       1.633e+01  2.084e+02   0.078   0.9375    
HasCrCard1          -6.967e-02  7.578e-02  -0.919   0.3579    
IsActiveMember1     -1.108e+00  7.409e-02 -14.958  < 2e-16 ***
EstimatedSalary      6.472e-07  6.060e-07   1.068   0.2855    
ageAtStartOfTenure          NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 18: Coefficients for logistics regression model

Confusion Matrix on `Trainset`:

n = 7000	Predicted: 0 - Not Exited	Predicted: 1 - Exited
Actual: 0 - Not Exited	5358	216
Actual: 1 - Exited	904	522

Table 2: Trainset confusion matrix for logistics regression model

Metric	**Value**
Precision	522 / (216+522) = 522/738 = 0.7073
Prevalence	(904+522)/7000 = 1426/7000 = 0.2037
Accuracy	(5358+522)/7000 = 5880/7000 = 0.84
Misclassification Error	(904+216)/7000 = 1120/7000 = 0.16

Table 3: Various performance metrics for logistics regression model's trainset

Based on the trainset confusion matrix and performance metrics in **Table 2** and **Table 3** , the model has a precision of 0.7073 which means that the model is correct for 70.73% of the time when it predicts that the customer will churn. The model also achieved a performance accuracy of 84% on the trainset.

Confusion Matrix on `Testset`:

n = 3000	**Predicted: 0 - Not Exited**	**Predicted: 1 - Exited**
Actual: 0 - Not Exited	2317	72
Actual: 1 - Exited	376	235

Table 4: Testset confusion matrix for logistics regression model

Metric	**Value**
Precision	235/(72+235) = 235/307 = 0.7654
Prevalence	(376+235)/3000 = 611/3000 = 0.2037
Accuracy	(2317+235)/3000 = 2552/3000 = 0.8507
Misclassification Error	(72+376)/3000 = 448/3000 = 0.1493

Table 5: Various performance metrics for logistics regression model's testset

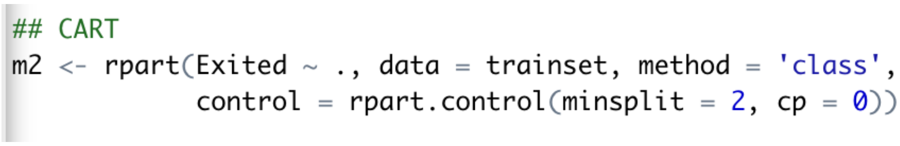
Based on the testset confusion matrix and performance metrics in **Table 4** and **Table 5**, the model has a precision of 0.7654 which means that the model is correct for 76.54% of the time when it predicts that the customer will churn. The model also achieved a performance accuracy of 85.07% on the testset, an increase of approximately 1% from its performance on trainset. This suggests that the model performs relatively better on unseen data and is not overfitted on the trainset data.

4.3 CART

There are three major steps involved in generating a model using CART and there are outlined as follows:

1. Grow the CART tree to the maximum
2. Prune the CART tree to the minimum
3. Select the most optimal CART tree as the final model

To grow CART tree, we use the `rpart` function in R, `'class'` for the method parameter and set the `minsplit = 2` and `cp = 0` so as to grow the tree to the maximum using the trainset data as shown in **Figure 19**.

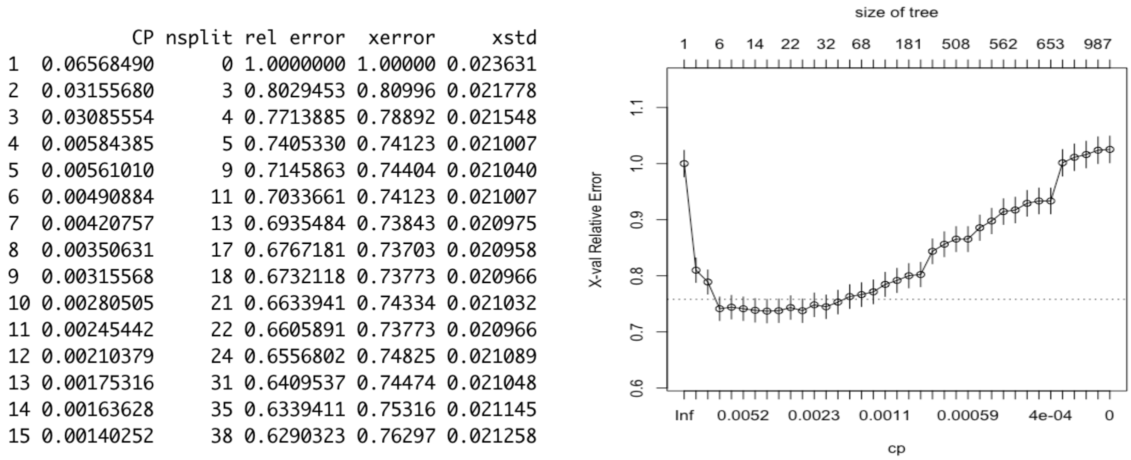


```
## CART
m2 <- rpart(Exited ~ ., data = trainset, method = 'class',
            control = rpart.control(minsplit = 2, cp = 0))
```

Figure 19: R code to generate a CART tree

4.3.1 Results for CART

With the CART tree generated, we obtain the complexity parameter (CP) table and plot as shown in **Figure 20**.



```
           CP nsplit rel error  xerror      xstd
1  0.06568490      0 1.0000000 1.00000 0.023631
2  0.03155680      3 0.8029453 0.80996 0.021778
3  0.03085554      4 0.7713885 0.78892 0.021548
4  0.00584385      5 0.7405330 0.74123 0.021007
5  0.00561010      9 0.7145863 0.74404 0.021040
6  0.00490884     11 0.7033661 0.74123 0.021007
7  0.00420757     13 0.6935484 0.73843 0.020975
8  0.00350631     17 0.6767181 0.73703 0.020958
9  0.00315568     18 0.6732118 0.73773 0.020966
10 0.00280505     21 0.6633941 0.74334 0.021032
11 0.00245442     22 0.6605891 0.73773 0.020966
12 0.00210379     24 0.6556802 0.74825 0.021089
13 0.00175316     31 0.6409537 0.74474 0.021048
14 0.00163628     35 0.6339411 0.75316 0.021145
15 0.00140252     38 0.6290323 0.76297 0.021258
```

(a) CP Table

(b) CP Plot

Figure 20: CP table and plot

Judging from both the table and plot, we employ the 1-SE rule to choose the optimal tree. We chose the 4th tree as it is the simplest tree that still performs well, with a relatively low 10-fold cross-validation (cv) error (`xerror` column) that lies below the dotted line of the plotted graph.

We decided to use the 1SE rule over the minimum CV error tree to choose the optimal tree as it is a more stable solution. For the minimum CV error tree, a small change in data could lead to a different solution, making it relatively unreliable.

Since the 4th tree is optimal, we choose a CP value between the 3rd and 4th tree CP values, i.e. `cp1 <- sqrt(0.03085554*0.00584385)`. Using the 1-SE rule, we use this specific value of CP to `prune()` the maximal tree to get a specific subtree. We then obtain the following optimal tree:

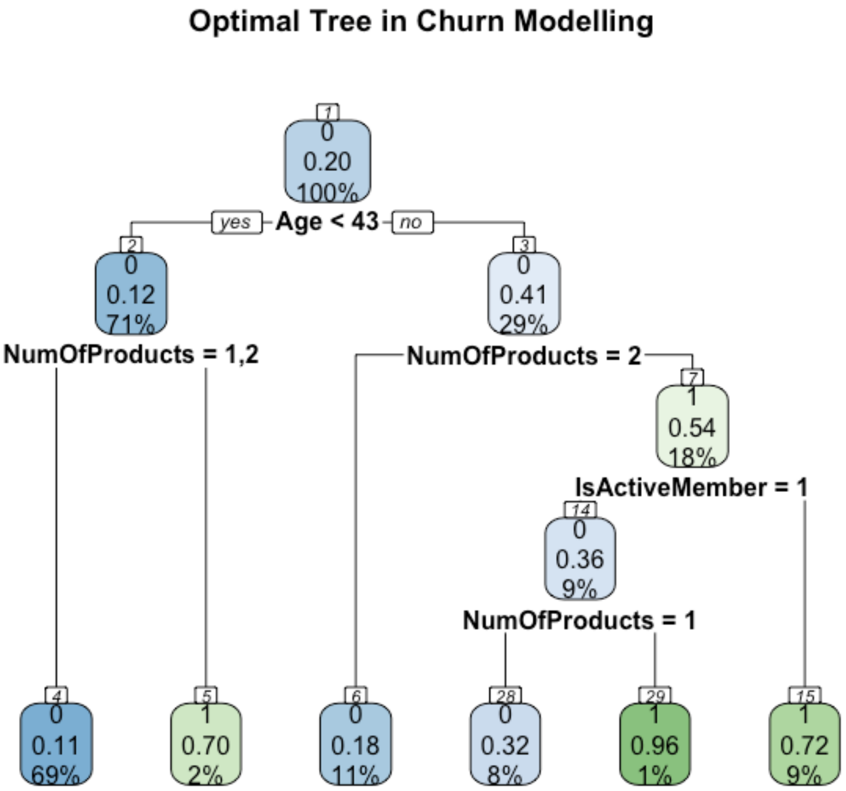
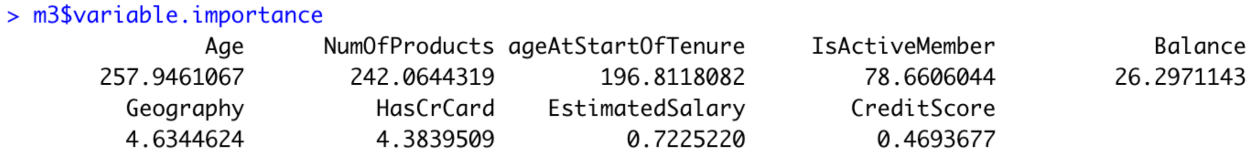


Figure 21: Optimal tree after pruning

Investigating the importance of all variables, we observe that `Age` is the most important, followed by `NumOfProducts` and then `ageAtStartOfTenure` as shown in **Figure 22**.



```
> m3$variable.importance
               Age      NumOfProducts ageAtStartOfTenure     IsActiveMember            Balance
       257.9461067        242.0644319        196.8118082         78.6606044         26.2971143
         Geography          HasCrCard    EstimatedSalary        CreditScore
         4.6344624          4.3839509          0.7225220          0.4693677
```

Figure 22: Variable importance

With the model in place, we tested it on the testset and obtained the confusion matrix in Figure (insert figure number).

n = 3000	Predicted: 0 - Not Exited	Predicted: 1 - Exited
Actual: 0 - Not Exited	2317	72
Actual: 1 - Exited	357	254

Table 6: Testset confusion matrix for CART model

Metric	Value
Precision	254/(72+254) = 254/326 = 0.7791
Prevalence	(357+254)/3000 = 611/3000 = 0.2037
Accuracy	(2317+254)/3000 = 2517/3000 = 0.857
Misclassification Error	(72+357)/3000 = 429/3000 = 0.143

Table 7: Various performance metrics for CART model's testset

Based on the testset confusion matrix and performance metrics in **Table 6** and **Table 7**, the model has a precision of 0.7791, an approximate 1% increase from the Logistics Regression model in Section 4.2. This means that the model is correct for 77.91% of the time when it predicts that the customer will churn. The model achieved a performance accuracy of 85.7% on the testset, which is comparable to the Logistics Regression model. From these results, we can conclude that both the Logistics Regression and CART model have similar performance on unseen data, with the CART model slightly outperforming the Logistic Regression model by 0.63%.

With the models developed, we know the profiles of those likely to churn. We can then adapt our strategies and products to better target those with a lower likelihood of churning, while developing new ones to better retain those who have a higher propensity to churn.

5 Remedy Recommendations

We initially planned to develop an engine that can identify customers with high churn risk and give personalised recommendations to the bank on how to best reduce their churn risk (perhaps by prompting a customer to deposit more money into the bank). However, our analysis and CART-based POC suggests that the biggest factor of churn is age and number of products. So, a recommendation engine may naively recommend changing the customers' age. Hence, we instead recommend the bank to investigate their niche services targeted at customers between 18 to 43 years old. Specifically, we recommend that the bank finds out in what way these services fall below customer expectations, and how they could improve upon them. Also, the bank should target its marketing towards prospective leads in a younger age group since their offerings appear to satisfy the financial needs of this age group better.

6 Appendix

6.1 Information on the Dataset

Name of Dataset: Churn_Modelling.csv

Number of Rows: 10,000

Legend: Original Variables, Derived Variables

Variable Name	(Assumed) Description	Default Data Type	Final Data Type	Example / Levels
RowNumber	Row number in dataset	integer	integer	0
CustomerID	ID to uniquely identify customers	integer	integer	15634602
Surname	Last name of customers	character	character	“Hargrave”
CreditScore	FICO credit system score	integer	integer	619
Geography	Possibly where the customer opened the account	character	Factor w 3 levels	“France”, “Germany”, “Spain”
Gender	Gender of customer	character	Factor with 2 levels	“Female”, “Male”
Age	Age of customer	integer	integer	42
Tenure	How many years ago the account was opened	integer	integer	2
Balance	Account balance	numeric	numeric	83808
NumOfProducts	Number of the bank’s products the customer uses	integer	Factor with 4 levels	“1”, “2”, “3”, “4”
HasCrCard	Whether they own at least 1 of the bank’s credit card	integer	Factor with 2 levels	“0”, “1”
IsActiveMember	Whether the bank considers the customer to be actively using their services	integer	Factor with 2 levels	“0”, “1”
EstimatedSalary	The customer’s self declared annual salary in USD	numeric	numeric	101349
Exited	Whether the customer has churned	integer	Factor with 2 levels	“0”, “1”
AgeAtStartofTenure	The age of customer when they first	N.A	integer	40

	opened the account			
hasManyProd	Whether the customer has more than 2 products	N.A	logical	TRUE
has4Prod	Whether the customer has 4 products	N.A	logical	TRUE
has3Prod	Whether the customer has 3 products	N.A	logical	TRUE
depositRate	The mean balance over the length of tenure. Obtained by dividing the customer’s balance by their tenure	N.A	numeric	41904
savingsToSalaryRatio	The ratio between a customer’s mean balance per tenure years and their estimated salary	N.A.	numeric	0.372
salaryToAgeSq	The ratio between a customer’s estimated salary and the squared of their age	N.A	numeric	57.2
salaryQuartile	The quartile that a customer’s salary falls in. Q1: EstimatedSalary < 51002 Q2: 51002 < EstimatedSalary < 100193 Q3: 100193 < EstimatedSalary < 149388 Q4: EstimatedSalary > 149388	N.A	Factor with 4 levels	"1Q", "2Q", "3Q", “4Q”