IME 451: Advanced Statistical Methods

Report on:

Data Analytics Project

1/2/2021

Prepared by:

Rahma Mohamed (120170027)

Rana Mohamed (120170001)

Rana Samir (120170008)

Submitted to:

Dr. Mohamed Gheith

Table of Contents

1. Introduction

A lot of customers are leaving their credit card services and the manager of the bank want to investigate the reasons behind their churn. The manger wants to predict which customers, from the currently existing, have a higher probability of being churned, so the decision makers can proactively make more programs, offers, and initiatives to change customers experiences and develop loyalty programs and retention campaigns so they don't leave, and maybe this would result a comeback from attrited customers.

The data set used from the bank was collected in a wide timeline to collect all customers data and whether they are existing or attritted customers. Our goal here is to model the probability of churn, conditioned on the customer features.

2. Data Description

A manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. [1]

We got this dataset from Kaggle [1]. We have been using this for a while to produce fruitful results.

Now, this dataset consists of 10,127 customers mentioning their different features. There are 21 variables showing the customers' different features. [1]

We have only 16.07% of customers who have churned. Thus, it was a bit difficult to train our model to predict churning customers.

The explanation of the variables included in the dataset is as follows:

1. CLIENTNUM: unique ID for each client.
2. Attrition_Flag: has 2 levels and shows whether a customer is an attrited customer or an existing customer.
3. Customer_Age
4. Gender
5. Dependent_count: The number of people who are under the responsibility of that customer.
6. Education_Level: has 7 possible values which are Unknown, Uneducated, High School, College, Graduate, Post-Graduate, Doctorate.
7. Marital_Status: has 4 possible values which are Divorced, Married, Single, Unknown.
8. Income_Category: has 6 possible values which are unknown, less than $40k, $40K-$60K, $60K-$80K, $80K-$120K, more than $120K.
9. Card_Category: has 4 possible values which are Blue, Gold, Platinum, Silver.
10. Months_on_book: period of relationship with bank.
11. Total_Relationship_Count: Total number of products held by the customers, e.g. cards, accounts, etc.)

12. Months_Inactive_12_mon: Number of months that customer is inactive in the last 12 months.
13. Contacts_Count_12_mon: Number of times the bank contacted the customer and/or vice versa in the 12 months.
14. Credit_Limit: Maximum amount that can be borrowed using a credit card or line of credit.
15. Total_Revolving_Bal: Total unpaid balance (dept) on the credit account).
16. Avg_Open_To_Buy: Average difference between the credit limit assigned to a cardholder account and the present revolving balance on the account. In other words, the money currently available for the cardholder to borrow using the credit card.
17. Total_Trans_Amt: Amount of all transactions in the last 12 months
18. Total_Trans_Ct: Number of all transactions in the last 12 months
19. Total_Amt_Chng_Q4_Q1: Change in customer's expenditure between the 4th and the 1st quarter. In other words, it is the total transaction amount in the 4^{th} quarter divided by that in the 1^{st} quarter.
20. Total_Ct_Chng_Q4_Q1: Change in the number of transactions made by customer between the 4th and the 1st quarter. In other words, it is the total transaction count in the 4^{th} quarter divided by that in the 1^{st} quarter.
21. Avg_Utilization_Ratio: Credit Utilization Ratio is how much the customer currently owe the bank divided by the credit limit. In other words, the average utilization ration is equal to the total revolving balance divided by the credit limit

In order to be able to further understand what these variables mean in the bank industry which was needed to interpret and analyse the results, further investigation and search was performed to better understand the dataset we have. The main important conclusion we reached are as follows:

1. A low credit utilization ratio is considered an indicator that the customer is doing a good job of managing their credit responsibilities because they're far from overspending. The general rule of thumb with credit utilization is to stay below 30 percent. Anything higher than 30 percent can make lenders worry that this customer is having trouble managing their finances and they maybe overspending and will have difficulty repaying new debts.[2] A high credit utilization indicates that the customer is probably spending a significant portion of their monthly income on debt payments, and this puts them at a higher risk of defaulting on credit payments (at least in the eyes of creditors). [3]

2. Average utilization ratios below 1% is also unrecommended. They may indicate that the customer is avoiding using the credit cards completely. Most credit scoring models view 1% utilization as better than 0%. If the average utilization ratio is 0%, it shows lenders and credit bank that the customer isn't making any purchases using credit card and the bank may eventually consider the account inactive. [4] When the account is idle, the bank makes no money from transaction fees or from interest rates. Banks may decide to cancel inactive accounts to give that line of credit to someone who will use it. [5]

3. There are many ways that can result in lower average utilization ratio, such as: using the credit card less, increasing the credit limit and paying the dept earlier [6]

3. Objective

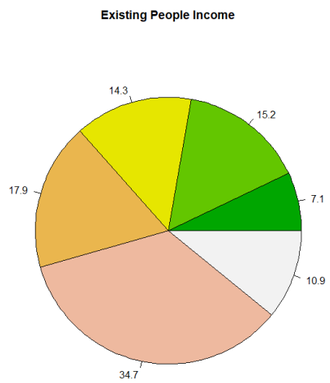
This project is carried out in a sequence of steps, the first of which consists of an exploratory analysis, where the objective is to know the behaviour of the variables and to analyse the relationship among different variables and the attributes that are highly correlated with the cancellation of credit card service customers.

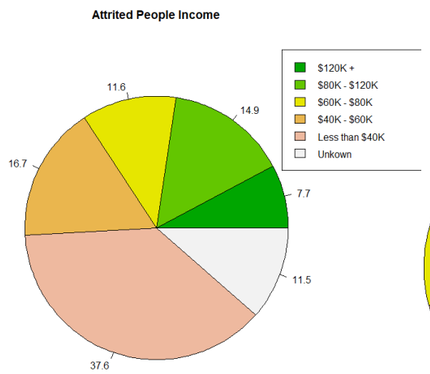
The second step is to build prediction model to accurately predict whether a customer will leave the credit card service or not.

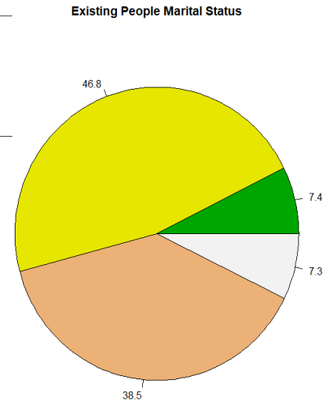
The third step is to divide our customer base into different segments where each customer segment has common features and behaviours to better understand our customers base and better customize our analysis and recommendation based on different customer trends.

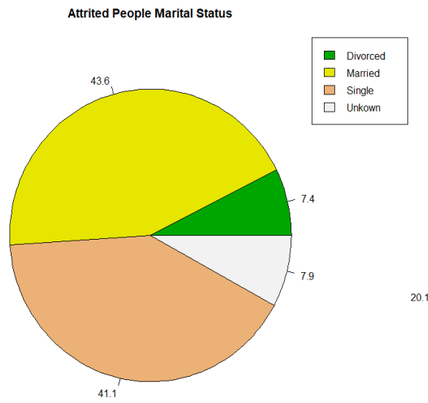
4. Data Exploration and Data Visualization

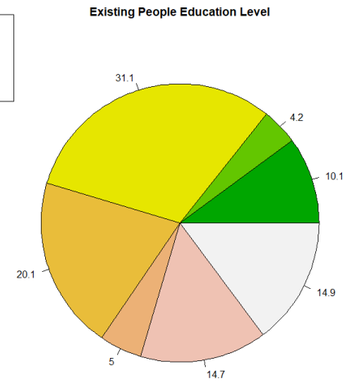
4.1. Categorial Data

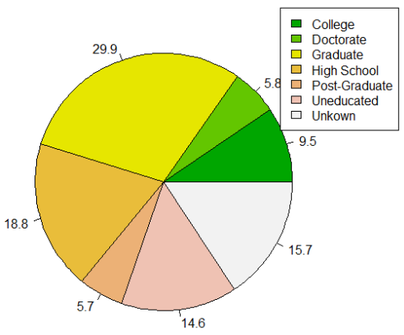












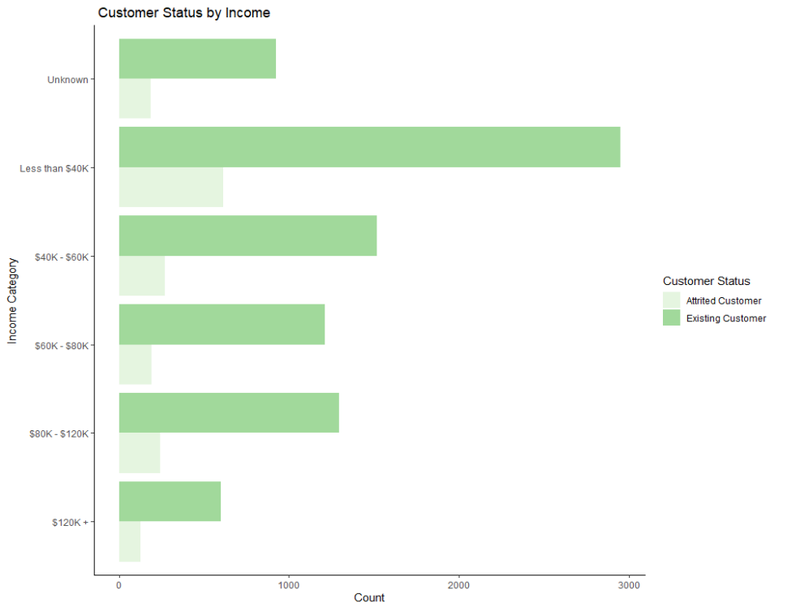


Figure (1): Customer Status by Income

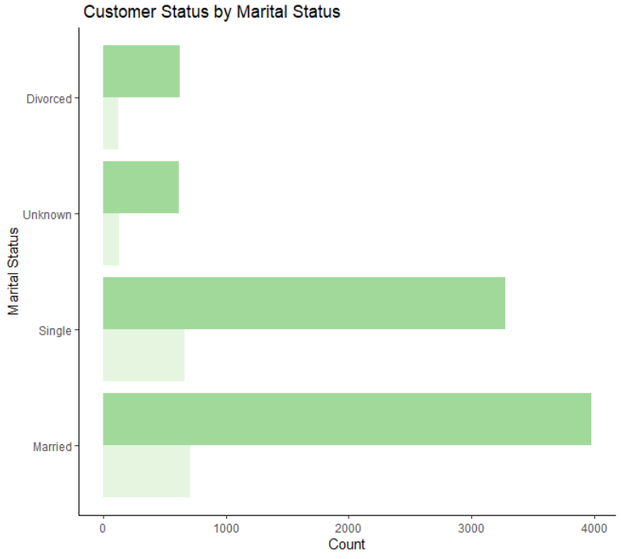


Figure (2): Customer Status by Marital Status

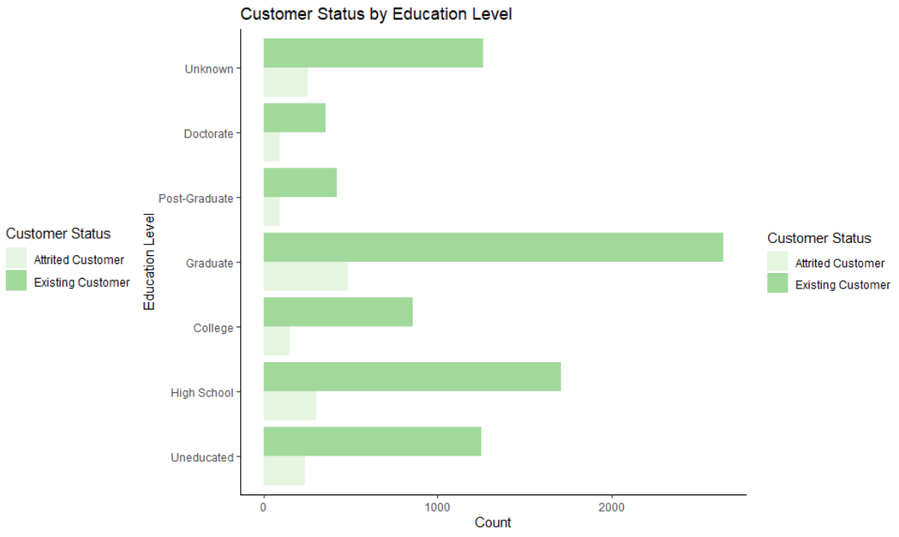
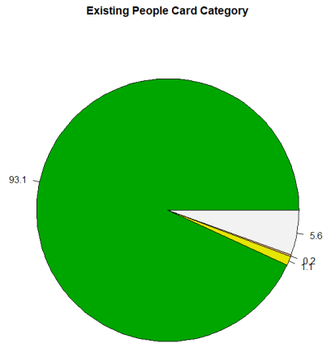


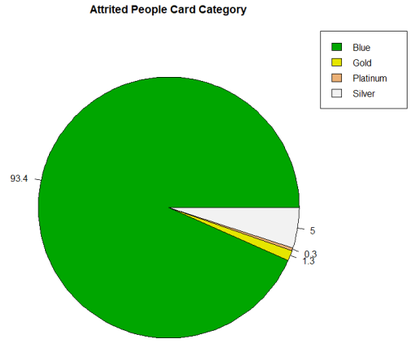
Figure (3): Customer Status by Education Level

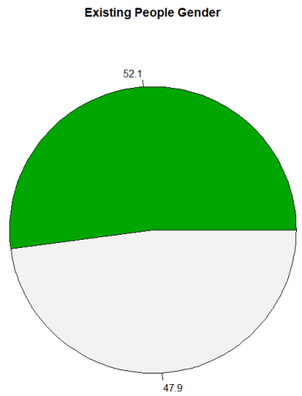
As shown in figure 1, for most of the existing customers, the income statues are less than $40k/ year. Compared to the attired customers, the ratio between the annual income is nearly the same. So, the income statues don't suggest a huge difference between the existing and attired customer.

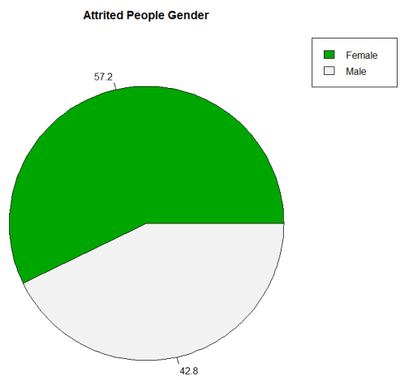
As shown in figure (2), most of the customers are married, however, the single customers have a large part between customer's marital status. We can also notice here that there is no great difference between the ratios of the marital statues between the existing and attired customers.

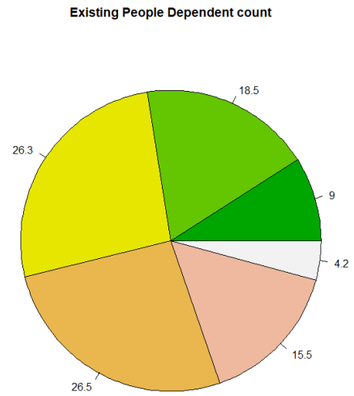
As shown in figure (3), main education level of the customers is graduated people, followed by high school. We can also notice here that there is no great difference between the ratios of the educational level between the existing and attired customers.

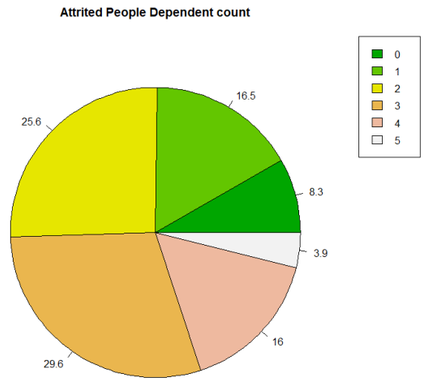












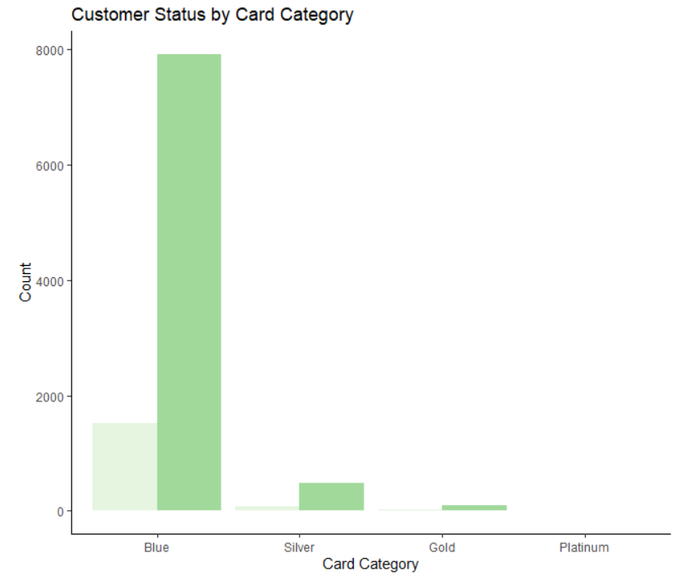


Figure (4): Customer Status by Card Category

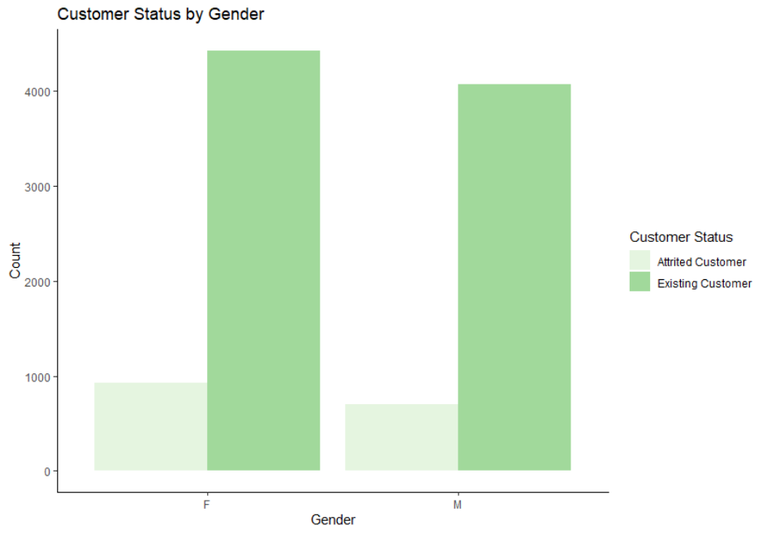


Figure (5): Customer Status by Gender

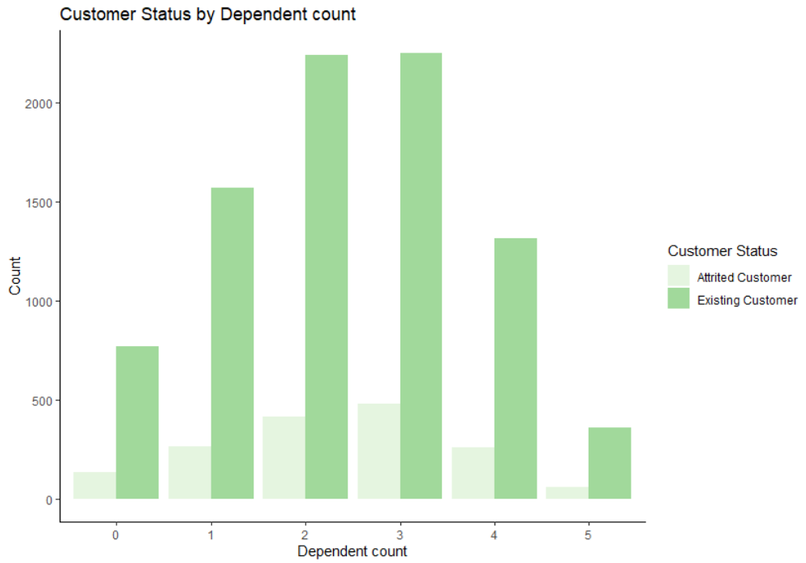


Figure (6): Customer Status by Dependent Count

As shown in figure (4), main card category of the customers is Blue With a great difference between other card categories. It's clear here that there is no great difference between the ratios of the card categories between the existing and attired customers.

As shown in figure (5), Most of the customer are female customers. It's well notable here that the same ratio between the gender of existing and attired customers apply in both genders.

As shown in figure (6), For most of the existing customers, the dependent count was the most at 2 and 3. Compared to the attired customers, the ratio between the dependent count is nearly the same. So, there is no great difference between the ratios of the dependent count between the existing and attired customers.

4.2. Numerical Data

For figure (7), both types of customers nearly have the same age distribution, with existing customers having more outliers affecting the median.

For figure (8), total transaction of existing customs has a huge difference between the attritted customers. Attired customers distribution is more positively skewed and have a mean value lower than 50, other than the existing customers who have a mean value around 75.

For figure (9), existing customers tend to have more count change than attired customers. with existing customers having more outliers affecting the median.

For figure (10), there is a huge difference between the distribution of the total revolving balance between the two types of customer. As existing customers tend to have more a mean total revolving count near 1500.

For figure (11), there is a huge difference between the distribution of the average utilization ratio between the two types of customer. As existing customers tend to have more a mean average utilization ratio near 0.25 and attired customers have near 0.

For figure (12), the credit card limit of both types of customers have nearly the same distribution with close mean values. Moreover, the credit card limit is positively skewed in the bank.

For figure (13), the total transaction amount distribution is fluctuating in both types of customers with a lot of outliers in the data. The only notable thing is that existing customers tend to make more transactions amounts than attired customers.

For figure (14), the number on book of both types of customers have nearly the same distribution with close mean values. Moreover, both of the distributions are centred with a close number of outliers.

For figure (15), the total relationship count of both types of customers is fluctuating (because it's a discrete data) with no outliers. The only notable thing is that attrited customers tend to have more dispersion other than the existing customers who's more concentrated in some numbers than other.

For figure (16), the average total months inactive for each customer in the last 12 months of both types of customers is fluctuating (because it's a discrete data) with some outliers in the attrited customers distribution. The notable thing is that attrited customers tend to have more concentrated other than the existing customers who's more speeded in some numbers than other.

For figure (17), the average total contacts count for each customer in the last 12 months of both customer status is fluctuating (since it's a discrete data) with some outliers in the existing customers distribution. In the existing customers' boxplot, all the data fall over the median.

For figure (18), The average open to buy for each customer is positively skewed in the bank with a notable number of outliers.

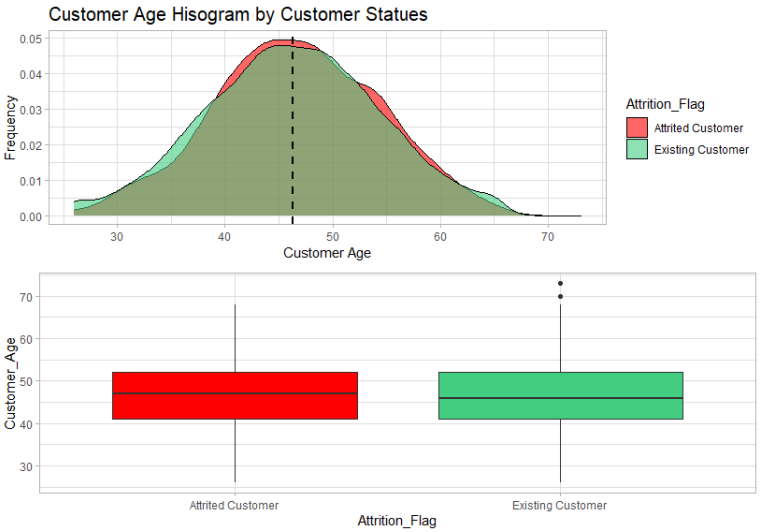


Figure (7): Customer Age Distribution by Customer Status

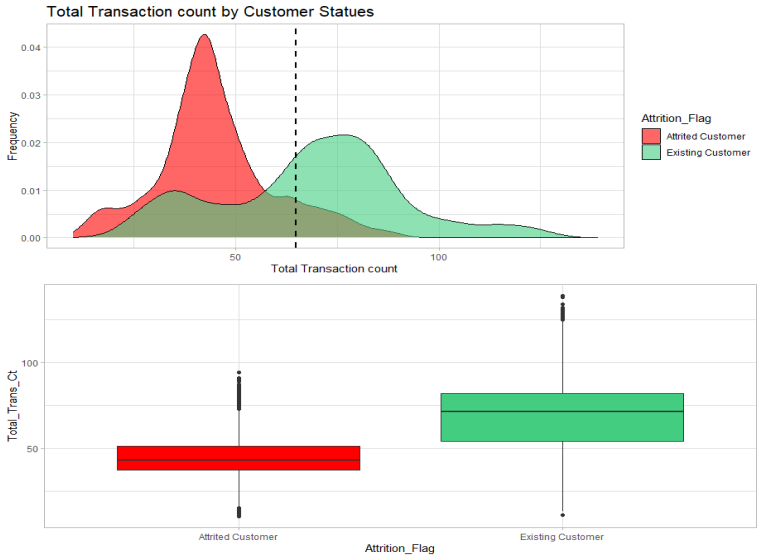


Figure (8): Total Transaction Count by Customer Status

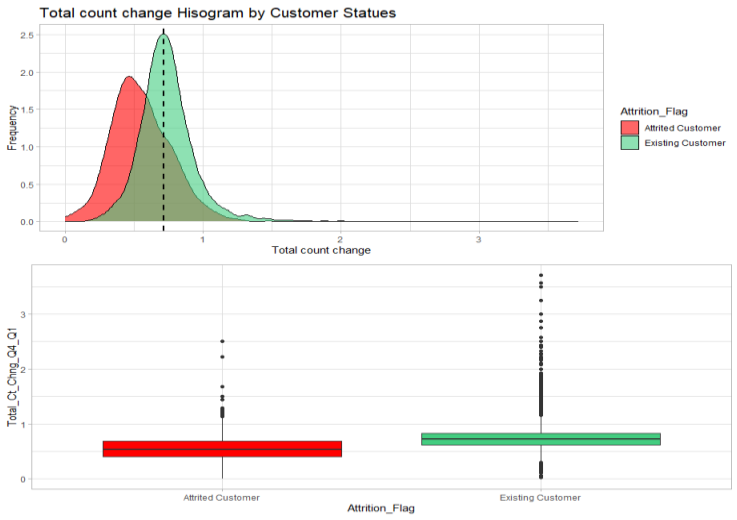


Figure (9): Total Transaction Count Change by Customer

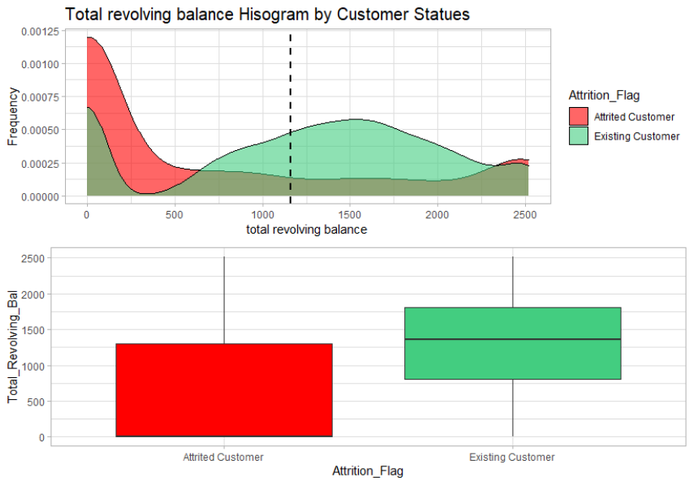


Figure (10): Total Revolving Balance by Customer Status

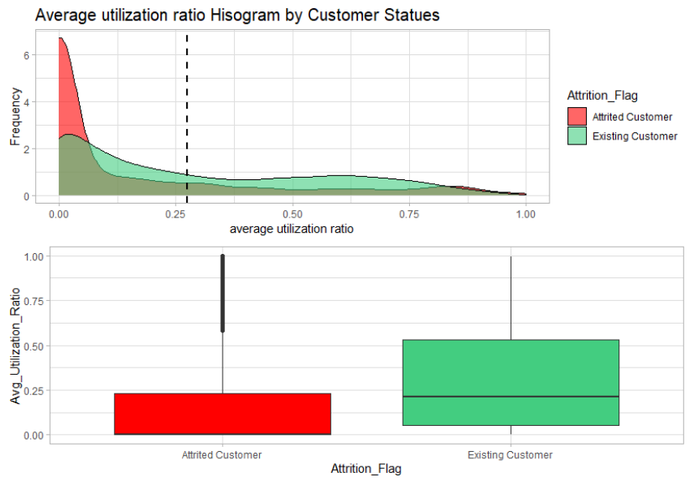


Figure (11): Average Utilization Ratio by Customer Status

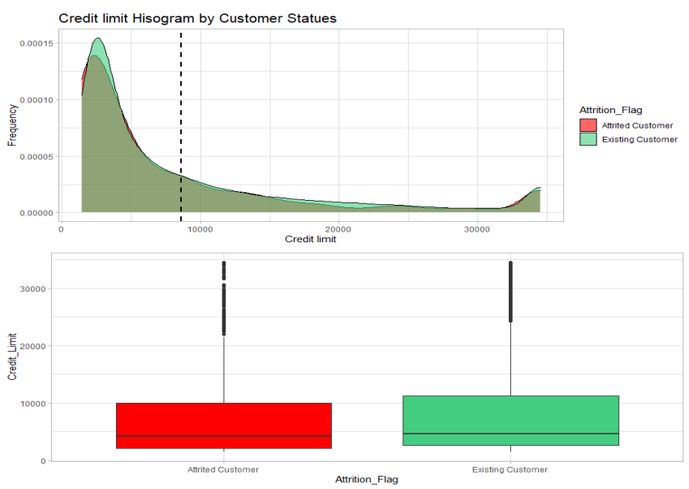


Figure (12): Credit Limit Ratio by Customer Status

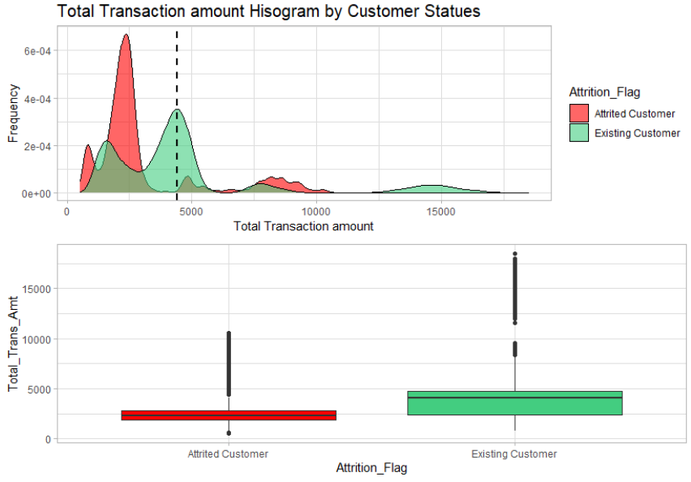


Figure (13): Total Transaction Amount by Customer Status

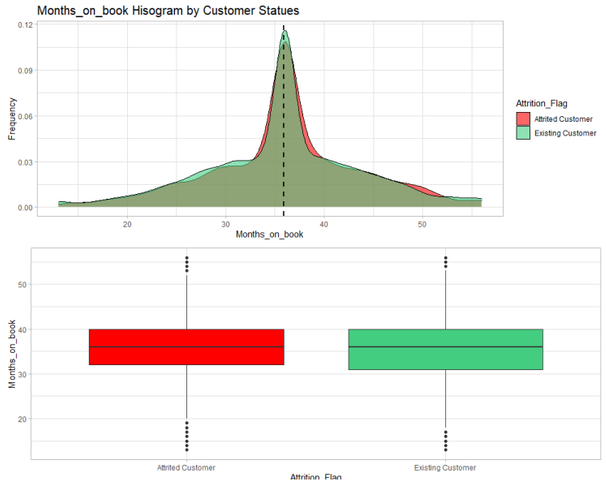


Figure (14): Months on Book by Customer Status

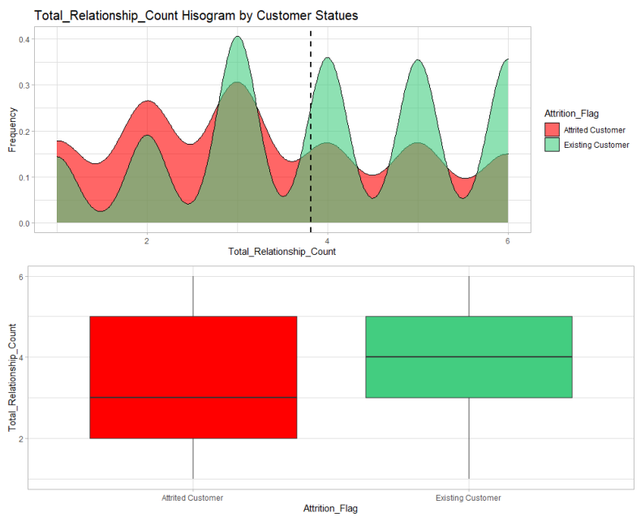


Figure (15): Total Relationship Count by Customer Status

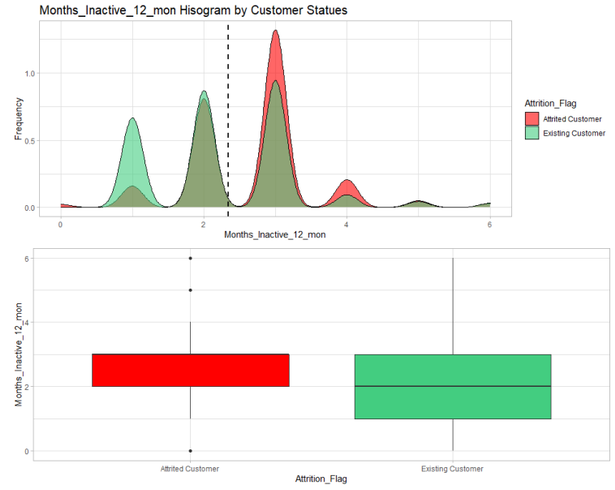


Figure (16): Months Inactive by Customer Status

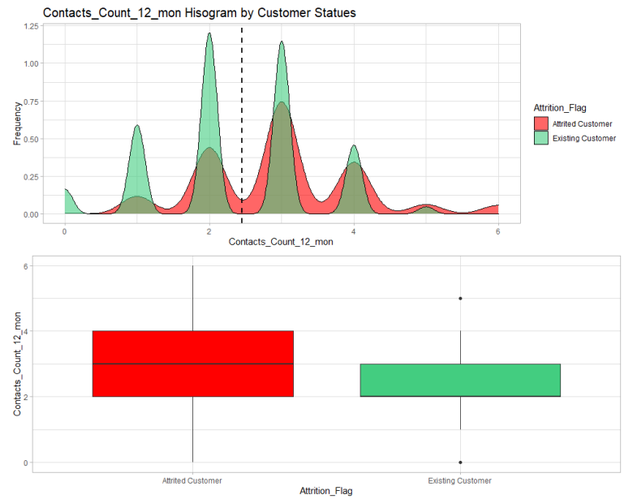


Figure (17): Contacts Count by Customer Status

Figure (18): Average Open to Buy by Customer Status

4.3. Variables Correlation Matrix

- Avg_Open_To_Buy is highly correlated to Credit_Limit
- Total_Trans_Ct amount is highly correlated to Total_Trans_Amt
- Total_Amt_Chng_Q4_Q1 is correlated to Total_Ct_Cng_Q4_Q1
- Avg_Utilization_Ratio is correlated to Avg_Open_To_BUy, Total_Revolving_Bal and Credit_Limit
- Months_on_book is highly correlated with Customer Age
- Total_relationship_count is correlated with Total_Trans_Amt and Total_Trans_Ct

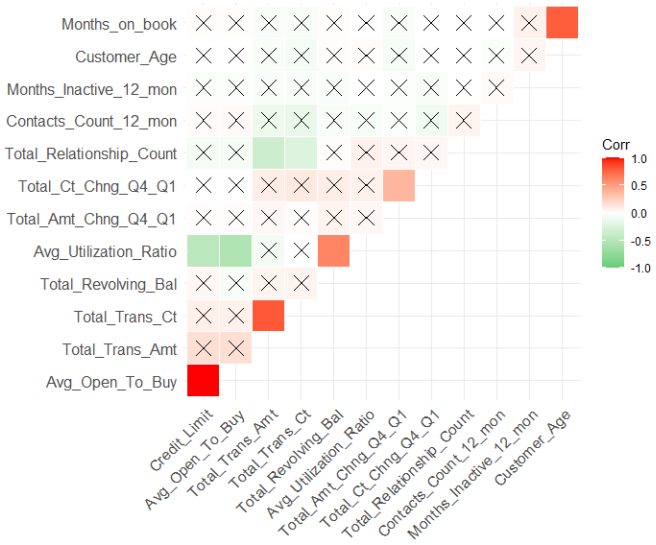


Figure (19): Variables Correlation Matrix

5. Data Analysis Techniques

5.1. Data Preparation

The dataset has been prepared in order to make it efficient for usage. The following preparation methods have been applied to specific variables according to what needed.

5.1.1. Removing Insignificant Variables

There are some variables that have been removed from the whole dataset as we have seen it does not affect our objective. The variables that have been removed are:

1. CLIENTNUM
2. Naive_Bayes_Classifier_Attritio
3. Naive_Bayes_Classifier_Attrit_1

The last 2 variables were removed because they are dependent variables whose values are calculated by applying so algorithm called "Naive Bayes" to the other variables in the dataset and thus they don't describe a specific feature for each customer and thus they were removed.

5.1.2. Converting Some Variables from Character Datatype into Numeric Datatype

There are some categorical variables in the dataset with a character data type which we see that they may be significant to our objective in the analysis technique used. Therefore, these categorical variables were converted into numerical type by giving each category a specific value.

5.1.2.1. Binary Categorical Variables

These are variables that can take on exactly one of two possible values (e.g. Attrition Flag and Gender), so they were converted into numeric binary variables.

The manipulated binary variables are as follows:

1. Attrition_Flag:
   - Existing Customer: 0
   - Attrited Customer: 1
2. Gender
   - Female: 1
   - Male: 0

5.1.2.2. Ordinal Categorical Variables

These are variables that contain categories can be ordered or ranked. Thus, their values were arranged in a specific order according to their meaning and given numbers as an indication to their order among the other values for the same category.

1. Education_Level
   - Unknown: 1
   - Uneducated: 2
   - High School: 3
   - College: 4
   - Graduate: 5
   - Post-Graduate: 6
   - Doctorate: 7
2. Income_Category
   - Unknown: 1
   - Less than $40K: 2
   - $40K - $60K: 3
   - $60K - $80K: 4
   - $80K - $120K: 5
   - $120K +: 6

5.1.2.3. Nominal Categorial Variables

These are variables that contain categories that can't be ordered or ranked. It would have no meaning, if they have been given weights like ordinal variables above and the results might be inaccurate because of those weights. Therefore, making each category a new binary variable reveals the effect of that category only all over the whole observations. Those nominal categorical variables were:

1. Marital_Status
   - Marital_Status_Unknown (New Binary Variable)
   - Single (New Binary Variable)
   - Married (New Binary Variable)
   - Divorced (New Binary Variable)
2. Card_Category
   - Blue (New Binary Variable)
   - Gold (New Binary Variable)
   - Silver (New Binary Variable)
   - Platinum (New Binary Variable)

After each category have been converted into a numeric binary variable the original variable containing the full categories were removed.

5.2. Prediction Models

Different prediction models were used to predict which customers are attrited and which are existing customers so that they can help us to predict the probability of customer being attrited in the future and use this information in the needed aspects.

These are three different prediction models that we experimented and performed on the dataset where each model uses a different technique. Therefore, each one has different accuracy and thus at the end they will be compared together.

In order to achieve that goal, the dataset is divided randomly into training data and testing data (with a split ratio of 0.7); the models are then built using the training data and tested using the testing data. After that, the accuracies have been computed.

Regarding our prediction objective, we are more concerned to predict the attrited customers (predict 1 more accurately than 0), thus we will choose the range of thresholds appearing in the ROC curve with relatively high true positive rate and at the same time not too large false positive rate. Then, within that range we will choose the threshold with the highest accuracy.

5.2.1. Logistic Regression Model

5.2.1.1. Logistic Regression Model Building

Firstly, a logistic regression model is built using all the variables. And the AIC value has been recorded.

AIC of the model: 3333.1

Therefore, we have decided to find out the correlation between the variables by removing a variable each time and notice the change in the AIC value. We started by removing the variables with the highest p-value and noticed the improvement in the AIC value. After this methodology was implemented, we have found that the best AIC is reached after removing the following insignificant variables:

- Marital_Status_Unknown
- Education level
- Single
- Customer_Age
- Platinum
- Divorced
- Gold
- Credit Limit
- Avg_Utilization_Ratio

AIC of the model: 3321.7

The significant variables are shown in table (1). It shown that some variables have positive coefficients and others have negative ones. All the variables that have positive coefficients mean that as the value of the variable increases, this indicates the increase in the probability that the predicted customers to be 1 (which is attired customer). For gender variable since it is a binary variable with 1 meaning female, this means that if a customer is female, she has higher probability of being attrited

For the variables that have negative coefficients, as their value decreases, the model is more likely to predict 1 (attrited customer). For the married variable, unmarried customers have higher probability of being attrited. For the blue variable, customers that has no blue cards are more likely to be attrited.

Table (1): Logistic Regression Model Significant Variables

Variable Name	Coefficients	Significance Code	p-value
Gender	1.091	***	1.50^{-12}
Dependent Count	0.1369	***	0.000128
Income Category	0.1920	***	0.000362
Total Relationship Count	-0.4500	***	$< 2^{-16}$
Months_Inactive_12_mon	0.4886	***	$< 2^{-16}$
Contacts_Count_12_mon	0.5383	***	$< 2^{-16}$
Total Revolving Balance	-0.0009798	***	$< 2^{-16}$
Total Transaction Amount	0.0004576	***	$< 2^{-16}$
Total Transaction Count	-0.1162	***	$< 2^{-16}$
Total_Ct_Chng_Q4_Q1	-3.017	***	$< 2^{-16}$
Married	-0.5597	***	2.22^{-09}
Months_on_book	-0.01333	*	0.018778
Blue	-0.8489	*	0.025503

5.2.1.2. Logistic Regression Model Prediction Accuracy

According to the ROCR curve shown in figure (20), table (2) summarizes the accuracy of the logistic regression prediction model at thresholds values between 0.1 and 0.3 (Suitable for our objective stated above. It is shown that a threshold of 0.3 results in the highest accuracy. Thus, table (3) shows the confusion matrix of the logistic regression model at threshold of 0.3.

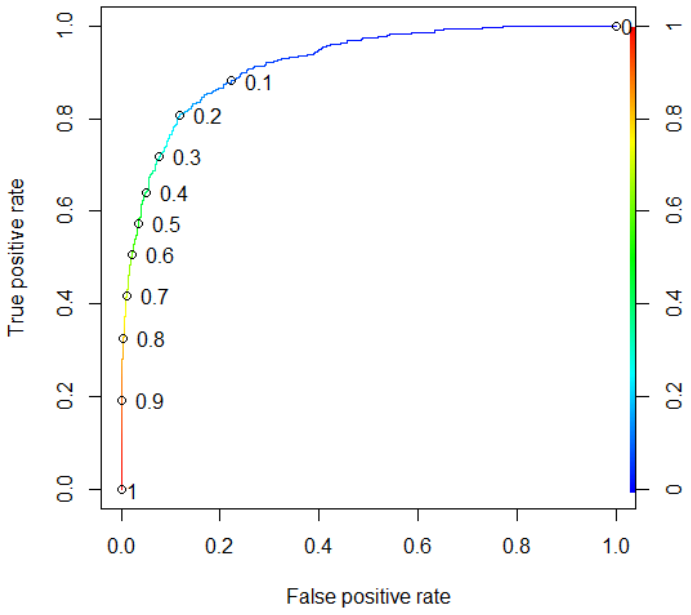


Figure (20): ROCR Curve for Logistic Regression Model.

Table (2): Model Accuracy at Different Thresholds for The Logistic Regression Model

Threshold	Accuracy
0.1	0.7943
0.2	0.8697
0.3	**0.8907**

Table (3): Confusion Matrix at Threshold of 0.3 for The Logistic Regression Model

	FALSE	TRUE
0	2355	195
1	137	351
Accuracy	**0.8907**	

5.2.2. Classification and Regression Tree (CART) Model

5.2.2.1. CART Model Building

There are some parameters that can be determined in order to enhance the accuracy of the CART model. The most commonly used parameter is the minbucket; which is the parameter that determine the minimum number of observations at each split. However, complexity parameter (cp) shows a very high effectiveness for the optimization of the CART model. The reason for that is that the (cp) ignore the number of splits that does not decrease the model lack of fit by that (cp) value.

From the code output and as shown in figure (21) a (cp) of 0.002 is the best to be used in our model. Therefore, the model is built again after adding the cp parameter equal to 0.002.

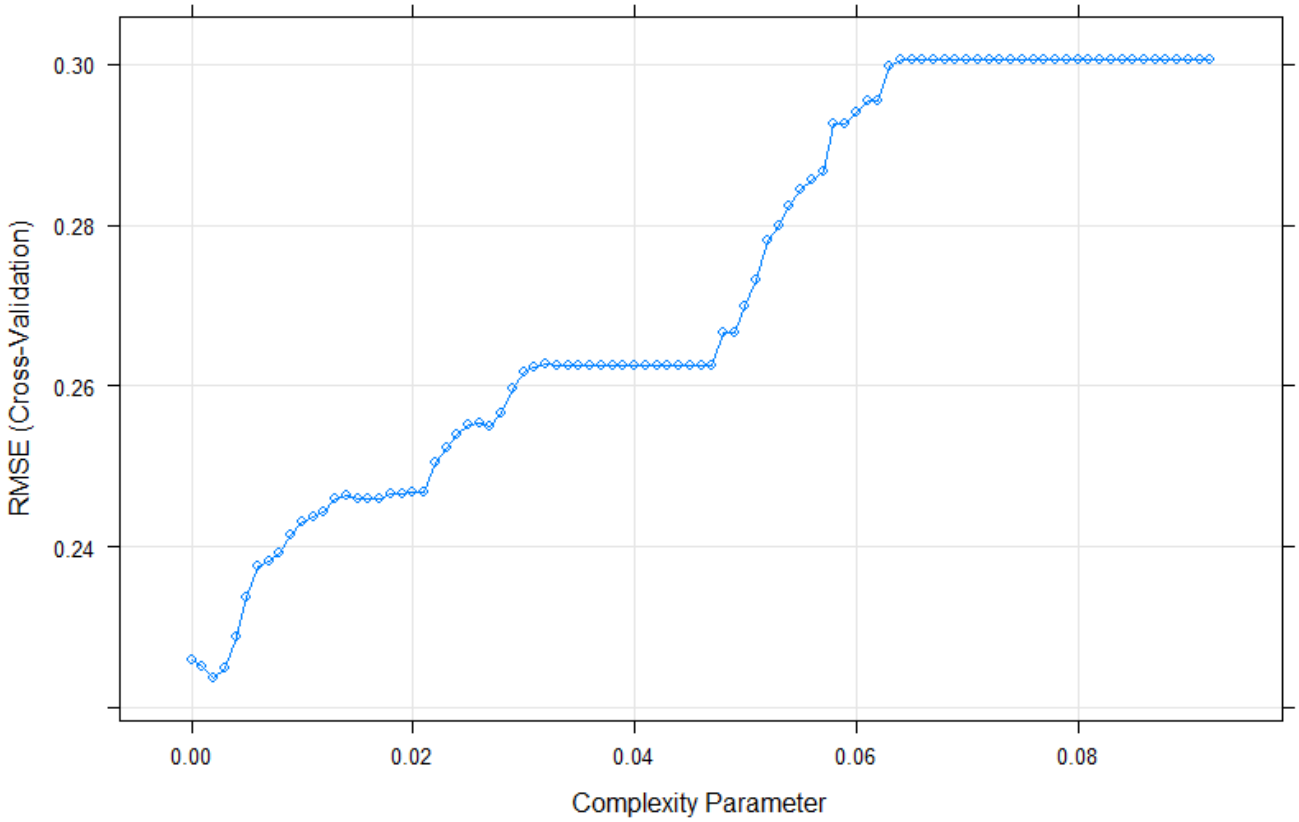


Figure (21). Complexity Parameter Vs. RMSE for The CART Model

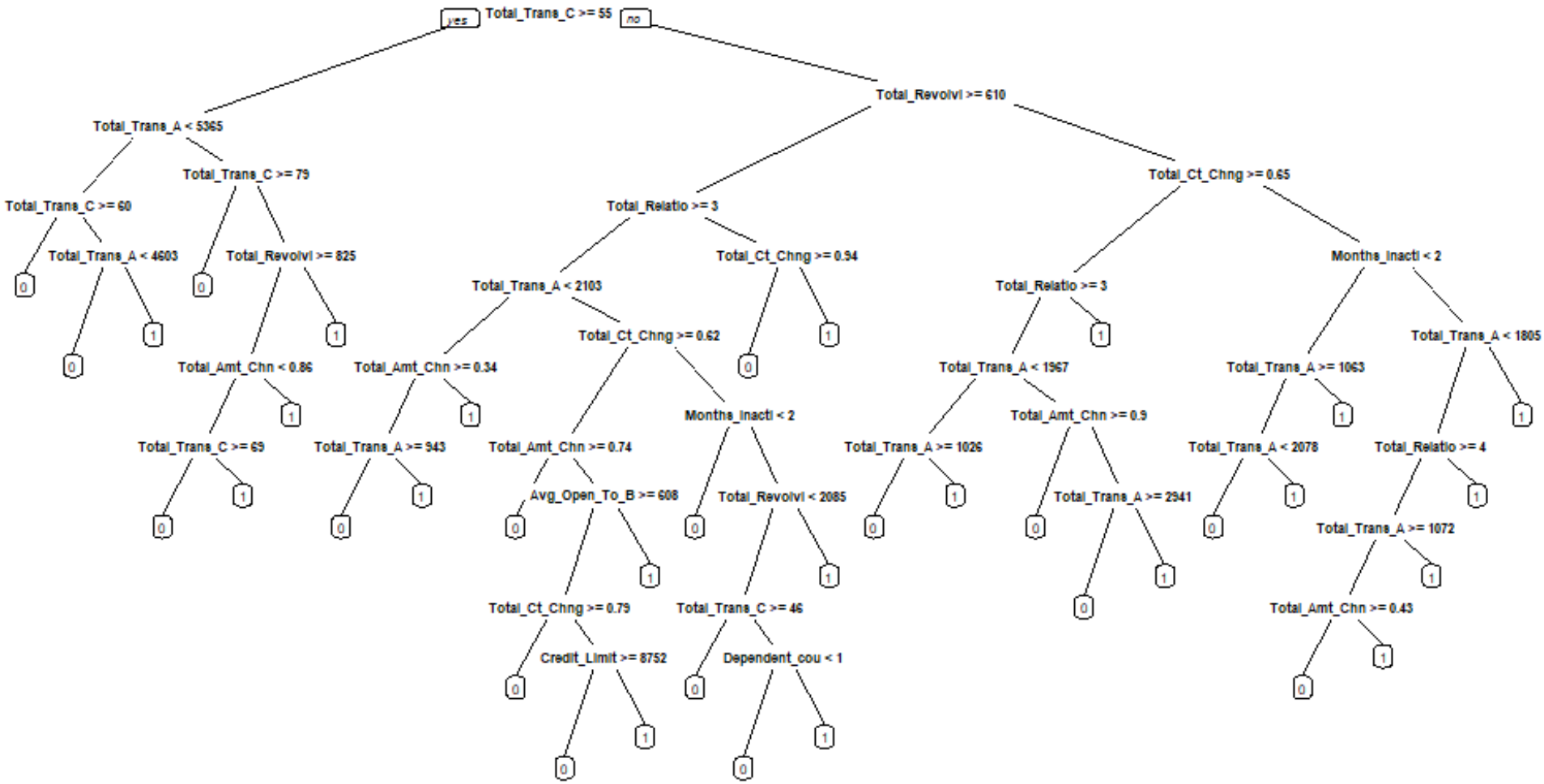


Figure (22): Shows The Tree of The CART Model Which Shows The Significant Variables of The Model.

5.2.2.2. CART Model Prediction Accuracy

According to the ROCR curve shown in Graph (), thresholds between 0.1 and 0.7 matches our objective where 0.3, 0.4, and 0.5 show the same accuracy. Also, thresholds of 0.6 and 0.7 show the same result. Therefore, they are not included in table (4) which summarizes the accuracy of the CART prediction model at each threshold.

It is shown that a threshold of 0.6 results in the highest accuracy. Thus, table (5) shows the confusion matrix of the CART prediction model at a threshold of 0.6.

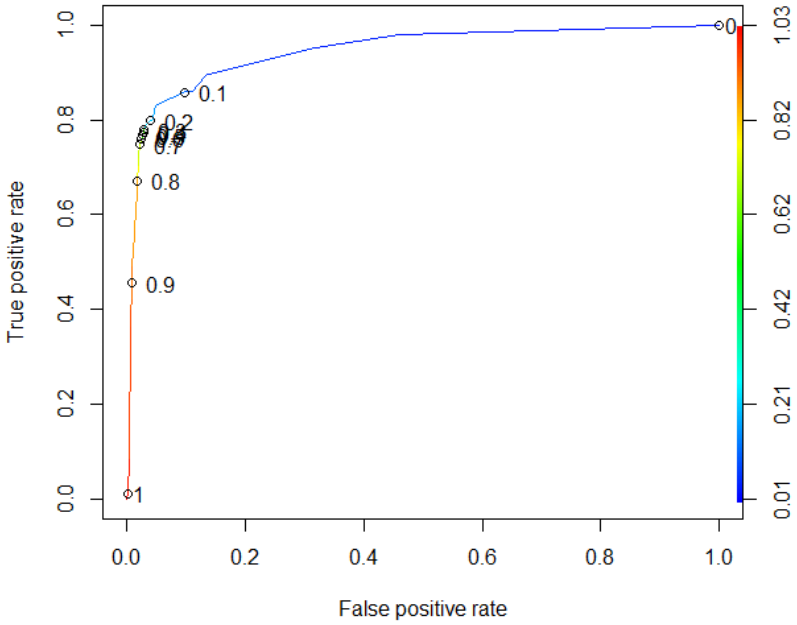


Figure (23): ROCR Curve for CART model.

Table (4): Accuracies Corresponding to Different Thresholds for the Cart Model

Threshold	Accuracy
0.1	0.9312
0.2	0.9361
0.3	0.9397
0.6	**0.9414**

Table (5): Confusion Matrix at Threshold 0.6 for The CART Model

	FALSE	TRUE
0	2489	61
1	117	371
Accuracy	**0.9414**	

5.2.3. Random Forest Model

5.2.3.1. Random Forest Model Building

There are some parameters needed to be determined for the random forest model in order to get higher accuracy. One of these parameters is the number of trees used to build the random forest.

Figure (24) trains different number of trees at the random forest model and plots the number of trees with the corresponding value of the mean square error. It is shown that number of trees equal to 500 has the lowest mean of squared residuals. Therefore, and argument has been added to the random forest model specifying the number of trees to be 500.

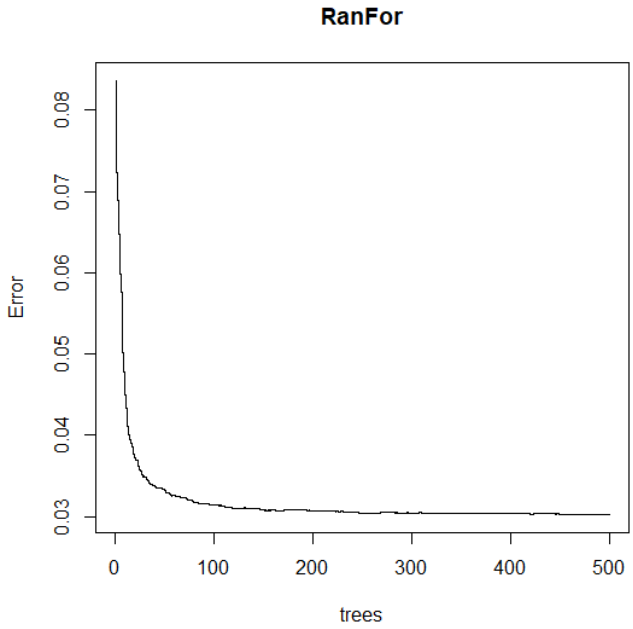


Figure (24): Number of Trees for The Random Forest Model Vs. Corresponding Error.

5.2.3.2. Significant Variables for The Random Forest Model

Figure (25) shows the %incMSE for each variable used in the random forest mode. %incMSE shows the percentage increase in the model mean squared error as a result of removing a specific variable. Therefore, the variables which show high %incMSE are very important for the random forest model. It is shown that the variables are ordered regarding their importance from the top down; thus, the most important variable is the one that is on the top.

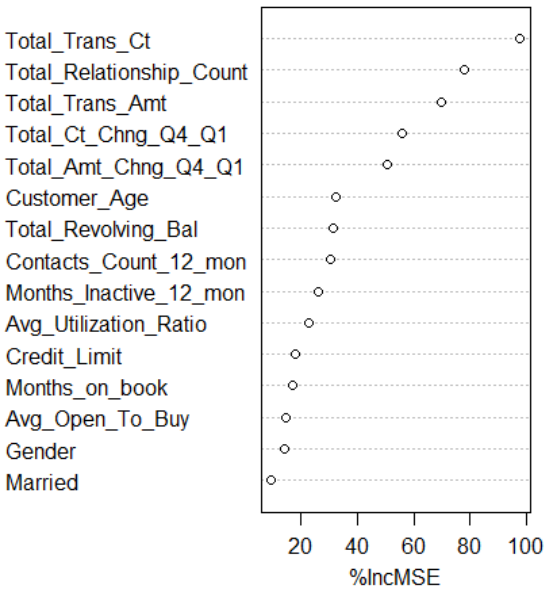


Figure (25): % Increase in MSE When The Corresponding Variable is Removes

5.2.3.3. Random Forest Model Prediction Accuracy.

Table (6) is built to show the corresponding accuracies for different threshold values between 0.1 and 0.5 (to match our objective). In accordance, table (7) shows the confusion matrix of the random forest prediction model at a threshold of 0.6

Table (6): Accuracies Correspond to different Thresholds of Random Forest

Threshold	Accuracy
0.1	0.8818
0.2	0.9358
0.3	0.9556
0.4	0.9641
0.5	**0.9674**

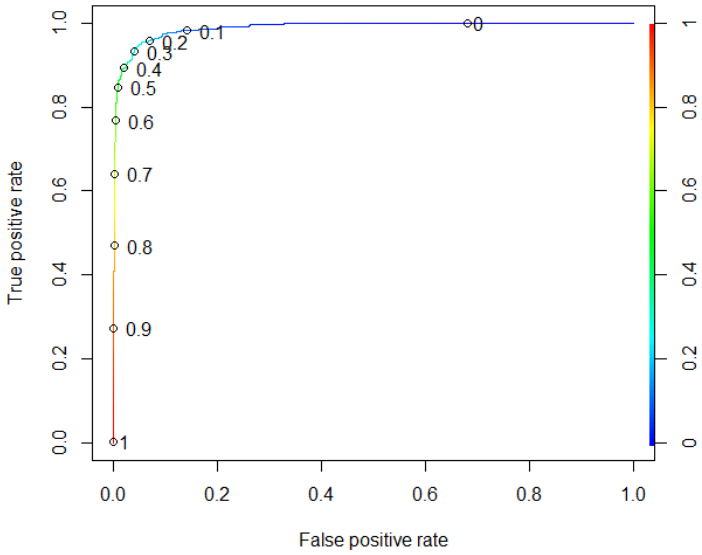


Figure (26): ROCR Curve for Random Forest

	FALSE	TRUE
0	2525	25
1	75	414
Accuracy	**0.9674**	

Table (7): Random Forest Model Confusion Matrix at threshold 0.5

5.2.4. Results

Table (8) summarize the overall accuracy for each prediction model. It is obvious that the random forest model produces the highest accuracy, followed by CART and then Logistic Regression comes at the end. It is also shown that the Random forest model has the highest sensitivity and specificity which means that it predict the attired customers with high accuracy (which are true positives) and also highly predicts the existing customers (which are true negatives).

Table (8): Overall Comparison between all the Prediction Models

	Logistic Regression	CART	Random Forest
Baseline Accuracy	0.8393		
Threshold	0.3	0.6	0.5
Accuracy	0.8907	0.9414	0.9681
Sensitivity	0.7193	0.7602	0.8484
Specificity	0.9235	0.9761	0.9902

From table (8) it is seen that all the models having sensitivity lower than specificity, in other words they predict the true positives (attrited customers) correctly with lower accuracy than that with true negatives (Existing Customers) and if lower threshold value is chosen it would result in lower overall accuracy and increased false positive rate, so it is a tradeoff and a clear objective should be decided for the prediction model.

5.3. Customer Segmentation using Clusters

Segmentation is a technique used to divide customers into groups based on attributes such as behavior or demographics. It is useful to identify segments of customers who may respond in a similar way to specific marketing techniques such as promotions, offers and advertising. As it gives businesses the ability to tailor marketing campaigns and timing to generate better response rates and provide improved consumer experiences.[7] Also, segmenting customers into groups help us better understand our customer base and be better able to take appropriate decisions regarding each segment as we will see in the analysis of our customer segmentation

5.3.1. Normalization

The variables having a maximum value more than 1 were normalized using min_max method

5.3.2. K-means Clustering

All variables (except for the variables removed in the data preparation section) were used in the clustering. Different number of clusters were tested to know the number of clusters that will be used: After plotting the total within cluster sum of squares against the number of clusters as shown in figure (27), 8 clusters were found appropriate since increasing the number of clusters above 8 will not cause a significant improvement in the total within cluster sum of squares. The 8 clusters are summarized in table (9)

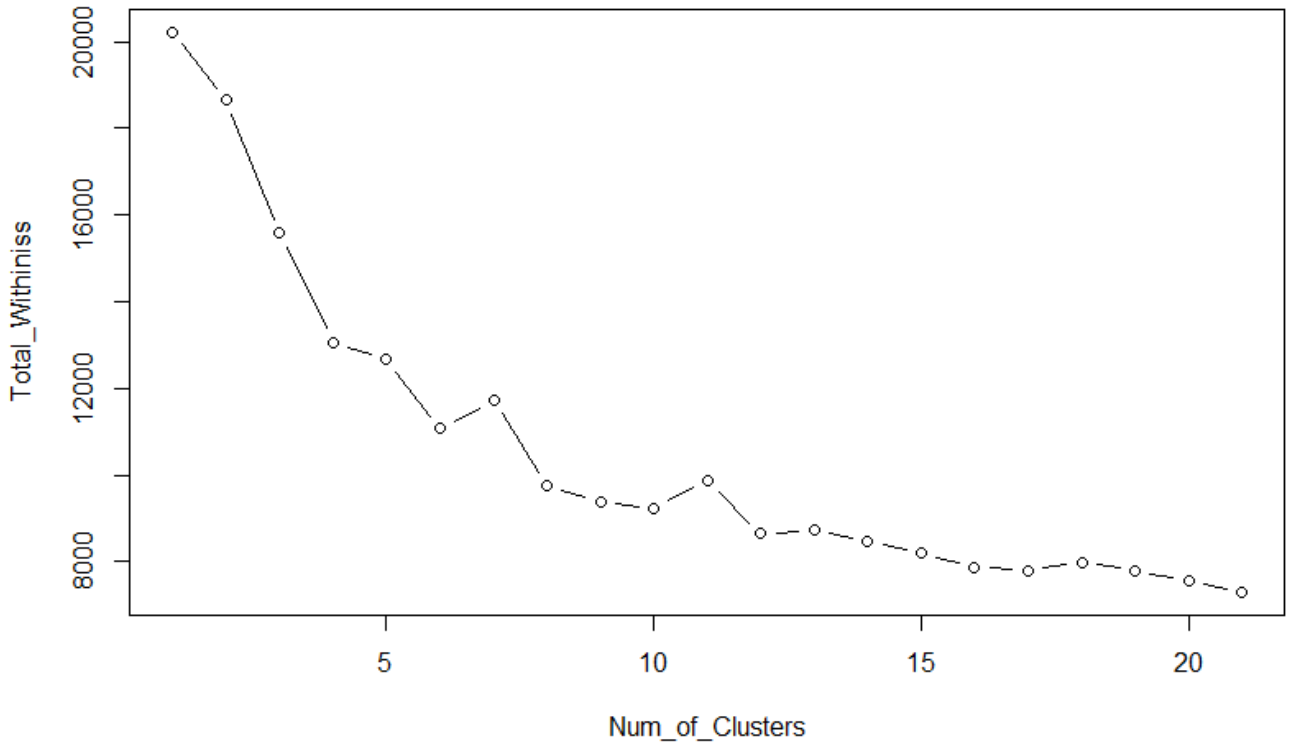


Figure (27): Total Within Cluster Sum of Squares Plotted Against The Number of Clusters

Table (9): Customer Segments with Their Behaviors and Profiles

Size	2020 (19.9%) 2nd Largest Customer Segment	750 (7.4%) 4th Smallest Customer Segment	638 (6.3%) Smallest Customer Segment	1633 (16.1%) 3rd Largest Customer Segment	692 (6.8%) 2nd Smallest Customer Segment	2068 (20.4%) Largest Customer Segment	705 (7.0%) 3rd Smallest Customer Segment	1621 (16.0%) 4th Largest Customer Segment
Attrition Flag	365 Attrited (18%) 1655 Existing (82%)	332 Attrited (44%) 418 Existing (56%)	95 Attrited (15%) 543 Existing (85%)	77 Attrited (5%) 1556 Existing (95%)	121 Attrited (17%) 571 Existing (83%)	273 Attrited (13%) 1795 Existing (87%)	114 Attrited (16%) 591 Existing (84%)	250 Attrited (15%) 1371 Existing (85%)
Contribution to Total Attritions	22.4%	20.4%	5.8%	4.7%	7.4%	16.8%	7%	15.4%
Customer Age	46	47	46	47	46	47	45	46
Gender	Females	745 Females (99.3%) 5 Males (0.7%)	210 Females (33%) 428 Males (67%)	Females	361 Females (52%) 331 Males (48%)	Males	389 Females (55%) 316 Males (45%)	Males
Dependent Count	2.3	2.4	2.5	2.3	2.6	2.4	2.4	2.2
Education Level	94 Doctorate (5%) 92 Post- Graduate (5%) 615 Graduate (30%) 193 College (10%) 389 High School (19%) 306 Uneducated (15%) 331 Unknown (16%)	36 Doctorate (5%) 38 Post-Graduate (5%) 226 Graduate (30%) 75 College (10%) 147 High School (20%) 110 Uneducated (15%) 118 Unknown (16%)	24 Doctorate (4%) 37 Post-Graduate (8%) 213 Graduate (33%) 68 College (11%) 117 High School (18%) 87 Uneducated (14%) 92 Unknown (14%)	84 Doctorate (5%) 86 Post-Graduate (5%) 526 Graduate (32%) 160 College (10%) 322 High School (20%) 225 Uneducated (14%) 230 Unknown (14%)	24 Doctorate (3%) 39 Post-Graduate (6%) 206 Graduate (30%) 71 College (10%) 143 High School (21%) 99 Uneducated (14%) 110 Unknown (16%)	77 Doctorate (4%) 107 Post-Graduate (5%) 645 Graduate (31%) 206 College (10%) 441 High School (21%) 292 Uneducated (14%) 300 Unknown (15%)	35 Doctorate (5%) 40 Post-Graduate (6%) 208 Graduate (30%) 84 College (12%) 123 High School 17%) 124 Uneducated (18%) 91 Unknown (13%)	77 Doctorate (5%) 77 Post-Graduate (5%) 489 Graduate (30%) 156 College (10%) 331 High School (20%) 244 Uneducated (15%) 247 Unknown (15%)
Marital Status	Single	Married	43 Divorced (7%) 236 Married (37%) 302 Single (47%) 57 Unknown (9%)	Married	Unknown	Married	Divorced	Single
Income Category	367 in $40K - $60K (18%) 1273 in less than $40K (63%) 380 Unknown (19%)	133 in $40K - $60K (18%) 441 in less than $40K (59%) 176 Unknown (23%)	81 in $120K + (13%) 140 in $80K - $120K (22%) 128 in $60K - $80K (20%) 108 in $40K - $60K (17%) 117 in less than $40K (18%) 64 Unknown (10%)	316 in $40K - $60K (19%) 1030 in less than $40K (63%) 287 Unknown (18%)	45 in $120K + (7%) 121 in $80K - $120K (17%) 88 in $60K - $80K (13%) 123 in $40K - $60K (18%) 241 in less than $40K (35%) 74 Unknown (11%)	315 in $120K + (15%) 689 in $80K - $120K (33%) 613 in $60K - $80K (30%) 330 in $40K - $60K (16%) 109 in less than $40K (5%) 12 Unknown (1%)	48 in $120K + (7%) 94 in $80K - $120K (13%) 99 in $60K - $80K (14%) 126 in $40K - $60K (18%) 247 in less than $40K (35%) 88 Unknown (12%)	238 in $120K + (15%) 491 in $80K - $120K (30%) 474 in $60K - $80K (29%) 284 in $40K - $60K (18%) 103 in less than $40K (6%) 31 Unknown (2%)
Card Category	2003 Blue (99%) 14 Gold (1%) 3 Platinum (0.15%)	740 Blue (99%) 5 Silver (0.7%) 4 Gold (0.5%) 1 Platinum (0.13%)	540 Silver (85%) 83 Gold (13%) 15 Platinum (2.4%)	1628 Blue (99.7%) 5 Gold (0.3%)	683 Blue (98.7%) 5 Silver (0.7%) 4 Gold (0.6%)	2065 Blue (99.9%) 3 Gold (0.1%)	696 Blue (98.7%) 5 Silver (0.7%) 3 Gold (0.4%) 1 Platinum (0.14%)	1621 Blue (100%)
Months on Book	36	36	35	36	35	36	35	35
Total Relationship Count	3.8	3.7	3.3	3.9	3.8	3.9	3.9	3.8
Months Inactive in last year	2.4	2.5	2.3	2.3	2.3	2.3	2.4	2.3
Contacts Count in Last Year	2.5	2.5	2.5	2.4	2.4	2.5	2.4	2.5
Credit Limit	$4,418	$4,391	$26,827	$4,136	$7,969	$10,424	$8,334	$11,334
Total Revolving Balance	$1,107	$144	$1,229	$1,594	$1,171	$1,254	$1,151	$1,127
Average Open to Buy	$3,311	$4,247	$25,597	$2,542	$6,798	$9,170	$7,183	$10,213
Total Transaction Amount	$4,354	$3,607	$6,898	$4,196	$4,521	$4,003	$4,414	$4,521
Total Transaction Count	68	58	77	65	66	58	67	66
Total Amount Change between 4th and 1st Quarter	0.74	0.73	0.77	0.78	0.75	0.78	0.76	0.75
Total Count Change between 4th and 1st Quarter	0.71	0.65	0.71	0.75	0.72	0.72	0.71	0.71
Average Utilization Ratio	0.35	0.04	0.05	0.51	0.27	0.23	0.27	0.20

5.3.3. Segmentation Analysis and Recommendations:

5.3.3.1. Clusters mainly don't differ much based on customer's age, educational level, dependent count, months on book, total relationship count, months inactive and contacts count. Thus, we won't use these factors in our analysis of customer segmentation and will depend only on other factors (bold factors in table (9)) where customer segments are exhibiting different behaviours

5.3.3.2. Cluster 2 has the least total transaction amount among the other segments so promotions such as low interest rates, vouchers and discounts may be offered to this group of customers to encourage their use of credit card. However, those customers have very low utilization ratio (below 1% is unrecommended). This low utilization ratio may imply inactive accounts as users are not making use of the credit limits given to them. Additional investigations may be needed to understand why this particular set of consumers are not utilizing their lines and whether promotions are needed or if their credit lines could in the future be assigned to a different set of consumers. If an account is inactive, the bank makes no money from transaction fees paid by merchants or from interest rate. Thus, it may be a better option for the bank to exploit the unused credit limit to extend to their customers by cancelling inactive accounts and giving that line of credit to someone who will use it.

5.3.3.3. Customers in cluster 3 have very low utilization ratio as well but they have the highest total transaction amounts and thus are less likely to be prone to have their credit card accounts being cancelled from the side of the bank as in the case of cluster 2 above. Customers in this segment are paying their depts but, they are not making use of the high credit limit given to them. Bank may decrease (not cancel) their credit limit to make use of these credit limits with other customers or they may encourage customers in this segment to use their credit cards more frequently (and also exploit the benefit that these customers belong to medium to high income groups) by offering them more rewards and benefits like vouchers, discounts and low interest rates.

5.3.3.4. Custer 5, 6, 7 and 8 have **good utilization ratios** (below 30% is recommended). Thus, it can be deducted that they use their Credit Cards effectively and are paying their depts and thus they pose little risk of being financially exhausted since low credit utilization ratio is considered an indicator that a customer is doing a good job of managing their credit responsibilities and they're far from overspending. Rewards, benefits, or Loyalty Program may keep them on using the Credit Card.

5.3.3.5. Cluster 1 and 4 have **high utilization ratios** (above 30%) which may mean that they pose higher risk of being unable to pay their depts. Further Analysis for customers in these 2 segments will be needed to decide whether their credit card and loan applications should be denied because they might be financially unable to pay their depts. Also, higher interest rates can be applied on these customers so that they reduce their reliance on credit and pay the depts they owe in order not to accumulate more depts (putting penalty on depts). Another possible reason for the high utilization ratio is that they might be in need to increase their credit limits but the bank must be cautious and further analysis and investigation regarding these customers' history and financial ability would be needed so that the bank can make sure that these customers have no risk of converting the new credit limit into depts!

5.3.3.6. Customers in cluster 2 pose a high risk of being attrited soon since the segment contains the **highest attrition percentage** within its customers. Also, Cluster 1 poses the highest number of attrited customers among all other customer segments (22.4% of total attritions). Customers in Cluster 6 and 8 also pose higher risk of attritions than others since they contribute to the percentage of total

attritions with 16.8%, and 15.4% respectively. The bank may decide to proactively go to the existing customers in those segments to provide them better services and work on turning customers' decisions in the opposite direction. Interest rates may be lowered to them. Rewards, vouchers, customized discounts and other benefits and promotions may be offered. However, **caution** must be considered towards customers in cluster 1 because they already have high utilization ratio as discussed above. Thus, further detailed analysis for customer profiles in this segment might be needed before any decision can be taken.

5.3.3.7. **Total transaction amount change** between 4th quarter and 1st quarter for all segments is less than 1 which means that credit card customers are using their credit cards more in the 1st quarter than that on the 4th quarter which may show the need to increasing the promotions for all customers at the 4th quarter of the year to boost its usage at the end of the year and to have more stable use of the credit card all over the year.

5.3.3.8. Demographic variables used in the segmentation may be used to better customize the promotions and offers to their target segments (e.g. offering discounts or vouchers on specific products when using the credit card on purchases to encourage the use of credit card).

5.3.4. Rebuilding Prediction Models for each Customer Segment

The same methodology of building the prediction models in section 5.2 were performed on each customer segment to be able to predict whether a customer belonging to a specific customer segment is an attrtited customer or existing customer

Table (10): Prediction Model Accuracies Among All Clusters

		Clusters							
Model	Model Accuracy and corresponding Threshold	1	2	3	4	5	6	7	8
Baseline Accuracy		0.8187	0.5555	0.8489	0.9531	0.8260	0.8679	0.8388	0.8456
Logistic Regression	AIC	555.12	314.08	228.12	199.55	235.36	722.19	214.1	554.67
	Threshold	0.4	0.4	0.2	0.4	0.25	0.3	0.3	0.5
	Accuracy	0.9159	0.8577	0.9018	0.9613	0.8985	0.9001	0.8815	0.8991
	Sensitivity	0.8	0.93	0.6551	0.6956	0.7777	0.6341	0.6176	0.6133
	Specificity	0.9416	0.904	0.9018	0.9743	0.9239	0.9406	0.9322	0.95
CART	Threshold	0.5	0.3	0.3	0.4	0.3	0.4	0.3	0.4
	Accuracy	0.9357	0.9466	0.9114	0.9674	0.9033	0.9114	0.8909	0.9094
	Sensitivity	0.7818	0.96	0.6551	0.6521	0.6111	0.5853	0.5588	0.7333
	Specificity	0.9698	0.936	0.9570	0.9829	0.9649	0.9610	0.9548	0.9635
Random Forest	Threshold	0.5	0.4	0.3	0.4	0.5	0.4	0.3	0.4
	Accuracy	0.9439	0.9333	0.9270	0.9796	0.9371	0.9436	0.9336	0.9506
	Sensitivity	0.8181	0.96	0.8275	0.6956	0.6944	0.7195	0.8235	0.8
	Specificity	0.9698	0.912	0.9447	0.9935	0.9883	0.9777	0.9548	0.9805

Comparing the results of the prediction models performed on the clusters given in table (10) and the prediction models performed on the whole data set given in table (8), we have found that:

1. Cluster 1 prediction models yielded better accuracies than that using the whole dataset. In addition, it yielded better sensitivity values (which is our main objective). It has also improved the specificity for all models except for the CART model that higher specificity when performed on the whole dataset.
2. Cluster 2 prediction model yielded lower accuracies than that using the whole dataset except for the CART model where cluster 2 Cart model yielded higher accuracy. However, cluster 2 prediction model yielded better sensitivity values (which is our main objective) for the logistic regression model and the CART model. Cluster 2 prediction models lowered the specificity values by small acceptable amounts
3. Cluster 3, cluster 6, cluster 7 and cluster 8 yielded lower values for the accuracies, sensitivity, and specificity for all the 3 models. Except that cluster 3 and 6 yielded a small improvement in the logistic regression model accuracy whereas cluster 7 improved the specificity in the logistic regression model. For cluster 8, it only improved the accuracy and specificity of the logistic regression model by a small value
4. Cluster 4 prediction models has yielded improvement in the overall accuracy and specificity but resulted in lower sensitivity values for all the 3 models applied.
5. Cluster 5 has yielded higher values for the accuracy, sensitivity, and specificity only in the logistic regression model but yielded lower values for these indicators for the other 2 models.
6. The random forest model has yielded the best results for all clusters except cluster 3 whose best results came from the CART model.
7. The AIC value for the logistic regression model of all cluster is much lower than that of the logistic regression model of the whole data set.

These results were somehow surprising as we supposed that clustering and grouping customers with common features together in the same group would enhance the accuracy for all clusters. However, there was some clusters with lower accuracies than the main prediction models performed on the whole dataset. The one common conclusion we can reach is that the percentage of attrited customers in each cluster largely affects the sensitivity values; if the cluster has higher percentage of attrited customers than the percentage of attrited customer in the whole dataset (16.06%), then this cluster shows improvement in the sensitivity values among all 3 models but this doesn't indicate that it will have higher accuracy, since it depends also on the specificity values.

6. Conclusion

From the data visualization part, we have found that for most of the variables the attrited and existing customers had the same patterns except for the numerical variables, however, most of the numerical variables also had the same distribution between the existing and attrited customers, we could spot some differences, which had a huge effect after building the models.

Three prediction models have been built in order to predict whether a customer is attrited or not. Those models were logistic regression model, CART model and random forest model. In order to guarantee a high accuracy of the model built, some methodologies have been followed and some

parameters have been added to the model. As a result, the random forest model yielded the best results.

Customer Segmentation using clustering has helped us better discover the customers of the bank, by looking through their behaviors and profiles while using Credit Card. It has also helped us to recommend different effective marketing campaigns or sales promotion to their targeted customers and be aware of customers segments posing higher risks than others. Also, better customized services can be offered to each segment based on our previous analysis

7. References

[1] S. Goyal, "Credit Card customers," 19-Nov-2020. [Online]. Available: https://www.kaggle.com/sakshigoyal7/credit-card-customers. [Accessed: 31-Jan-2021].

[2] Resources.display, "What is a Credit Utilization Rate?," Experian, 20-Oct-2020. [Online]. Available: https://www.experian.com/blogs/ask-experian/credit-education/score-basics/credit-utilization-rate/#:~:text=Your%20credit%20utilization%20rate%2C%20sometimes,generally%20expressed%20as%20a%20percent. [Accessed: 31-Jan-2021].

[3] L. T. Irby, "What Is the Credit Utilization Ratio?," The Balance, 07-Oct-2020. [Online]. Available: https://www.thebalance.com/what-is-a-good-credit-utilization-ratio-960548. [Accessed: 31-Jan-2021].

[4] "How Does the Credit Utilization Percentage Impact My Credit Score?," Credit Card Insider, 04-Nov-2020. [Online]. Available: . [Accessed: 31-Jan-2021].

[5] S. Rathner, "Credit Card Closed for Inactivity? What You Need to Know," NerdWallet, 22-Sep-2020. [Online]. Available: https://www.nerdwallet.com/article/credit-cards/credit-card-cancelled-due-inactivity#:~:text=When%20your%20account%20is%20idle,someone%20who%20will%20use%20it. [Accessed: 31-Jan-2021].

[6] L. DeNicola, "How Important Is Credit Card Utilization to Your Credit Score?," Experian, 09-Jun-2020. [Online]. Available: https://www.experian.com/blogs/ask-experian/how-important-is-credit-card-utilization/. [Accessed: 31-Jan-2021].

[7] R. Vickery, "Segmenting Credit Card Customers with Machine Learning," Medium, 24-May-2019. [Online]. Available: https://towardsdatascience.com/segmenting-credit-card-customers-with-machine-learning-ed4dbcea009c. [Accessed: 31-Jan-2021].