Master's Thesis – MSc Data Science and Marketing Analytics

Customer Segmentation and Churning Analysis

A Quantitative Study on Credit Card Customers

Name: Emilia Setyanda

Student number: 447371

Supervisor: Prof. dr. Andreas Alfons

Second assessor: Prof. dr. Erjen van Nierop

Date of submission: 15 August 2021

Abstract

Popularity and preference for payment using credit card have been on the rise since 2016. Since customers could own more than 1 credit card, banks face the problem of churning. Churning brings losses to the bank as attracting new customers costs more than retaining current ones. Longer customers also tend to spend more, which brings higher revenues to the bank. Many previous research for segmentation and churning prediction has been done, but little to none has gone in-depth to discover more about the variables of importance and the values that indicate churning potentials. This paper attempts to conduct segmentation using k-means, rough k-means, and k-medoids clustering, where rough k-means with 2 clusters turn out to have the highest silhouette score. Both logistic regression and random forest are used for churning prediction, where the latter obtains higher rates in all four performance metrics: accuracy, recall, precision, and F1 score. In both clusters, total transaction count, amount, and their changes from Q4 to Q1, as well as total revolving balance and relationship count with the bank turn out to be the most important variables in improving the accuracy and purity of the nodes in the churning prediction. The length of the relationship and contacts count between the bank and its customers are more important for improving the accuracy and purity for big spender customers in cluster 1. Meanwhile, the customers' age and months of inactivity are better indicators of churning for the passive customers in cluster 2.

Acknowledgments

This thesis symbolizes my academic and personal growth throughout the 4 years I spent at Erasmus University Rotterdam, both as a bachelor's and master's student. It stands as a proof that through hard work and perseverance, I could conduct research independently using the knowledge that I have gained during my master's courses.

First, I want to thank my supervisor, Professor Andreas Alfons. These past 12 months have been a challenge for me to balance between finishing my master's and working full-time as an intern. However, Professor Alfons has been understanding, pointing me towards the right direction in my thesis structure, and encouraging me when I have doubts about myself.

I would also like to thank my friends who are both in Europe or abroad, for checking in on me, accompanying me to refresh my mind, and supporting me. This academic experience has been far more memorable with them by my side.

Most importantly, I want to thank my parents who have been my constant supporter throughout all phases of my life, even when they are more than 10,000 kilometers away in Indonesia. Words cannot describe how grateful and lucky I am to have them as my parents.

Rotterdam, 6 August 2021

Emilia Setyanda

List of Abbreviations

Abbreviation	Description
ALE	Accumulated Local Effects
AUC	Area Under Curve
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CRM	Customer Relationship Management
DT	Decision Tree
FCM	Fuzzy C-Means
FN	False Negative
FP	False Positive
KNN	K-Nearest Neighbors
LR	Logistic Regression
LRFM	Length, Recency, Frequency, and Monetary
LRFMP	Length, Recency, Frequency, Monetary, and Periodicity
MDA	Mean Decrease Accuracy
MDG	Mean Decrease Gini
NB	Naïve Bayes
OOB	Out-of-bag
PP	Percentage Points
RF	Random Forest
RFM	Recency, Frequency, and Monetary
RST	Rough Set Theory
SOM	Self-Organizing Map
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VIF	Variance Inflation Factor

Table of Contents

Abstract ii

Acknowledgements iii

List of Abbreviations iv

Table of Contents v

1. Introduction 1

2. Theoretical Framework and Hypotheses Development 2

2.1. Customer segmentation 2

2.2. Customer churning 4

2.3. Combination of customer segmentation and churning 5

2.4. Hypotheses development 6

3. Data 7

3.1. Data description 7

3.2. Descriptive statistics 9

3.3. Data splitting 11

4. Methodology 11

4.1. Customer segmentation methods 12

4.1.1. K-means clustering 12

4.1.2. Rough k-means clustering 13

4.1.3. K-medoids clustering 14

4.1.4. Methods to determine the number of clusters 14

4.2. Churning prediction methods 16

4.2.1. Logistic regression 16

4.2.2. Random forest 18

4.2.3. Evaluation metrics for predictive classification methods 22

5. Results 23

5.1. Clustering 23

5.2. Churning prediction 26

5.2.1. Logistic regression 26

5.2.2. Random forest 29

6. Conclusion 36

References 39

Appendices 44

Appendix A. Boxplot of descriptive statistics using the 5 segmentation variables for both clusters 1 and 2 44

1. Introduction

Results from The Diary of Consumer Payment Choice in 2020 showed that preference for credit card payments has been increasing since 2016 while those for cash and debit card have decreased and remained stable respectively. Aligned with the preferences, the share of payment instrument usage with credit cards has also increased since 2016 (see Figure 1 below).

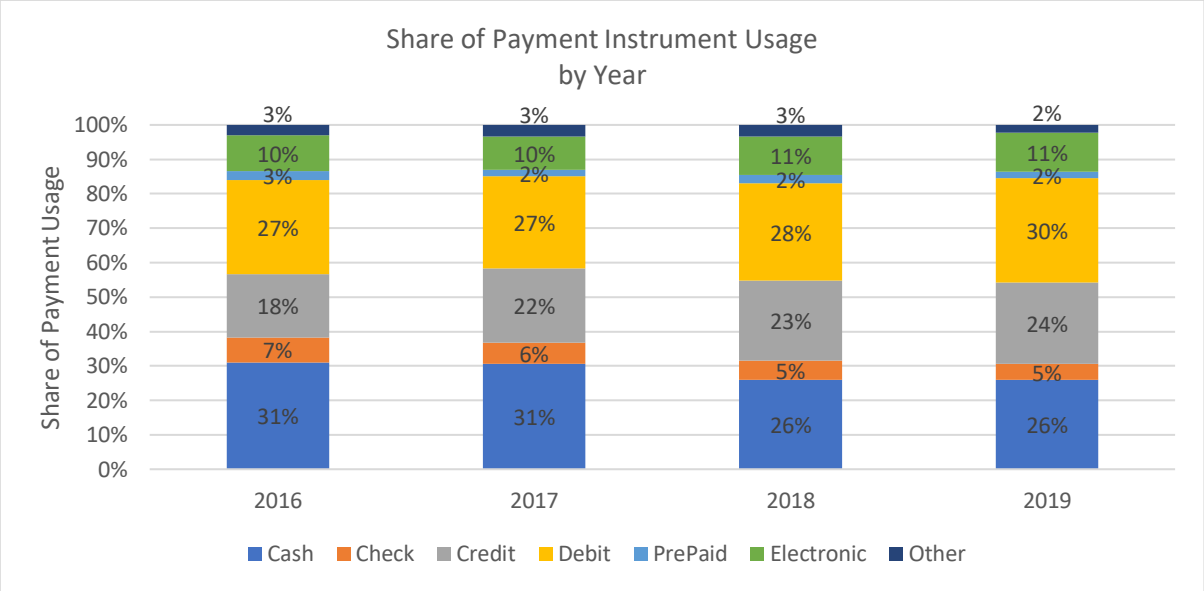

Figure 1. The proportion of payment instrument usage from 2016 to 2019. Reprinted from *2020 Findings from the Diary of Consumer Payment Choice* (p.5) by L. Kim, R. Kumar, and S. O'Brien, 2020, San Francisco: Federal Reserve System. Copyright 2020 by the authors.

However, in some countries, individuals typically own more than 1 credit cards in their wallets; for example, in Japan and the United States, the average person has 2 and 4 credit cards respectively (Frankel, 2020; Jiang et al., 2020). This fosters the customers' ease of churning. While the general acquisition cost for customers in retail banking is approximately $200, a typical customer generates $150 of revenue in a year (Lax, 2016). This means that banks would need to retain their customers for more than 1 year before hitting the breakeven mark. Moreover, the acquisition costs of new customers are typically 5 times higher than the retention costs of current ones (Wertz, 2018). Thus, more effort should be directed towards customer retention. A possible solution is to create customer segments and check which segments are most profitable or have the highest likelihood of churning. The focus of this research would be to find customer segments based on their demographics and credit card details. Hence, the following research question is proposed:

What customer segments could be found within credit card customers?

To expand the research, the following sub-questions would also be explored:

1. What characteristics differentiate the segments with high and low proportions of churning customers?
2. Which factors are most indicative of customers' likelihood to churn in different segments?

Previous research combining both segmentation and churning prediction has been done by Rajamohamed and Manokaran (2017), where their research focuses on finding the model with the best performance metrics. However, little to no research has been done to describe which variables are important and at which point do the values point towards an increasing likelihood of churning. This paper attempts to bridge that gap for future research by using model-agnostic method for black box interpretation. Aside from the financial advantages of retaining customers described in the previous paragraph, the detailed explanation on variables of importance would benefit the bank's stakeholders as it could set up a cautionary alert whenever customers indicate potential signs of churning. Retained customers are also more likely to bring higher profits, spend more, and spread positive word-of-mouth due to the increasing trust (Griffin, 2002).

In the next section, discussion on previous literatures revolving around customer segmentation and churning prediction would be given. A deeper explanation of the primary and secondary research questions is done in the hypotheses development section. This is followed by the data section which includes data description of all variables in the data set, descriptive statistics, and data preparation or splitting. Afterward, the methodology section dives into the machine learning models chosen for both the segmentation and churning prediction. Next, results from the best-performing models and their marketing implications are elaborated. Lastly, the conclusion would answer the research question, summarize the main findings, and state the limitations and suggestions for further research.

2. Theoretical Framework and Hypotheses Development

In this section, previous research that is related to this paper would be discussed. First, similar research on customer segmentation in various industries is presented. This is followed by a section that revolves around customer churn prediction.

2.1. Customer segmentation

Customer segmentation is the process of dividing customers into different groups, where customers within a group share more similarities with one another compared to those outside of the group. The customer segments should also display differences between groups which are important for businesses.

The practice of customer segmentation in a marketing strategy context was first mentioned by Smith (1956). The paper mentioned several benefits from segmentation, namely to satisfy the different needs of customers. Through the recognition of different demand schedules, businesses could also profit from segmentation through more precise marketing campaigns; this would consequently lead to better brand awareness and higher revenues.

Early research about customer segmentation using credit card data has shown that demographic variables such as age and income are frequently encountered. Meadows and Dibb (1998) conducted a case study of 4 banks in England and found that all 4 banks used age and income to indicate the customers' life stage and value. Using only demographic features are inadequate to define and segment purchase behavior for financial products (Brooks & White, 1996). However, these simple forms of segmentation were done during a time where data availability and statistical computing power were not as advanced as they are now. Thus, the researchers suggested for further studies to include variables such as lifestyle, family circumstances, and spending behaviors to better predict the product needs and design marketing efforts. Tracking spending patterns is also beneficial because when customers have not used your products for a period of time, they are more likely to churn.

One of the most known segmentation techniques nowadays is the recency, frequency, and monetary (RFM) method. The RFM technique was first introduced by Hughes (1994) and has since become a widespread and critical tool for marketers to measure their customers' relations. Research using RFM segmentation has been done in various industries such as electronic (Cheng & Chen, 2009), online retail (Chen et al., 2012), and hospitality (Dursun & Caber, 2016).

Other researchers have attempted to conduct studies by including new variables or excluding parts of the RFM corresponding to the product or service's nature. Chang and Tsay (2004) created the LRFM (Length, Recency, Frequency, and Monetary) model by adding length into the feature as a measure of customer loyalty. In their study about the Turkish grocery retail store, Peker et al. (2017) proposed the addition of periodicity and slight alteration of the recency variable resulting in LRFMP (Length, Recency, Frequency, Monetary, and Periodicity) segmentation using k-means. In their paper, recency is computed as the average difference in days between the end date of the observation period and the customer's last *n* visits. Periodicity is defined as the regularity of a customer's visit and is computed using the standard deviation of the number of days in between the customer's consecutive visits. Mo et al. (2010) conducted a multi-region segmentation using k-means with 6 input variables namely the customers' monetary amount and frequency of transaction along with 4 customer loyalty attributes. Between the 4 segments found, each attribute showed significance below 1%-level, indicating that the segments are significantly different from one another.

2.2. Customer churning

Churning in this research is defined as the act of customers voluntarily ending business with certain entities. Nevertheless, different definitions of churning are used in the following works of literature discussed. Literature that uses another definition of churning would be explicitly stated.

Kumar and Ravi (2008) predicted churning for a Latin American bank where the variables consist of the customers' sociodemographic and transaction information. The methods used include multilayer perceptron, radial basis function, logistic regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM). Four criteria are applied to choose the best model, namely sensitivity, specificity, accuracy, and Area Under Curve (AUC). RF outperforms all the other models in all four criteria.

Some researchers have also tried to find which variables are most informative of customers' churning likelihood. Lin et al. (2011) conducted a study using rough set theory (RST) to predict churning credit card customers in a Taiwanese bank. In their study, churning is divided into voluntary and involuntary; voluntary churning means that the customers end relationships with the bank while involuntary churning happens when the banks end relationships with the customers due to unpaid debt, failure of payment, or fraudulent activities. Using the number of supports in flow network graphs, the most important patterns in the data are that males, married, high annual pay-off time, and high purchase amount customers tend not to churn and vice versa. Additionally, customers aged between 30-39 are more likely to fall in the voluntary churning group while involuntary churning customers tend to show no consumption increase in the past 6 months. Therefore, the researchers suggested using both demographic and transaction information to predict and classify customer churning.

Nie et al. (2011) used LR and DT to predict churning for credit card customers in a Chinese bank. The definition of churning used here is when customers did not do any transaction within 12 months. The best LR model contains the variables describing customer, basic card, transaction, and risk information. The result shows that larger intervals between transactions, larger intervals between the last transaction date and the beginning of the research date, and higher ratio of debit transactions done in point of sale to all channels increase the odds of customers churning. A similar DT model with the same variables as the LR reveals the most important variables to be the last transaction amount, ratio of debit transactions done via the internet to all channels, and interval between the card issue date and beginning of the research date. When the three aforementioned variables are higher, the less probable the customers churn. These findings are in line with the reality faced by financial institutions. The marketing department from the bank could then focus on internet banking as customers frequently using the internet to conduct transactions are more likely to be loyal, similar to long-standing

customers. The comparison of the best performing LR and DT model indicates that the LR has lower average error and misclassification costs.

2.3. Combination of customer segmentation and churning

Rajamohamed and Manokaran (2017) used the combination of both unsupervised and supervised learning techniques to cluster and classify the credit card customers of a Taiwanese bank. For clustering, they tried 3 methods namely k-means, fuzzy c-means, and rough k-means. Meanwhile, the methods tried for classification include k-nearest neighbor (KNN), naïve bayes (NB), DT, RF, and SVM. Based on the sum of squared errors (SSE), the best algorithm for clustering is the rough k-means while for classification, SVM has the highest accuracy. When both clustering and classification are combined, the hybrid classifier with rough k-means and SVM performed the highest accuracy with the lowest misclassification errors.

Hung et al. (2006) did a study using customers' transaction data in a Taiwanese wireless telecommunication company to predict the probability and identify causes of churning. Aside from demographics, variables denoting customers' bills and payments, call detail records, and interactions with customer service are included in the model. Supervised learning using decision trees with and without segmentation was done; subsequently, a T-test was performed to compare the decision tree models with the backpropagation neural network. Higher hit and capture ratios were obtained using the decision tree without segmentation; however, the backpropagation neural network showed even better performance metrics than the DT without segmentation. The researchers noted that the highly unbalanced proportion of the churn vs. non-churn customers might be driving the segmented decision tree's poor predictive performance, as each segment does not contain a sufficient number of churned customers.

Bose and Chen (2009) also did a two-stage study consisting of unsupervised clustering and supervised churn prediction for a mobile service provider. The clustering techniques used include k-means, k-medoids, Self-Organizing Map (SOM), Fuzzy C-Means (FCM), and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). The variables relating to the customers' mobile service usage and revenue generation are used for clustering. The chosen churn prediction method is decision tree with boosting. Their results showed that including cluster labels as a variable in the churn prediction improved the top decile lift of the model; hence, clustering is useful to detect patterns that would enhance churning detection. The researchers suggested for marketers to use the hybrid of SOM and DT with boosting for predicting churn at an immediate or nearby point in time since it performs best with the current data; when using future data for decision making that comes at a later stage, BIRCH

provides better performance. K-means and DT with boosting prevail as the best combination when the decision-making time frame is not so clear for the marketers.

2.4. Hypotheses development

In the previous subsections, existing results and methods related to customer segmentation, churning, and the combination of both are discussed. Based on these aforementioned literature, different hypotheses would be elaborated in this subsection.

Since customer segmentation is an unsupervised learning, there is no hard and definitive rule to define what segments should be created or how many segments should be retained. The main objective of conducting segmentation is that it should be beneficial for the business. Previous research has segmented credit card customers based on profitability and credit standing through the customers' income, transaction amount, and credit scores (Li et al., 2010) or their transaction frequency, amount, and satisfaction (Mo et al., 2010). However, the bank's relationship with customers and the customers' inactivity are important features to estimate churning. To the best of the author's knowledge, these variables denoting the length, depth, or inactivity in the relationship have not been included in past research. Hence, the segments created would be assessed based on 2 factors, namely the different levels of revenues received and relations. Thus, the first hypothesis is:

H1: There exist segments within the customers that are characterized by revenue generation and relations.

Further elaborating on the characteristics, revenues from the credit card holders could be observed through their consumption, specifically their transaction information (e.g. frequency, amount). When consumption increases, the bank receives higher commission charges and revenues. Customer relations deal with the degree of loyalty and customer satisfaction shown through the number of products owned, relationship duration, and inactivity. The most apparent evidence of customer satisfaction is whether a customer churns or not. This leads to the second hypothesis:

H2: Segments with higher proportions of churning tend to have lower frequency and monetary transactions, fewer products owned, shorter relationship length, and higher inactivity.

Retention efforts should be executed for the segments with higher churning proportions as a 5% increase in customer retention is proven to reduce a bank's operating expenses by 18% (Karakostas et al., 2005). Although various research in predicting customer churn has been conducted, not many have provided an explanation on which variables are important in explaining the behaviors or characteristics of customers who are likely to churn. A study by Lin et al. (2014) demonstrated that performing clustering prior to prediction results in better performance. Therefore, after the segmentation is done,

the next step would be to predict churning and find out which variables are indicative of a customer's churning likelihood in the different customer segments so that customized anti-churn strategies could be prepared. This results in the third hypothesis:

H3: The segments are characterized by different variables of importance for churning prediction

3. Data

3.1. Data description

The credit card customers dataset is from Analyttica TreasureHunt (ATH) Leaps, a platform for individuals to learn, apply, and solve data science and machine learning cases with real-life datasets. The dataset is available on Kaggle[1] and contains 10,127 credit card accounts with 21 variables ranging from demographics, account, to transaction details. The variables descriptions are in Table 1 below.

Table 1. Variable names, types, and description for the credit card dataset.

Variable	**Type**	**Description**
Clientnum	Numerical	Unique identifier for the customer holding the account
Attrition_Flag	Character	"Attrited Customer" if the customer churns and "Existing Customer" if the customer stays
Customer_Age	Numerical	Customer's age in years
Gender	Character	M=Male, F=Female
Dependent_count	Numerical	Number of dependents
Education_Level	Character	Educational qualification of the account holder with 7 levels (e.g. uneducated, high school, college, graduate, post-graduate, doctorate, and Unknown)
Marital_Status	Character	Single, married, divorced, Unknown
Income_Category	Character	Annual income category of the account holder (< $40K, $40K - $60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category	Character	Type of card (Blue, Silver, Gold, Platinum)
Months_on_book	Numerical	Months on book (time of relationship)

[1] Link to the dataset: https://www.kaggle.com/sakshigoyal7/credit-card-customers

Total_Relationship_Count	Numerical	Total number of products held by the customer
Months_Inactive_12_mon	Numerical	Number of months inactive in the last 12 months
Contacts_Count_12_mon	Numerical	Number of contacts in the last 12 months
Credit_Limit	Numerical	Credit limit on the credit card
Total_Revolving_Bal	Numerical	Total revolving balance on the credit card
Avg_Open_To_Buy	Numerical	Open to buy credit line (average of the last 12 months)
Total_Amt_Chng_Q4_Q1	Numerical	Change in transaction amount (Q4 over Q1)
Total_Trans_Amt	Numerical	Total transaction amount (last 12 months)
Total_Trans_Ct	Numerical	Total transaction count (last 12 months)
Total_Ct_Chng_Q4_Q1	Numerical	Change in transaction count (Q4 over Q1)
Avg_Utilization_Ratio	Numerical	Average card utilization ratio

Note. Adapted from "Predict Customer Attrition Using Naïve Bayes Classification" by Analyttica TreasureHunt (ATH) Leaps, n.d. (https://leaps.analyttica.com/sample_cases/11). Copyright n.d. by Analyttica TreasureHunt (ATH) Leaps

As *Clientnum* is a unique ID for each customer and no two rows in the data set share the same ID, this variable would be removed prior to analysis. The variables *Attrition_Flag* and *Gender* are transformed into dummy variables called *churn* and *female* to make them easier to interpret in the results section. The dummy variables, *Education_Level, Marital_Status*, *Income_Category,* and *Card_Category* are then turned into factor variables to categorize the data and ensure that the statistical function would treat the data correctly. Three demographic variables contain missing values which are indicated with the category "Unknown" in their answers, namely *Education_Level, Marital_Status*, and *Income_Category*. The proportions of missing values are 15%, 7.4%, and 11% for the three aforementioned variables respectively. As the 3 variables with missing values are categorical and have more than 2 levels, the imputation is done using multinomial logit.

After some data pre-processing, the descriptions of variables used to conduct the segmentation are briefly discussed below.

Inactivity. This variable informs how active the customer-business interaction is. In previous research, the activity level of a customer is measured by recency which is the time interval (typically the number of days) between the observation period and the customer's last usage date. However, as information about the last date of usage is not available in the dataset, recency would be substituted by the number of months where the customer has been inactive in the past 12 months (*Months_Inactive_12_mon*).

Length. Length is typically defined as the duration of the relationship between the business and the customer. Therefore, the variable *Months_on_book* would be used as a feature of length.

Relations. This feature is reflected in the number of the bank's products that the customers possess (*Total_Relationship_Count*). Customers with higher number of relations counts are likely to be more loyal and satisfied with the entity. Therefore, segments with higher relations count are expected to have lower proportions of churning customers.

Frequency. It is expected that customers who often use credit cards would have higher loyalty and lower probabilities of churning. Thus, frequency refers to the number of transactions done with the credit card in the last 12 months (*Total_Trans_Ct*).

Monetary. The amount of monetary transactions paid using the credit card in the past 12 months (*Total_Trans_Amt*) is the measure of monetary used. This variable would be used as an indicator of the customer's contribution to the store's revenues.

3.2. Descriptive statistics

The dataset contains an unbalanced proportion of churning and non-churning customers with 16% and 84% respectively. The descriptive statistics in Figure 2 below shows that non-churning customers tend to have a higher average of *Total_Relationship_Count*; this means that non-churning customers own more products from the bank than churning ones. It is interesting to note that there is no noticeable difference seen in *Months_on_book* for churning vs. non-churning customers. Revolving balance is the amount of credit that is unpaid at the end of the billing period and is taken to the next one. The average *Total_Revolving_Bal* for non-churning is nearly twice as high as that of churning clients. Both the averages and ranges for transaction amount and count (*Total_Trans_Amt* and *Total_Trans_Ct* respectively) of non-churning customers are higher. This is as expected since customers who frequently use credit cards are more likely to believe that they derive value from the ownership of those cards and become unlikely to churn (Lin et al., 2011). Non-churning customers also have a lower average of *Months_Inactive_12_mon* and a higher average of *Avg_Utilization_Ratio*.

Figure 2. Comparison of the descriptive statistics between non-churning (churn = 0) and churning (churn = 1) customers for numerical variables.

The descriptive statistics for categorical variables could be seen in Table 2 below. Both groups do not show much difference in terms of education level, marital status, or income level. The categories which often appear are those that hold graduate degrees, are married, and earn less than $40K in a year. Remarkably, although graduate degree holders are the category with the most observations in this dataset, the income level category of less than $40K does not seem to reflect this. Most of the card category seems to be heavily dominated by blue card holders in both groups, as the total percentage adds to more than 93% of the observations.

Table 2. Comparison of the descriptive statistics between non-churning and churning customers for categorical variables.

		Non-churn		**Churn**	
Variable	**Categories**	**Frequency**	**Percent (%)**	**Frequency**	**Percent (%)**
Education_Level	Uneducated	1,654	16.33	310	3.06
	High School	1,953	19.29	347	3.43
	College	999	9.86	181	1.79
	Graduate	3,001	29.63	568	5.61
	Post-Graduate	491	4.85	110	1.09
	Doctorate	402	3.97	111	1.10

Marital_Status	Single	3,632	35.86	743	7.34
	Married	4,202	41.49	751	7.42
	Divorced	666	6.58	133	1.31
Income_Level	Less than $40K	3,599	35.54	756	7.47
	$40K - $60K	1,706	16.85	304	3.00
	$60K - $80K	1,241	12.25	191	1.89
	$80K - $120K	1,323	13.06	246	2.43
	$120K +	631	6.23	130	1.28
Card_Category	Blue	7,917	78.18	1,519	15.00
	Silver	473	4.67	82	0.81
	Gold	95	0.94	21	0.21
	Platinum	15	0.15	5	0.05

3.3. Data splitting

After the segmentation is done, the data set is split based on the observations' cluster members. For the churning prediction, each cluster set would be further divided into 80% training and 20% testing sets. Due to the unbalanced proportion of churning and non-churning samples, stratified sampling would be used to ensure that the training and testing sets contain similar proportions of churning samples as the original cluster sets. The stratified sampling is done using the function *partition()* from the *splitTools* package in R.

4. Methodology

After data cleaning, imputation, and transformation have been done, the next step is to do the clustering and churning prediction. To answer the main research question, customer segmentation using the ILRFM (Inactivity, Length, Relations, Frequency, and Monetary) with k-means and rough k-means clustering would be conducted. Segmentation with all variables available is also attempted; as some variables in the dataset are categorical and numerical, k-medoids clustering is used. Then, the data set would be divided into the number of clusters chosen, where each set contains only members of that specific cluster. Afterward, variables that are most indicative of churning likelihood in the different segments would be examined through logistic regression and random forest. The quality of the LR and RF would be measured using several evaluation metrics. The conceptual framework of this research could be seen in Figure 3 below.

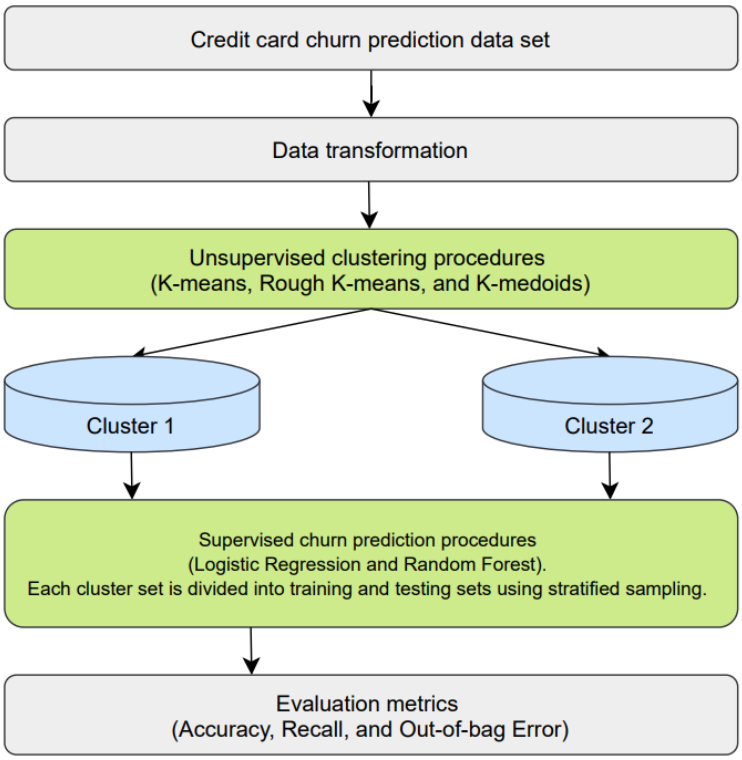

Figure 3. Conceptual framework of the research using clustering and churning prediction procedures.

4.1. Customer segmentation methods

Clustering is an unsupervised learning problem as it aims to find patterns in the data. The data is then divided into homogenous groups, where observations within a group are more similar to one another while those between groups are more distinct.

4.1.1. K-means clustering

K-means clustering is one of the most widely used clustering methods in machine learning. Suppose there are n observations in the dataset represented with the vector x_i where $i = 1, 2, \dots, n$. Then the number of K clusters is specified, where $K \leq n$ as each cluster needs to contain at least 1 observation. The clusters are denoted as $C_1, C_2, C_3, \dots, C_K$. Afterward, K data points are randomly selected as the cluster centroids, $c_1, c_2, c_3, \dots, c_K$. The next steps are repeated until a local optimum is found and the cluster assignments do not change: assign each observation into a cluster based on its distance to the centroid and re-compute the centroid by calculating the average of all data points in that cluster. The clustering mechanism of k-means aims to satisfy the following properties: each observation belongs to one cluster, the clusters are non-overlapping, and within-cluster variations are as small as possible. To define the within-cluster variation, Euclidean distance is used. The equation for k-means to minimize the sum of within-cluster variations for all K clusters look as follows:

$$J = \min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \right\}$$

In each k cluster, the within-cluster variation is computed as the Euclidean distance between all observations and the cluster centroids, divided by $|C_k|$ which is the number of observations in the kth cluster. Most importantly for k-means, when the Euclidean distance function is used, the values of features included should be numerical and on the same scale; data transformations such as normalization or standardization need to be done prior to the analysis.

4.1.2. Rough k-means clustering

Rough sets are used when there is insufficient information to precisely define each observation into at most one cluster. Rough k-means clustering has the following properties: each cluster is represented by a lower ($\underline{B}C_k$) and upper ($\overline{B}C_k$) approximation, the lower approximation of each cluster has to contain at least 1 observation in the beginning, if an observation belongs to a cluster's lower approximation then it also becomes a member of the cluster's upper approximation, and an observation belongs to more than 1 upper approximations if it does not belong to any lower approximation. The number of objects in the rough boundary area is denoted by $\left|\overline{B}C_k - \underline{B}C_k\right|$ while the weights of the lower and upper approximations are w_l and $w_u = 1 - w_l$ respectively.

Similar to regular k-means, the number of K clusters needs to be specified as a first step, where $K \leq n$. The second step is where the observations are randomly assigned to K clusters. The third step is to compute the cluster centroids with the following formula:

$$c_k = \left(w_l \times \frac{\sum_{x_n \in \underline{B}C_k} x_n}{\left|\underline{B}C_k\right|} \right) + \left(w_u \times \frac{\sum_{x_n \in \overline{B}C_k} x_n}{\left|\overline{B}C_k\right|} \right)$$

Where the first bracket shows the lower approximation's sum of values divided by the number of observations in the lower approximation (denoted with $\left|\underline{B}C_k\right|$), multiplied by the weight; the same applies for the second bracket but for the upper approximation. Then, the Euclidean distances between the observations and each cluster centroids are computed as the fourth step. To determine if an observation belongs to a lower or upper approximation of a certain cluster in the fifth step, the formula below is used:

$$T = \left\{ k : \frac{d(x_i, c_k)}{d(x_i, c_h)} \leq threshold \text{ and } h \neq k \right\}$$

The notation $d(x_i, c_h)$ denotes the distance between observation x_i to the cluster center c_h, where c_h is the closest cluster center for observation x_i compared to all other cluster centers. The default

threshold in R is 1.5, which is also the threshold used for this analysis. If T turns out to be an empty set (or in other words, $T = \emptyset$), then x_i belongs to the lower approximation of cluster h. Otherwise, if $T \neq \emptyset$, it means that x_i is located in proximity to at least one other cluster center c_k in addition to c_h. Thus, x_i belongs to the upper approximations of both clusters h and k, where $k \in T$. Then the algorithm repeats the third to fifth steps until there is convergence, which is when cluster centers and assignments remain stable. Thus, the optimization problem for rough k-means aims to minimize the observations' weighted sum of Euclidean distances in both the lower and upper approximations to their cluster centroids; the equation for this is shown below:

$$J = \min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \left(\frac{w_l}{|\underline{B}C_k|} \sum_{x_i \in \underline{B}C_k} \|x_i - c_k\|^2 + \frac{w_u}{|\overline{B}C_k|} \sum_{x_i \in \overline{B}C_k} \|x_i - c_k\|^2 \right) \right\}$$

4.1.3. K-medoids clustering

K-medoids has a similar objective and algorithm as k-means. The difference is that the cluster centers in k-medoids are actual points (known as medoids) as opposed to the average of the observations within clusters in k-means. K-medoids has the advantage of being more resilient to outliers or noise in the data because it minimizes the sum of pairwise dissimilarities whereas k-means typically use the sum of squared Euclidean distances (Arora et al., 2016).

As both numerical and categorical variables are present in the dataset, Gower distance is the dissimilarity measure used. Gower is versatile since it could handle categorical, numerical, logical, and character variables. Gower coefficient (denoted as $d(i,j)$) has a range of [0,1] indicating the minimum and maximum distance respectively; it is computed with the following formula:

$$d(i,j) = \frac{1}{p} \sum_{i=1}^{p} d_{ij}^{(f)}$$

where $d_{ij}^{(f)}$ is the partial dissimilarity between observations x_i and x_j for variable f and p is the number of variables in the data. In the case of numerical variables, $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$ where x_{if} is the value of observation i for variable f while R_f is the range of values for variable f. When the variables are categorical, $d_{ij}^{(f)} = 0$ if x_i and x_j are in the same category and $d_{ij}^{(f)} = 1$ if the two observations are in different categories.

4.1.4. Methods to determine the number of clusters

Several methods are available to assist in determining the optimal number of clusters. This section would elaborate more on the silhouette method. However, it is important to keep in mind that this

method is a not strict performance requirement and should not be taken as such. In many applications, the number of clusters is chosen based on subject-matter knowledge, business requirements, and research motivation. Without any previous knowledge of the industry or business, the rule of thumb suggests setting K equal to $\sqrt{n/2}$. This is problematic when there is a substantial pool of observations as the number of clusters could become unmanageably large (Lantz, 2013).

The silhouette approach defines how well an object sits in its cluster. It is measured through two parts, namely cohesion (a_i) and separation (b_i). Cohesion is computed by the average distance of observation i with all other observations in the same cluster; the lower a_i is, the more it shows that the assignment of observation i to its cluster is suitable. The formula for cohesion is shown below:

$$a_i = \frac{1}{|C_i| - 1} \sum_{i,j \in C_i \text{ and } i \neq j} d(i,j)$$

where $|C_i|$ is the number of observations falling into the cluster C_i and $d(i,j)$ is the distance between observation i and j.

Separation is determined by the minimum average distance between i and all points that sit in any other cluster. It is defined in the formula below:

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

where $|C_k|$ is the number of observations falling into the cluster C_k and $d(i,j)$ is the distance between observation $i \in C_i$ and $j \in C_k$. As seen in the formula above, to compute separation, the assumption that there exists another cluster apart from C_i and $K > 1$ has to hold.

Therefore, the formula for silhouette value for observation i is:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

The average silhouette ranges from -1 to +1, where a value close to +1 shows a good clustering and high certainty that the observations have been assigned to the correct cluster. Meanwhile, a value close to -1 indicates that the observations are assigned to the wrong clusters and a value of 0 indicates that the object lies within proximity to the neighboring cluster. Thus, the chosen number of K should be one where the average silhouette is highest.

For clusters that have only a single member, it is unclear how to define the cohesion or the distance between an object with itself in the same cluster; therefore, the silhouette coefficient here is undefined. Similarly, the silhouette score when $K = 1$ is also undefined as there is no other cluster to

compute for the separation. In the original paper for silhouette, Rousseeuw (1987) suggests setting the silhouette to 0 in such cases, as it is the most neutral value between a range of -1 to +1. This is reasonable as it is hard to define whether an observation is clustered "well" or "poorly" when clusters only consist of a single member or when there is only 1 cluster.

4.2. Churning prediction methods

Logistic regression (LR) is a widely-used statistical model that has been applied to various industries. It has also shown relatively good performance despite its simplicity and training efficiency (Nie et al., 2011). The coefficients provide measures on both the direction and size of the relationship; those with statistically significant coefficients could be used as indicators of variable importance. Numerous research has shown random forest (RF) to be one of the best-performing methods for churn prediction (Buckinx & Van den Poel, 2005; Kumar & Ravi, 2008; van Wezel & Potharst, 2007); some even achieved an accuracy of 90% or higher (Sinha & Huraimel, 2021). RF is often used as it typically handles large data with less computational time than other methods and gives an overview of which variables are important predictors along with an internal estimation of the error rate from the out-of-bag (OOB) samples.

4.2.1. Logistic regression

Binomial logistic regression, or simply referred to as logistic regression, is a statistical model that is often used for prediction analysis when the dependent variable is binary (e.g. churn or no churn, win or lose). The outcome is a probability that falls between the range [0,1] and is done with the logistic function below:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

where $p(X) = pr(Y = 1|X)$, p is the total number of variables in the data, β_0 is the intercept, and β_p is the coefficient for the pth variable. In order to have an equation that is linear to X, some transformations need to be done to the equation above. First, the odds are computed with the following equation:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$$

where $\frac{p(X)}{1-p(X)}$ denotes the odds, which is the probability of an event occurring divided by the probability that an event does not occur. Odds could range from 0 to ∞, showing a very low or very high probability of the event occurring respectively. Taking the logarithms on both sides would result in the equation below:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The left-hand side denotes the log odds, which is also called logit. Therefore, the logistic regression has a logit that is linear to X. When the coefficients are positive, then it shows that the variable has a positive effect in increasing the odds of the event happening. For example, the effect of a unit change in X_p is equal to $\left(e^{\beta_p} - 1\right) * 100\%$ change in probability towards $Y = 1$.

However, in order to have a valid interpretation of the LR results, the data set needs to be verified for several assumptions. The first one is to check for linearity between each numerical predictor variable and the logit-transformed response variable. As shown in Figure 4 below, the scatter plots seem to show linear associations between all predictor values and the logit-transformed churn variable.

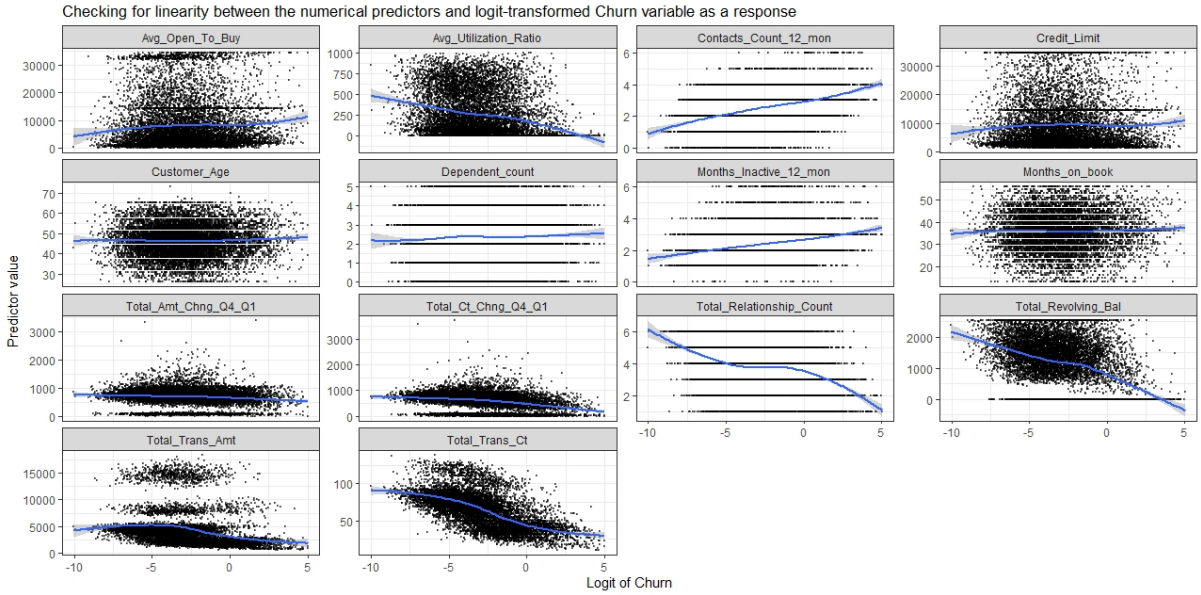

Figure 4. The plot to check for linearity assumption between each numerical independent variable and the logit-transformed dependent variable.

The second assumption is to check for high correlations or multicollinearity between the independent variables. High correlations between the predictors would usually cause the beta coefficients of these variables to have large standard errors. In R, this is done with the *vif()* function in the *car* package. The rule of thumb is to remove the variables with high collinearity, which is those with Variance Inflation Factor (VIF) values larger than 10 (Stoltzfus, 2011). There seems to be no major signs of concern except for the high collinearity between the variables *Credit_Limit* and *Avg_Open_To_Buy* as their VIF values are well above 10 (see Table 3 below). This is intuitive as the average open to buy is the average of credit left available for the credit card user, which is computed by deducting the present credit balance from the credit limit. Therefore, the *Avg_Open_To_Buy* is removed from the logistic regression

estimation. After the removal of *Avg_Open_To_Buy*, the variable *Credit_Limit* no longer shows a VIF value above 10.

Table 3. Variance Inflation Factor (VIF) values for the independent variables.

Variable	VIF
Avg_Open_To_Buy	101.76
Credit_Limit	95.33
Total_Trans_Ct	4.37
Income_Category	4.25
Total_Trans_Amt	4.12
Female	3.35
Customer_Age	2.78
Months_on_book	2.77
Avg_Utilization_Ratio	2.47
Total_Revolving_Bal	2.46
Card_Category	1.50
Total_Relationship_Count	1.18
Marital_Status	1.08
Dependent_count	1.05
Months_Inactive_12_mon	1.05
Education_Level	1.04
Contacts_Count_12_mon	1.04
Total_Ct_Chng_Q4_Q1	1.04
Total_Amt_Chng_Q4_Q1	1.03

4.2.2. Random forest

Random forest is a supervised machine learning that consists of an ensemble of decision trees. The class that has the highest majority vote (classification) or the average prediction (regression) is the output given from the individual trees. An unpruned DT that is grown to its full depth becomes too complex and would risk overfitting, as it memorizes the noise and is very sensitive to changes in the training data, overlooks important patterns, and does not generalize well to unseen data sets.

RF typically performs better than DT as it combines multiple relatively uncorrelated trees, which would lead to a reduction in overfitting and improvement of accuracy. It does this through a slight modification of bootstrap aggregating (commonly referred to as bagging). The first step is to create a

bootstrapped dataset where samples are randomly selected with replacement, so the samples could be picked more than once. This allows each tree in RF to be trained on different sets of samples; hence, it is less sensitive to the training data. The bootstrapped dataset created has the same size as the original one.

In the second step, a DT is created where a random sample of m out of the total p number of variables is considered during each splitting process in the tree. This means that different random selection of variables are used at each split point. This step assists in reducing the influence of dominating variables, correlations of predictions from the trees, and variance. For classification problems, the default settings and rules of thumb are to take $\sqrt{p}$ as the number of variables randomly selected and use Gini impurity index as a splitting criterion to choose a variable for splitting that would result in the lowest impurity or highest proportion of correctly classified observations. A Gini impurity index ranges between 0 and 1 and is computed with: $1 - \sum_{q=1}^{Q}(p_q)^2$, where p_q is the proportion of observations falling into class q and Q is the total number of classes. A Gini index of 0 signifies a pure terminal or leaf node, where all observations are predicted into the same class. A Gini of 1 means that all observations are randomly distributed while 0.5 means that the observations are equally distributed into the classes available. Each tree is grown until the terminal nodes contain less than the minimum number of samples specified, denoted as *nodesize*; the default *nodesize* for a classification RF is 1 sample. The 2 aforementioned steps are repeated until the number of trees to be made, denoted as M, is fulfilled. As the last step, the prediction from all of the trees created is aggregated. The class with the highest vote becomes the final prediction.

The rule of thumb for the number of trees is to start with 10 times the total number of features (Boehmke & Greenwell, 2019). Probst and Boulesteix (2017) stated that tuning the number of trees for classification RF is not recommended as it usually results in a very small and negligible gain in the performance metric. Therefore, this paper would use the default number of trees in RF packages, which is set to 500.

According to Probst et al. (2019), there are three influential hyperparameters in RF that would need to be tuned with grid search, namely the number of variables randomly selected for splitting (*mtry*), the number of samples to be drawn for each tree (*sampsize*), and the minimum number of samples in each leaf/terminal node (*nodesize*). Tuning other hyperparameters outside of these three would not bring substantial gain in terms of performance; thus, other hyperparameters would be left in their default values.

In their research, Probst et al. (2019) found that tuning *mtry* results in the highest average improvement for the model's performance. For *mtry*, there is a tradeoff between stability and

accuracy. Stability is better achieved when the value for *mtry* is lower as it helps the creation of uncorrelated trees. It also helps unmask variables with moderate effects that would have otherwise been overshadowed by those with stronger effects. On the other hand, this could cause suboptimal variables to be chosen as splitting candidates which leads to trees with lower accuracy.

The tuning of *sampsize* also presents the same compromise between accuracy and stability. Lowering the number of samples drawn for training each tree would break the correlations between trees and allow for the creation of more dissimilar trees. The aggregation of these trees could then result in better prediction accuracy. On the other hand, reducing the number of observations used for training would result in lower accuracy of the individual trees. When Martínez-Muñoz and Suárez (2010) conducted a study about the effect of sample size on the model's performance, they found that the optimal values for sample size depend on the problem on hand and could be estimated with out-of-bag error. The result of their study also shows that using a smaller sampling size than the default, which is to sample all observations in the data set with replacement, usually leads to both a better performance in most data sets and a lower computational time.

Similarly, *nodesize* also has a tradeoff between accuracy and computational time. Lower values produce trees with more depth, indicating that more splits are done until the tree reaches the terminal nodes; this leads to higher accuracy. Although *nodesize* has a default value of 1 for classification, it is recommended to set it higher as it exponentially decreases computational time without a major loss in prediction performance, especially in large data sets.

To find the variable importance in RF, Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) are used. MDA shows the decrease in OOB accuracy when other variables are kept constant except for one certain variable, which is randomly permuted or shuffled without replacement. The random permutation imitates the omission of a certain predictor from the model. Higher MDA means that the variable is important as the permutation of the variable results in a high decrease of accuracy. For M number of trees, the formula of MDA is as follows:

$$Mean\ Decrease\ Accuracy\ (x_p) = \frac{1}{M}\sum_{t=1}^{M}\sum_{i\in OOB}\frac{I(y_i = f(x_i)) - I(y_i = f(x_i^p))}{|OOB|}$$

where the importance of variable x_p in tree t is the sum of the difference in class predictions between the original (x_i) and permuted (x_i^p) observation i in variable p, divided over the number of OOB samples. Then, this average is further divided by the M number of trees.

MDG gives the sum of the decrease in a node's impurity for all trees when variable p is used for splitting, divided by the number of trees, as seen with the formula below:

$$Mean\ Decrease\ Gini\ (x_p) = \frac{1}{M}\left[1 - \sum_{t=1}^{M} Gini\ (p)^t\right]$$

Similar to MDA, higher MDG means a variable is more important in contributing to purer terminal nodes. However, MDG suffers from bias when variables have different scales of measurement; it tends to prefer variables with larger ranges or more categories. Therefore, it might not show the most reliable result in indicating the importance of a variable. This problem is especially important when the model is used for variable selection instead of for prediction purposes only (Strobl et al., 2007).

Black box models acquire their names due to the fact that they are created directly by an algorithm that consists of such complicated functions, to the point that even those who design the algorithms could not understand how the variables are utilized and combined to create the predictions. Since the internal processes of RF during training could not be unveiled in a human-comprehensible form, RF falls into a "black box" supervised learning method (Rudin & Radin, 2019). To extract the results in an interpretable form, a model-agnostic method is needed. Therefore, Accumulated Local Effects (ALE) plot is used to visualize the average change in in-sample predictions of the target variable due to certain predictors. As seen on the VIF table (see Table 3), some predictors seem to be correlated. Compared to other similar model-agnostic methods such as Partial Dependence Plots (PDP), ALE is faster to implement and unbiased when predictors are correlated. Since ALE plot is centered at 0 on the y-axis, each point on the line could be interpreted as how the effect of a certain predictor results in the deviation from the mean prediction. The x-axis indicates the range of the predictor variables. The ALE formula for feature x_p is shown below:

$$\widehat{ALE}\ (x_p) = \left[\sum_{k=1}^{k(x_p)} \frac{1}{n(k)} \sum_{i:x_p^{(i)} \in N(k)} \left[f\left(z_{k,p}, x_{\setminus p}^{(i)}\right) - f\left(z_{k-1,p}, x_{\setminus p}^{(i)}\right)\right]\right] - c_p$$

The index for the interval $N(k)$ where x_p falls into is shown with $k(x_p)$. The total number of observations in the interval $N(k)$ is denoted $n(k)$. The algorithm starts by estimating the local effect through partitioning the range for x_p into grids or intervals with similar sample sizes. Then, the feature's observations that fall into the interval $N(k)$ would be replaced by the minimum ($z_{k-1,p}$) and maximum value ($z_{k,p}$) from the grid. This could also be seen through the notation $f\left(z_{k,p}, x_{\setminus p}^{(i)}\right)$ which displays that the value for observation i for x_p is replaced by the maximum or rightmost value of the interval (vice versa with $z_{k-1,p}$ as the minimum or leftmost value). The difference in in-sample predictions is summed up for all points within the interval and then divided by $n(k)$. The uncentered ALE is achieved through the equation inside the big square bracket. It is subtracted by c_p, which is a

constant denoting the average prediction, to arrive at the centered ALE where the average effect is zero.

4.2.3. Evaluation metrics for predictive classification methods

Accuracy, recall, precision, and F1 score would be the evaluation metrics used for the LR and RF algorithms. In order to describe the metrics used, an example of the confusion matrix (Figure 5 below) and the explanation of its terms would be first given.

		Actual	
		Not churn	**Churn**
Predicted	**Not churn**	True negative (TN)	False negative (FN)
	Churn	False positive (FP)	True positive (TP)

Figure 5. Confusion matrix for the credit card churn data set.

When the algorithm correctly predicts that a customer would fall in the churning category and the customer does churn in reality, then it is called a true positive (TP). On the other hand, true negative (TN) is when a customer is correctly predicted to be a retained customer. False positive (FP) happens when a customer is predicted to churn despite the reality that the customer does not churn. Lastly, a customer who is predicted to stay even though the customer actually falls into the churning category is called false negative (FN).

Accuracy is one of the most prevalent metrics to evaluate a machine learning algorithm. It is defined as the proportion of correct predictions over the total sample. It is calculated with the formula below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

However, predicting a customer will not churn when the customer actually churns (false negative) is likely to be costlier for the bank (Borbora et al., 2011). If the model has a high recall and a low false negative, the bank could create anti-churn strategies in advance to retain the customers. Thus, recall is also included as an evaluation metric. Recall is the proportion of correctly predicted churn cases (TP) divided by the total actual churn cases. The formula for recall is shown below:

$$Recall = \frac{TP}{TP + FN}$$

Precision is the proportion of correctly predicted churned customers out of the total customers that are predicted to churn. Precision is useful to minimize false positives, which would lead to a decrease in the unnecessary costs and efforts allocated to churning prevention for customers who are incorrectly predicted to fall in the churning category. It is computed with the formula below:

$$Precision = \frac{TP}{TP + FP}$$

Since the data set contains unbalanced proportions of churning and non-churning customers, F1 score might be a better metric than accuracy. It is the harmonic mean of precision and recall; hence, it is preferable when researchers would like to emphasize on reducing false positives and false negatives as much as possible. The formula for F1 score is:

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Random forest is equipped with an internal and unbiased measurement for the model error, which is called the out-of-bag (OOB) error rate. Since RF uses a bagging method, typically a third of the total samples are not taken during the construction of the trees; these observations would be called the OOB samples. The trees built without these OOB samples could be used to predict the classes of the OOB observations; the class with the most votes would be the final prediction. The proportion of OOB samples incorrectly classified by the RF is called the OOB error.

5. Results

5.1. Clustering

Figure 6 below shows that all clustering methods have their highest average silhouette when there are 2 clusters. When comparing the methods, it seems that rough k-means clustering generally performs best for 2 and 3 clusters. While for 4 clusters and more, regular k-means clustering performs better with a higher average silhouette than rough k-means or k-medoids. In ranking the average silhouette score, rough k-means clustering with 2 and 3 clusters come as first and second respectively; this is followed by k-means clustering with 7 clusters as third highest. As a higher silhouette means that the observations are well-placed into their clusters, the number of clusters with the highest score would be interpreted and used for churning prediction. In this case, the rough k-means with 2 clusters is the optimal number of clusters.

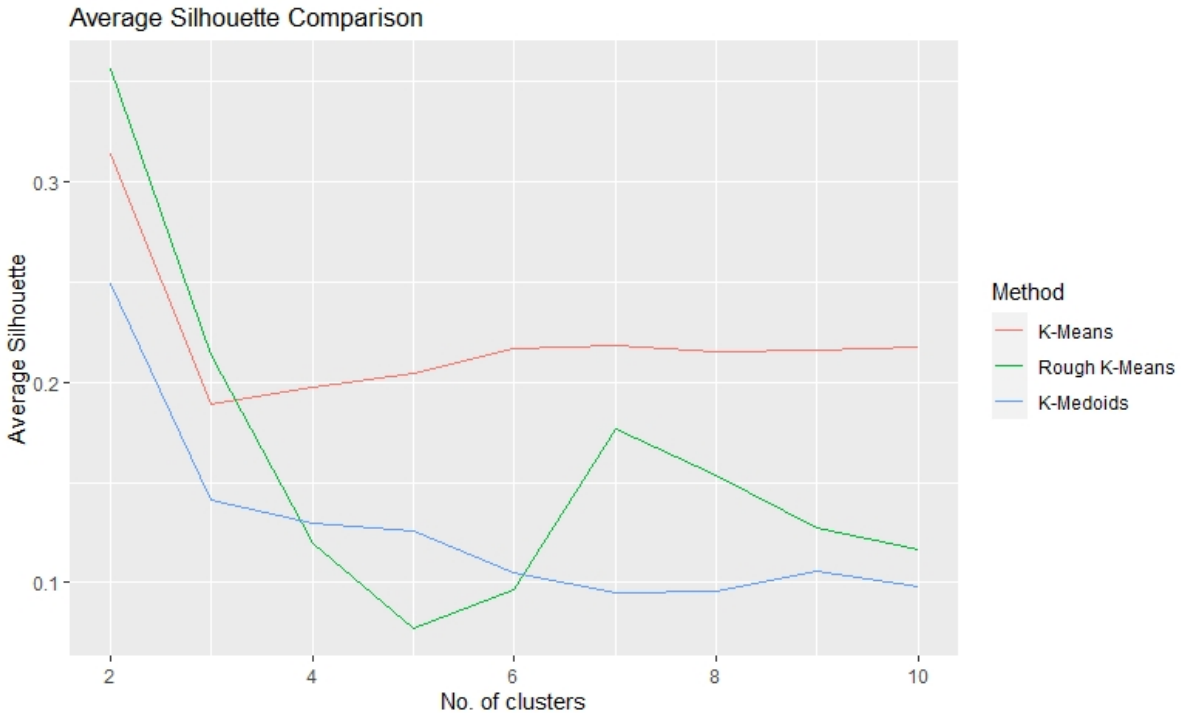

Figure 6. Average silhouette comparison between k-means, rough k-means, and k-medoids clustering.

As explained in the methodology section, rough k-means would place observations into lower and upper approximations. Observations that are in a cluster's lower approximation become a definite member of that cluster. There are also observations that do not fall into any lower approximation, so these samples could fall into the upper approximations of both clusters. Hence, as seen in Table 4 below, the total number of customers increased from 10,127 to 12,590 since some are present in both clusters. The resulting plot of rough k-means clustering could be seen in Figure 7 below. From both Table 4 and Figure 7, it could be clearly seen that cluster 2 (colored green) has more observations in the lower approximation than cluster 1 (colored red). The star-shaped observations which are colored black are those that fall in the boundary region or the upper approximations of both clusters.

Table 4. Number of customers in each cluster with rough k-means and 2 clusters.

	C1	C2	Total
Customers in lower approximation	1,153	6,511	7,664
Customers in upper approximation	2,463	2,463	4,926
Customers in each cluster	3,616	8,974	12,590

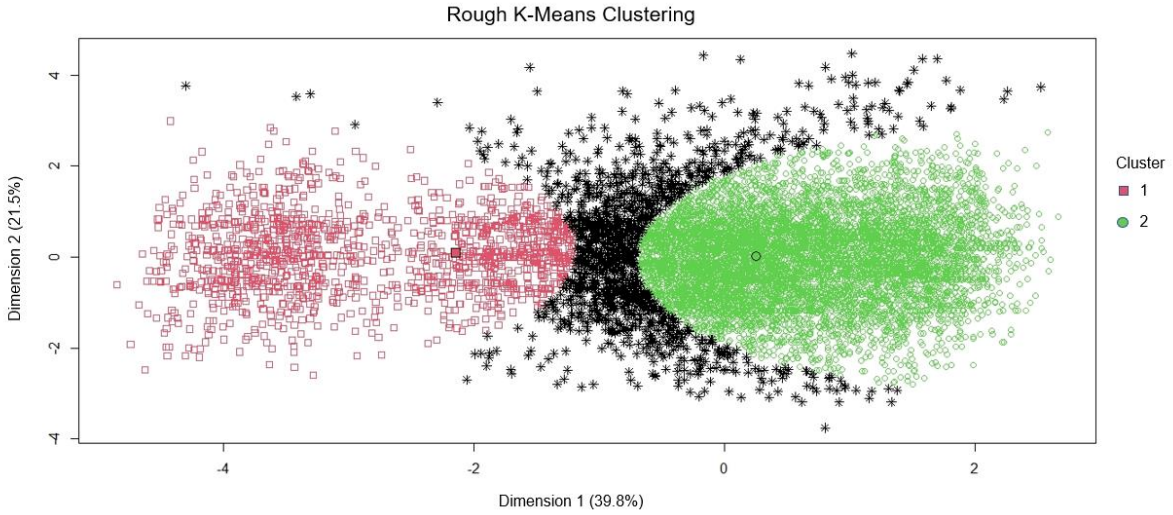

Figure 7. Visualization of the rough k-means clustering with 2 clusters. The black star-shaped observations are those that fall in the boundary region.

The averages of variables used in each cluster for rough k-means could be seen in Table 5 below (for a more detailed descriptive statistics of both clusters, see Figure A1 in Appendix). Both clusters depict similar values for the average length of relationship (*Months_on_book*) and inactivity level (*Months_Inactive_12_mon*). In cluster 1, the customers seem to have a lower average for the number of the bank's products owned (*Total_Relationship_Count*) but substantially higher average transaction amount (*Total_Trans_Amt*) and count (*Total_Trans_Ct*) than cluster 2. In fact, the average transaction amount in cluster 1 is more than twice higher than that in cluster 2; the minimum value of cluster 1 is notably not too far from the maximum value of cluster 2. Additionally, the proportion of churning customers in cluster 1 is more than 6 percentage points (pp) lower. Nonetheless, the maximum relationship count for cluster 2 is twice higher than that of cluster 1. Therefore, cluster 1 would be called "Spender" for customers who are actively using their cards for transactions and cluster 2 would be called "Passive" due to the lower rate of transactions done despite having a higher average relationship count with the bank.

Table 5. The mean of variables used for rough k-means with 2 clusters.

Variable	**Mean**	
	C1	**C2**
Months_on_book	35.63	36.00
Total_Relationship_Count	2.82	4.04
Months_Inactive_12_mon	2.39	2.37
Total_Trans_Amt	7,132.04	3,415.53

Total_Trans_Ct	84.01	60.05
Proportion of staying customers	88.69%	82.46%
Proportion of churning customers	11.31%	17.54%

5.2. Churning prediction

5.2.1. Logistic regression

The logistic regression analysis is done using the *glm()* function in *stat* package. The following section would discuss the results from logistic regression as seen in Table 6. Although all coefficients are shown in Table 6, only those which are statistically significant at 5%-level would be interpreted. The characteristics that have similar effects on both clusters, whether those are driving or reducing churning behaviors, would be discussed first. Then, the discussion for features that are distinctive to only either cluster 1 or 2 would follow.

As shown by the positive coefficients, the likelihood of churning increases when the dependent count is higher for both clusters. Variables that have similar effects of fueling churning likelihood in both clusters when their coefficients rise include the number of contacts the customers have with the bank in the past 12 months (*Contacts_Count_12_mon*) and total transaction amount (*Total_Transaction_Amt*). From the coefficients, increasing the number of contacts by 1 unit would increase the churning likelihood by 106.89% for cluster 1 and 67.87% for cluster 2. In both clusters, increasing the transaction amount by 1 unit would lead to a 0.1% increase in churning probability. Higher contact counts indicate that the customers are reaching out to the bank more frequently or vice versa. If the contacts are initiated by the customers, then these customers are presumably not happy with certain aspects of the products. Thus, it is reasonable that a higher number of contacts indicate unsatisfied customers who are likely to churn. This could indicate poor customer service, which is typically one of the biggest contributors to credit card churning (Krishnan, 2020; Sinha & Huraimel, 2021). However, the positive coefficient for transaction amount is not expected as previous research indicates that higher transaction amount is often a characteristic with a negative impact on the customers' intention to churn (Lin et al., 2011; Nie et al., 2011).

Customers whose *Marital_Status* are married have lower likelihood of churning by 39.89% and 43.90% for clusters 1 and 2 respectively, in comparison to those whose marital status are single. This is aligned with the findings of Lin et al. (2011). Other features such as having higher average for revolving balance (*Total_Revolving_Bal*), more relationships with the bank (*Total_Relationship_Count*), and higher frequency of transactions done using the credit card (*Total_Trans_Ct*) negatively affect customers' churning probability in both clusters. In particular, the latter two features seem to have quite large

impacts on churning. When the number of the bank's products owned by the customer increases by 1, then the probability of churning falls by 25.92% in cluster 1; in cluster 2, the effect is even more pronounced with 37.94%. Similarly, boosting the frequency of transactions done using the card by 1 would lower churning by 17.72% and 12.54% in the first and second clusters. It is logical that higher usage of the credit cards (as indicated by the higher *Total_Revolving_Bal* and *Total_Trans_Ct*) and deeper relationships with the bank (as indicated by the greater *Total_Relationship_Count*) would lead to lower churning.

The first cluster depicts an unexpected characteristic whereby customers who have longer relationship time with the bank have higher tendency to churn. This is shown by an increase of churning probability by 4.5% when customers increase their relationship duration by 1 month. The result here goes against previous papers which stated that longer customers tend to be more loyal, due to the lower likelihood of these customers being influenced by competitors' marketing campaigns (Colgate et al., 1996). A possible explanation for the unexpected result is the relatively smaller number of observations in cluster 1. For instance, although it is not statistically significant, the coefficient for *Months_on_book* in cluster 2 has the anticipated negative effect on churning.

In the second cluster specifically, several findings about the credit card usage details are aligned with previous research discussed. For instance, higher periods of inactivity increase churning, whereas an increase of transaction amount and count from the first to fourth quarter would decrease the customers' prospects of churning. Several demographical features are statistically significant in predicting an increase in churning probability such as being female, having a post-graduate education degree, earning an annual income above $120K, and having a silver credit card. Amongst all the demographic variables, being a female seems to have the most notable effect in churning possibility with an increase of 150.93%. Lin et al. (2011) had a similar result where male customers are less likely to churn. An interesting result is how increasing credit limit seems to have a negative effect in churning, while higher income categories indicate positive effects. The contrast is noteworthy as a higher credit limit is most likely linked to higher earning.

Table 6. Logistic regression results for clusters 1 and 2.

Dependent variable:	**Cluster 1**		**Cluster 2**	
Churn	**Number of obs: 2892**		**Number of obs: 7221**	
Independent variables	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Female	0.062	0.845	0.920***	0.000
Customer_Age	-0.032	0.128	-0.010	0.234
Dependent_count	0.144*	0.039	0.140***	0.000
Education_Level (High School)	-0.097	0.730	-0.015	0.908
Education_Level (College)	-0.210	0.530	0.095	0.552
Education_Level (Graduate)	-0.165	0.532	0.135	0.266
Education_Level (Post-Graduate)	0.440	0.289	0.406*	0.038
Education_Level (Doctorate)	0.409	0.388	0.348	0.083
Marital_Status (Married)	-0.509**	0.010	-0.578***	0.000
Marital_Status (Divorced)	-0.115	0.773	-0.078	0.646
Income_Category ($40K - $60K)	-0.058	0.842	-0.169	0.177
Income_Category ($60K - $80K)	0.079	0.854	-0.077	0.705
Income_Category ($80K - $120K)	0.754	0.074	0.271	0.181
Income_Category ($120K +)	0.824	0.084	0.720**	0.003
Card_Category (Silver)	-0.069	0.854	0.555*	0.016
Card_Category (Gold)	0.857	0.136	0.645	0.162
Card_Category (Platinum)	1.728	0.259	0.093	0.945
Months_on_book	0.044*	0.017	-0.005	0.582
Total_Relationship_Count	-0.300***	0.000	-0.477***	0.000
Months_Inactive_12_mon	0.030	0.687	0.436***	0.000
Contacts_Count_12_mon	0.727***	0.000	0.518***	0.000
Credit_Limit	0.000	0.613	-0.000**	0.004
Total_Revolving_Bal	-0.001***	0.000	-0.001***	0.000
Total_Amt_Chng_Q4_Q1	0.000	0.269	-0.001***	0.000
Total_Trans_Amt	0.001***	0.000	0.001***	0.000
Total_Trans_Ct	-0.195***	0.000	-0.134***	0.000
Total_Ct_Chng_Q4_Q1	0.001	0.082	-0.001***	0.000
Avg_Utilization_Ratio	-0.001	0.052	-0.000	0.592
Intercept	6.874***	0.000	4.541***	0.000

Note: Coefficients with *** are significant at 0.1%-level, **1%-level, *5%-level

Based on Table 7 below, LR results in a relatively high accuracy of more than 90% for both clusters on the testing set. The recall, precision, and F1 score for cluster 1 are all approximately 75%. However, the recall rate for cluster 2 is only 64.26%, which means that only 64.26% of the total churned customers were detected by the LR. This could result in significant losses for the bank if measures are not in place to retain the customers. Although cluster 2 has a slightly higher precision rate than cluster 1, the low recall rate consequently results in a lower F1 score.

Table 7. Performance metrics for clusters 1 and 2 using logistic regression on the testing set.

	Logistic Regression	
Evaluation metric	**Cluster 1**	**Cluster 2**
Accuracy	95.03%	90.86%
Recall	75.61%	64.26%
Precision	79.49%	80.08%
F1 score	77.50%	71.30%

5.2.2. Random forest

The tuning of random forest's hyperparameters is done with grid search using the *ranger()* function from the *ranger* package due to the fast computing time in comparison to other packages (Wright & Ziegler, 2017). The performance metrics of the tuned random forest for clusters 1 and 2 could be found in Table 8 below.

Table 8. Performance metrics for clusters 1 and 2 using random forest.

Data set	**Metric**	**Cluster 1**	**Cluster 2**
Train	Out-of-bag error	3.01%	4.82%
Test	Accuracy	97.38%	95.35%
	Recall	93.90%	84.95%
	Precision	84.62%	88.27%
	F1 score	89.02%	86.58%

For the first cluster, the best-performing tuned training set returns an OOB error rate of 3.01%. The hyperparameters for this model are as follows: 10 random variables selected for splitting each time, 6 samples minimum in each terminal node, and a sample size for each tree grown consisting of 80% of total churn and 20% of total non-churn observations. The tuned model is run using the *randomForest()* function in the *randomForest* package and then tested with unseen samples from the testing data set.

The tuned RF for cluster 1 has a high recall rate of 93.90%, meaning that it has a low FN and is good at correctly identifying the customers who actually churned (see Table 9 below).

Table 9. Confusion matrix for cluster 1 testing set using tuned random forest.

		Actual	
		Not churn	**Churn**
Predicted	**Not churn**	628	5
	Churn	14	77

In the second cluster, the hyperparameters that result in the lowest OOB error of 4.82% are: random selection of 8 variables for splitting, minimum node size of 4 samples, and a sample size containing 70% of total churn samples with 30% of total non-churn samples. The precision rate for cluster 2 is 88.27%, which is nearly 4 pp better than that of cluster 1. This means that cluster 2 has a lower proportion of FP, which could lower the unnecessary costs that go into retention efforts for customers who actually fall into the non-churning category. The confusion matrix for the tuned RF of cluster 2 is seen below in Table 10.

Table 10. Confusion matrix for cluster 2 testing set using tuned random forest.

		Actual	
		Not churn	**Churn**
Predicted	**Not churn**	1,451	48
	Churn	36	271

The importance of each predictor variable using Mean Decrease Accuracy (MDA) for both clusters could be seen in Figure 8. Across both clusters 1 and 2, the variables *Total_Trans_Ct*, *Total_Trans_Amt*, *Total_Ct_Chng_Q4_Q1*, *Total_Amt_Chng_Q4_Q1*, and *Total_Relationship_Count* bring the highest contribution to accuracy. The first 2 variables, *Total_Trans_Ct* and *Total_Trans_Amt,* stand out the most in both clusters as their MDA are much higher compared to the rest of the variables; cluster 2 has an additional standout variable of *Total_Relationship_Count*. Variables that seem to be relatively more important for cluster 1's prediction accuracy than cluster 2 are *Months_on_book* and *Contacts_Count_12_mon*. This could be seen through the higher positions of both variables on the MDA for cluster 1. On the other hand, *Total_Relationship_Count* and *Total_Revolving_Bal* seem to be more significant in improving cluster 2's accuracy.

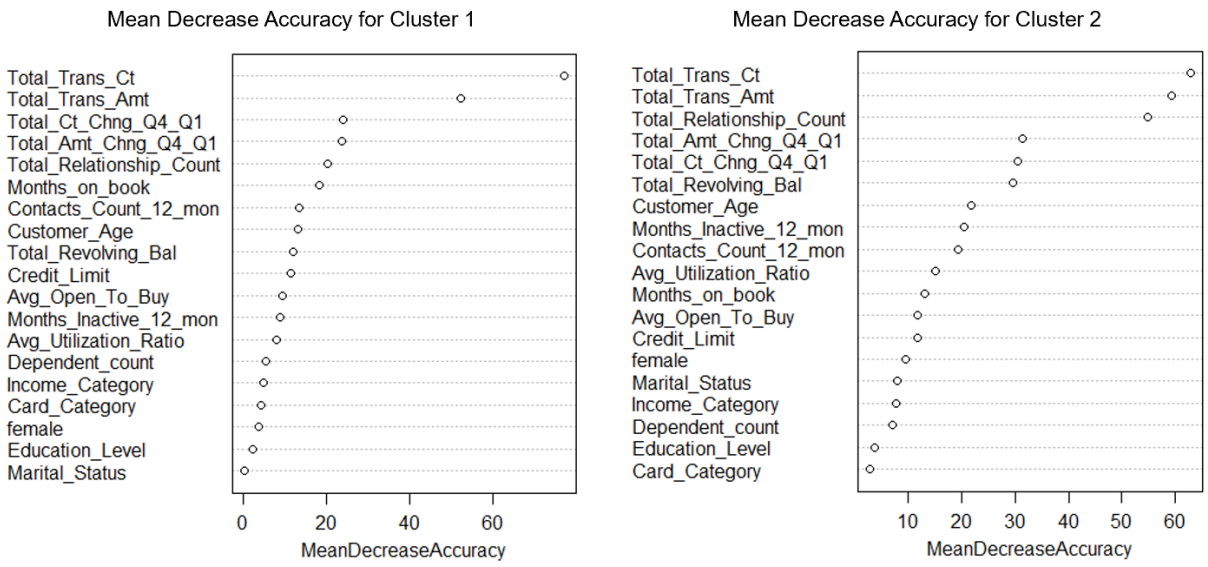

Figure 8. Mean Decrease Accuracy for the tuned random forest of clusters 1 (left) and 2 (right).

According to Mean Decrease Gini (MDG) seen in Figure 9, the top 5 variables for reducing impurity in both clusters are *Total_Trans_Ct, Total_Trans_Amt, Total_Revolving_Bal, Total_Amt_Chng_Q4_Q1*, and *Total_Ct_Chng_Q4_Q1.* Both *Total_Trans_Ct* and *Total_Trans_Amt* have markedly higher MDG than other variables in the 2 clusters; in particular, the former seems to be substantially higher in importance for cluster 1's Gini impurity than cluster 2's (as shown by the MDG of nearly 150 vs. 70 respectively). For the second cluster, another variable that stands out based on its MDG score is *Total_Revolving_Bal.* The same occurrence as with MDA, *Months_on_book* and *Contacts_Count_12_mon* seem to improve the purity of nodes in cluster 1 better than cluster 2. Meanwhile, *Total_Relationship_Count*, *Customer_Age*, and *Months_Inactive_12_mon* are the three variables that stand higher on the MDG for cluster 2.

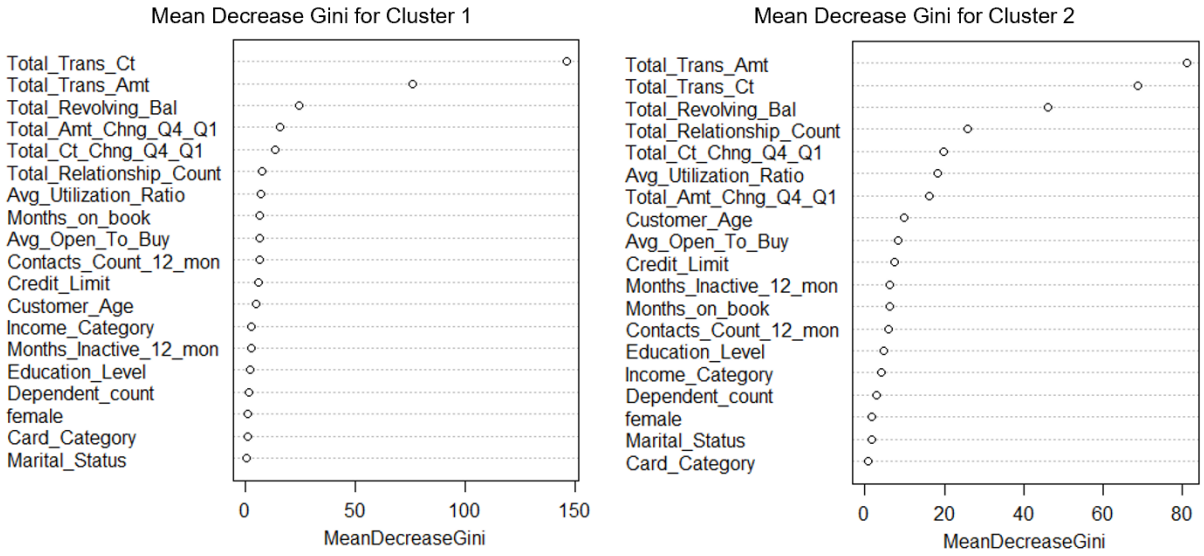

Figure 9. Mean Decrease Gini for the tuned random forest of clusters 1 (left) and 2 (right).

Since RF is a black box method, the effects of aforementioned features that are important in improving the accuracy or purity of the forest would be further explored using ALE plots. The ALE plots are created using the training set; thus, when the following section regarding ALE plots mentions the term "predictions", it is specifically referring to the shortened term for in-sample predictions. The ALE plots for cluster 1's most important variables are shown in Figure 10 below. Distribution of the observations is shown through the black lines on the bottom of each graph. Sparse areas indicate that there are little to no observations in those ranges; hence, interpretation for those ranges is not recommended

and often deemed unreliable. As RF also achieves higher performance metrics than LR, the marketing implications to the bank would also be elaborated in the following section.

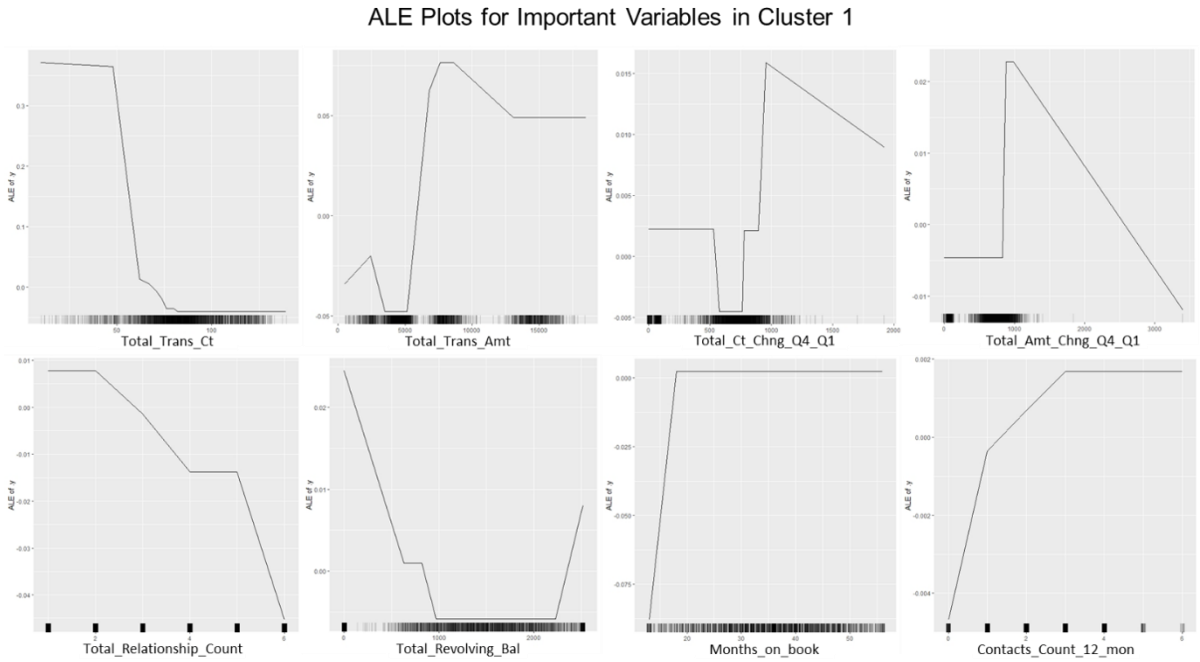

Figure 10. ALE plots for important predictors of churning using random forest in cluster 1.

Two variables seem to have fully negative effects on the average prediction of customers churning as their values increase, namely *Total_Trans_Ct* and *Total_Relationship_Count*. For *Total_Trans_Ct* lower than 50 times, the ALE score is approximately 0.38; this means that the prediction of churning is higher by 0.38 in comparison to the average prediction. When *Total_Trans_Ct* is higher than 70 or *Total_Relationship_Count* is more than or equal to 3, then the churning prediction starts falling below average. This shows that for cluster 1 which consists of big spenders, the bank should set up measures to cross-sell their products and services as well as increase the frequency of usage amongst the credit card customers to at least 70 times annually. This could be done through arranging attractive offers and discounts for restaurants, travel bookings, or other payments. With 79% of customers stating that loyalty programs induce them to continue doing business with companies, another possibility is to create a rewards or loyalty program where the customers could collect points and redeem them in a certain time period (Morgan, 2020).

Both *Months_on_book* and *Contacts_Count_12_mon* show similar occurrences where there is a drastic increase in churning likelihood at the beginning that levels out over time. Nevertheless, customers who are on the book for approximately more than 18 months still seem unlikely to churn as its ALE value is near 0. When customers contact the bank more than or equal to twice, then the churning probability increases to a level above the average prediction. Krishnan (2020) found that the main driver of

customers leaving is poor service; a study even revealed that 86% of customers would pay higher fees in return for better customer experience. Thus, the bank should look into more ways of improving contact with its customers through the creation of a chatbot to help answering the easy and basic questions while leaving the experts to handle more complicated matters. The chatbot could also be used to create new leads by offering customers related new offerings or determining the appropriate offers for their needs (Sinha & Huraimel, 2021).

Unexpectedly, *Total_Trans_Amt*, *Total_Ct_Chng_Q4_Q1*, and *Total_Amt_Chng_Q4_Q1* tend to have on average positive effects on churning, reaching their peaks at around 7500, 920, and 800 respectively. On the other hand, *Total_Revolving_Bal* has the expected negative relationship with churning, except for values higher than 2250. For observations with *Total_Trans_Amt* less than 5000, *Total_Ct_Chng_Q4_Q1* between 500 to 750, *Total_Amt_Chng_Q4_Q1* below 800, and *Total_Revolving_Bal* between 970 and 2250, the predictions fall below 0, which means that the features have negative effects on churning for those values.

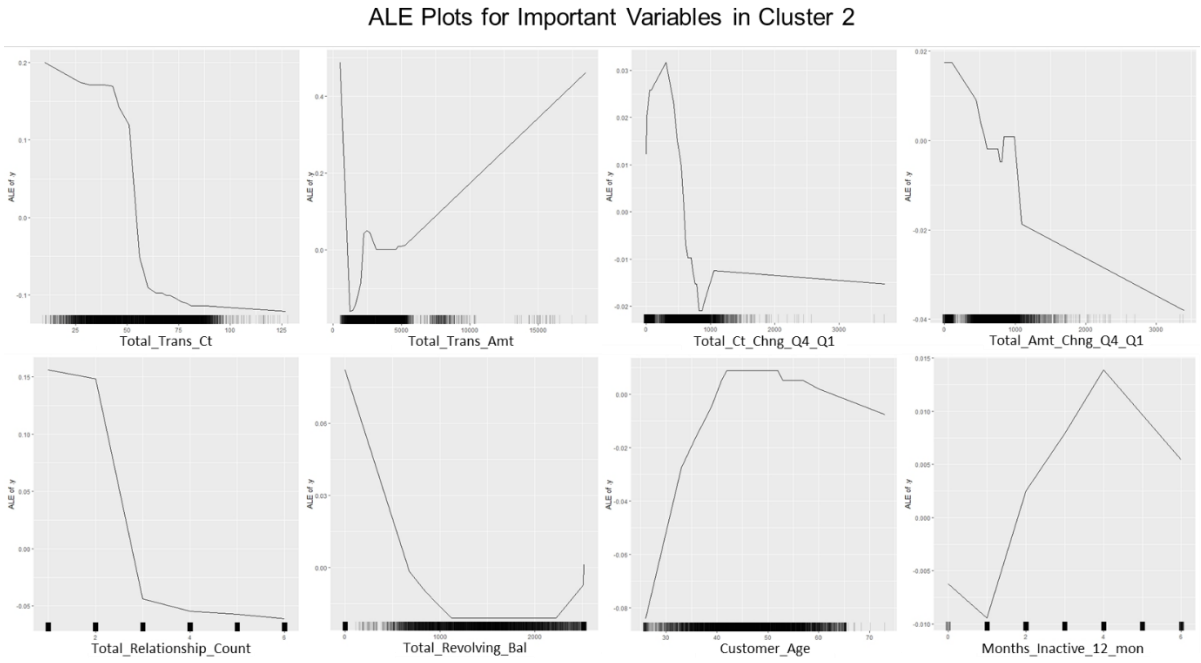

Figure 11. ALE plots for important predictors of churning using random forest in cluster 2.

Figure 11 above depicts the ALE plots for cluster 2's most important variables. As *Total_Trans_Ct*, *Total_Amt_Chng_Q4_Q1*, and *Total_Relationship_Count* get higher, the churning likelihood decreases; the likelihood becomes lower than the average at approximately 50, 650, and 3 for each variable consecutively. Meanwhile, *Total_Ct_Chng_Q4_Q1*, *Customer_Age*, and *Months_Inactive_12_mon* are inverted V-shaped, initially rising and reaching their peaks at 306, 42-52, and 4 respectively before

decreasing. Customers with *Total_Ct_Chng_Q4*_Q1 > 617, *Customer_Age* ≤ 40 or ≥63, and *Months_Inactive_12_mon* < 2 have below average prediction of churning likelihood.

Similar to cluster 1, cluster 2 customers could be retained through efforts that would increase the transaction and relationship count. However, the focus on cluster 2 should be directed more towards improving the transaction count and amount as the customers are "Passive" due to the lower usage rate despite the higher relationship count. This is also supported by the fact that an increase in the change of transaction count and amount from Q4 to Q1 leads to a decline in churning likelihood. In this case, the bank should ensure that it has attractive offers ongoing in sectors where the customers seem to be using credit cards as a form of payment. For example, if the customers in cluster 2 mostly use their cards for grocery payments, then the bank could arrange deals or partnerships with certain supermarkets. The bank could also partner up with telecom companies to increase knowledge about the services and incentives offered by sending regular communications through text messages. This is particularly useful for customers in the rural parts of Africa or the Middle East where trust and knowledge on digital payments are still low (Salazar & Monteverde, 2019). Mastercard did this through the introduction of text messaging options where customers could get answers on the bank's offerings or their financial pain points. Better knowledge and higher trust from the customers lead to usage growth; as a result, the inactivity rate would also decrease.

The finding with age is in line with previous ones where mature customers aged approximately above 55 are less likely to churn due to their inclination for building and maintaining relationships with the bank, which in turn results in stability and reliability (Ansell et al., 2007; Dias et al., 2020). To retain these customers, the bank should train the staffs on communicating with the elderly, maintain accessibility by keeping some branches open for offline interactions, keep the choice of options for paper-based statements, and design a user-friendly interface to ease reading and understandability. Possible reasons for younger customers below 30 years of age having less likelihood to churn than the average prediction are less knowledge about finance best practices and less choice compared to the older age groups who earn higher income and have established their credit scores. Their needs for finding the best amenities between banks are also lower as the offers are usually similar across the industry for this age group. For this age group, the bank could best retain them by developing personal financial robo-advisors, as they are mainly focused on saving and investing more. A recent study by Vanguard shows that millennials are twice as likely as the older generations to seek digital advisors, with the average age for clients in Vanguard Digital Advisor being 37 years old compared to the Personal Advisor Services which is 57 years old (Iacurci, 2020). Meanwhile, middle-aged customers from 40-55 years old have higher likelihood to churn as they are more well-informed of the benefits offered from each bank and typically has higher income which leads to higher credit rating (Mavri &

Ioannou, 2008). These customers are usually in the stage of preparing for retirement and inheritance and thus, often look for the best deals regardless of the bank. Therefore, to retain this age group, the bank needs to come up with plans that ensure asset security and protection.

Total_Trans_Amt and *Total_Revolving_Bal* have V-shaped and U-shaped plots respectively, with below average churning prediction for values between 1245 to 2016 for the former and 680 to 2514 for the latter. For the transaction amount, the line lies around the average prediction from 2500 up to 5000, at which point it starts increasing; nevertheless, those ranges above 5000 have sparse observations. In both clusters, the positive effect shown by *Total_Trans_Amt* could either indicate that the bank does not possess attractive arrangements for customers who are big spenders or the bank should build measures to prevent big spenders from churning. Although the generally negative effect of *Total_Revolving_Bal* gives some reassurance that those with increasing carryover debt stay in the bank, the upward slope beyond 2250 presents some risk. In all of the implications described above, care should be taken so that the efforts to increase usage in certain customers are aligned with their income and credit ratings.

6. Conclusion

This study attempts to uncover the segments that are present in credit card customers and the variables that influence churning likelihood. The main research question is: "*What customer segments could be found within credit card customers?*". The segmentation is done with 5 variables, namely *Months_on_book* as the length of relationship between customers and the bank, *Total_Relationship_Count* as the depth of relationship, *Months_Inactive_12_mon* as the inactivity indicator, *Total_Trans_Amt* as the usage in monetary amount, and *Total_Trans_Ct* as the frequency of usage. Rough k-means with 2 clusters have the best silhouette score compared to those of regular k-means and k-medoids. The 2 clusters found could be differentiated based on their revenue generation and relationship with the bank.

To answer the research sub-question of "*What characteristics differentiate the segments with high and low proportions of churning customers?*", the descriptive statistics of the clusters found are investigated. The first cluster is called "Spender" and it is characterized by a lower proportion of churning as well as noticeably higher total transaction amount and count than the second cluster "Passive". In fact, the minimum transaction amount for cluster 1 is nearly the same as the maximum amount for cluster 2. However, the second cluster has a deeper relationship with the bank in terms of the average number of products hold by each customer; the range is also larger with the maximum relationship count being twice as high as cluster 1 (6 vs. 3 products).

After the segmentation, the churning prediction is conducted using logistic regression and random forest due to their interpretability and accuracy. This is done to answer the second research sub-question: "*Which factors are most indicative of customers' likelihood to churn in different segments?*". The sub-question aims to find the important variables that could be used as indicators of churning customers so that preventative measures could be created. Random forest performs extensively better than logistic regression in all performance metrics for both clusters. In particular, the recall rates for random forest are 93.90% and 84.95% for clusters 1 and 2, which are substantially better than those of logistic regression (75.61% and 64.26% for clusters 1 and 2). This is beneficial for the bank as it reduces the likelihood of customers lost due to them being incorrectly classified as falling into the non-churning group. In both models, the results show similarity with that of Nie et al. (2011) where the demographic variables do not provide much influence to the churning prediction and most contributions originate from the usage behavior and card information.

The most important variables in both clusters using RF are similar, namely *Total_Trans_Ct, Total_Ct_Chng_Q4_Q1*, *Total_Trans_Amt, Total_Amt_Chng_Q4_Q1*, *Total_Revolving_Bal*, and *Total_Relationship_Count*. Variables *Months_on_book* and *Contacts_Count_12_mon* show more importance for cluster 1; meanwhile, *Customer_Age*, and *Months_Inactive_12_mon* are more important in churning prediction for cluster 2. For customers in cluster 1, the bank could develop a chatbot that could handle simple inquiries from the customers efficiently to reduce the number of contacts count to below 2. The chatbot could also be trained to detect customers' satisfaction and needs or inform the customers about ongoing offers so that it could also cross-sell different products which the customers might need. This would help increase relationship count with the bank and decrease customers' likelihood to churn. For cluster 2 customers, the bank should encourage usage through better offers and communications as well as develop the services for asset security and protection to retain middle-aged customers.

The main limitation of this research stems from the data set. The data set does not provide any information about the locations of operation or the name of the bank. This makes it impossible to understand the credit card usage and industry situation in the country where the information is gathered. Moreover, comparison between the bank and its competitors' financial situation, advantages and disadvantages, and offerings is also not feasible. Lastly, each row only consists of summarized transaction information for each customer instead of each transaction that the customer does. Having more details on each transaction would allow a larger number of observations to be processed in the model and hence, a more comprehensive result that could potentially account for factors such as seasonality.

Another limitation is that the variable importance might be biased due to the different range in each variable. Some continuous variables have larger ranges than the categorical ones. This could result in variable selection bias when constructing the trees and thus, irrelevant variables being listed as the important ones. Some solutions suggested by Strobl et al. (2007) are to use the *cforest* function in R as it forms an unbiased classification tree and constructs the random forest without sampling replacement.

Therefore, it is recommended for future research to replicate the research using a dataset that has the aforementioned features, such as the bank's name, country of location, and detailed customer transactions. Results from banks in different regions could then be compared and analyzed for differences and similarities in the most important variables and their effects on churning. Future research could also account for different ranges that the variables possess and create the random forest using sampling without replacement.

References

Ansell, J., Harrison, T., & Archibald, T. (2007). Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis. *Marketing Intelligence & Planning, 25*(4), 394-410. doi:10.1108/02634500710754619

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science, 78*, 507-512. doi:10.1016/j.procs.2016.02.095

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Boca Raton: Chapman and Hall/CRC.

Borbora, Z., Srivastava, J., Hsu, K. W., & Williams, D. (2011, October). Churn prediction in MMORPGs using player motivation theories and an ensemble approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 157-164). IEEE.

Bose, I., & Chen, X. (2009). Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn. *Journal of Organizational Computing and Electronic Commerce, 19*(2), 133-151. doi:10.1080/10919390902821291

Brooks, R., & White, D. (1996). Don't copy the competition--lead with new products. *Bank Marketing*, *28*(5), 13-17.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252-268. doi:10.1016/j.ejor.2003.12.010

Chang, H. H., & Tsay, S. F. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management, 11*(4), 161-203.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197-208. doi:10.1057/dbm.2012.17

Cheng, C., & Chen, Y. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176-4184. doi:10.1016/j.eswa.2008.04.003

Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: A study of the student market in Ireland. *International Journal of Bank Marketing, 14*(3), 23-29. doi:10.1108/02652329610113144

Dias, J., Godinho, P., & Torres, P. (2020). Machine Learning for Customer Churn Prediction in Retail Banking. *Computational Science and Its Applications – ICCSA 2020 Lecture Notes in Computer Science,* 576-589. doi:10.1007/978-3-030-58808-3_42

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives, 18*, 153-160. doi:10.1016/j.tmp.2016.03.001

Frankel, M. (2020, July 16). Credit Card Debt Statistics for 2020: The Ascent. Retrieved May 4, 2021, from https://www.fool.com/the-ascent/research/credit-card-debt-statistics/

Griffin, J. (2002). *Customer loyalty: How to earn it, how to keep it*. San Francisco: Jossey-Bass.

Hughes, A. M. (1994). *Strategic database marketing*. Chicago: Probus Publishing Company

Hung, S., Yen, D. C., & Wang, H. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515-524. doi:10.1016/j.eswa.2005.09.080

Iacurci, G. (2020, October 14). Young investors are going digital. Financial advisors need to adapt with them. *CNBC*. Retrieved August 3, 2021, from https://www.cnbc.com/2020/10/14/millennials-gen-z-want-robo-advisors-and-digital-financial-advice.html

Jiang, J., Kasamatsu, K., & Ainoya, T. (2020). Research on Payment UX Status During the Share Cycle Services Between Japan and China. *Lecture Notes in Computer Science HCI International 2020 – Late Breaking Papers: Interaction, Knowledge and Social Media,* 522-534. doi:10.1007/978-3-030-60152-2_39

Karakostas, B., Kardaras, D., & Papathanassiou, E. (2005). The state of CRM adoption by the financial services in the UK: An empirical investigation. *Information & Management, 42*(6), 853-863. doi:10.1016/j.im.2004.08.006

Kim, L., Kumar, R., & O'Brien, S. (2020, July 31). *2020 Findings from the Diary of Consumer Payment Choice* [PDF]. San Francisco: Federal Reserve System.

Krishnan, K. (2020). *Building big data applications*. London: Elsevier.

Kumar, D. A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies, 1*(1), 4. doi:10.1504/ijdats.2008.020020

Lantz, B. (2013). *Machine Learning with R*. Birmingham: Packt Publishing.

Lax, H. (2016, July 12). New Customer Retention: A Fundamental in Retail Financial Services. Retrieved May 4, 2021, from https://customerthink.com/new-customer-retention-a-fundamental-in-retail-financial-services/

Li, W., Wu, X., Sun, Y., & Zhang, Q. (2010). Credit Card Customer Segmentation and Target Marketing Based on Data Mining. *2010 International Conference on Computational Intelligence and Security*. doi:10.1109/cis.2010.23

Lin, C., Tzeng, G., & Chin, Y. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications, 38*(1), 8-15. doi:10.1016/j.eswa.2010.05.039

Lin, W., Tsai, C., & Ke, S. (2014). Dimensionality and data reduction in telecom churn prediction. *Kybernetes, 43*(5), 737-749. doi:10.1108/k-03-2013-0045

Martínez-Muñoz, G., & Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition, 43*(1), 143-152. doi:10.1016/j.patcog.2009.05.010

Mavri, M., & Ioannou, G. (2008). Customer switching behaviour in Greek banking services using survival analysis. *Managerial Finance, 34*(3), 186-197. doi:10.1108/03074350810848063

Meadows, M., & Dibb, S. (1998). Assessing the implementation of market segmentation in retail financial services. *International Journal of Service Industry Management, 9*(3), 266-285. doi:10.1108/09564239810223565

Mo, J., Kiang, M. Y., Zou, P., & Li, Y. (2010). A two-stage clustering approach for multi-region segmentation. *Expert Systems with Applications, 37*(10), 7120-7131. doi:10.1016/j.eswa.2010.03.003

Morgan, B. (2020, May 7). 50 Stats That Show The Importance Of Good Loyalty Programs, Even During A Crisis. Retrieved August 4, 2021, from https://www.forbes.com/sites/blakemorgan/2020/05/07/50-stats-that-show-the-importance-of-good-loyalty-programs-even-during-a-crisis/

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications, 38*(12), 15273-15285. doi:10.1016/j.eswa.2011.06.028

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence & Planning, 35*(4), 544-559. doi:10.1108/mip-11-2016-0210

Probst, P., & Boulesteix, A. (2017). To Tune or Not to Tune the Number of Trees in Random Forest. *Journal of Machine Learning Research*, *18*(1), 6673-6690.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, *20*(1), 1934-1965.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

Rajamohamed, R., & Manokaran, J. (2017). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing, 21*(1), 65-77. doi:10.1007/s10586-017-0933-1

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi:10.1016/0377-0427(87)90125-7

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review, 1*(2). doi:10.1162/99608f92.5a8a3a3d

Salazar, D., & Monteverde, C. (2019, February 4). 3 Ways to advance usage and drive impact in financial inclusion. Retrieved August 2, 2021, from https://blogs.worldbank.org/allaboutfinance/3-ways-advance-usage-and-drive-impact-financial-inclusion

Sinha, S., & Huraimel, K. A. (2021). *Reimagining businesses with AI*. Hoboken, NJ: Wiley.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing, 21*(1), 3. doi:10.2307/1247695

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine, 18*(10), 1099-1104. doi:10.1111/j.1553-2712.2011.01185.x

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1). doi:10.1186/1471-2105-8-25

Wertz, J. (2018, September 12). Dont Spend 5 Times More Attracting New Customers, Nurture The Existing Ones. *Forbes*. Retrieved March 31, 2021, from

https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/

van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research, 181*(1), 436-452. doi:10.1016/j.ejor.2006.05.029

Wright, M. N., & Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C and R. *Journal of Statistical Software, 77*(1). doi:10.18637/jss.v077.i01

Appendices

Appendix A. Boxplot of descriptive statistics using the 5 segmentation variables for both clusters 1 and 2

Figure A1. Comparison of descriptive statistics between clusters 1 and 2 using rough k-means clustering.