Credit Card Customers and Churn Rate

Erin Scharnweber, Ashley Cicero, Gabriella Dicroce, Vedant Ghavate

San Diego State University, Fowler College of Business

Executive Summary

Modeling Goal: Every business wants to retain its customers, and most importantly prevent customer attrition. In the "BankChurners" data set provided by Kaggle, a bank is experiencing attrition for its credit card customer users. The objective of our research is to identify and visualize which of these factors are contributing to customer churn. Our hypothesis is that total transaction amount and the total number of transactions will have a large impact on customer attrition. The analysis presented herein will use four prediction models to analyze customer churn and use machine learning techniques to ensure the highest prediction accuracy.

Data Preparation: The datafile BankChurners.csv contains 1,000 bank clients and includes various demographic information including gender, income level, and education level, as well as the behavior of those customers including the number of transactions the customer made during a certain period of time and the transaction amounts. Details concerning the data selection are provided in Section 2.

Modeling Methodology: The dependent variable that is being predicted is attrition. The independent variables that we found to influence attrition are described in Table 1. A series of regression and decision tree models were developed to determine which variables had the highest correlation to customer attrition. The four predictive models that we used are linear regression, logistic regression, simple decision tree, and bagged decision tree. The data was tuned and dummy coded to fit these individual models and determine if a relationship exists between the different variables. To determine the accuracy of each model, cross validation was completed by dividing the data into training and test data using the caret package (see Appendix 2 for R code). Details concerning the modeling and selection are presented in Section 3.

Results and suggested actions: The results of this investigation are summarized in Figure 1. As Table 1 in Section 3 demonstrates, the most significant variable in determining customer churn is the total transaction amount from the customer, followed by the total revolving balance. Table 1 also shows that total transaction amount and the total number of transactions have a high significance on customer attrition with an overall score of 1,039.32 and 711.71 respectively.

The highlighted blue section in Figure 1 below shows the total transaction count and transaction amount of attrited customers, while the orange highlighted section shows the total transaction count and transaction amount of existing customers. The attrited customers have a higher transaction amount during transaction count from 30-50. However, after 50 transactions, attrited customers will show decline in transaction amount. In comparison, existing customers will show a five-fold increase after a total of 50 transactions during the same period. It is important to note that existing customers have a higher possibility of attrition if the customer has less than 50 transactions which is important in customer retention. This will be discussed further in Section 5.

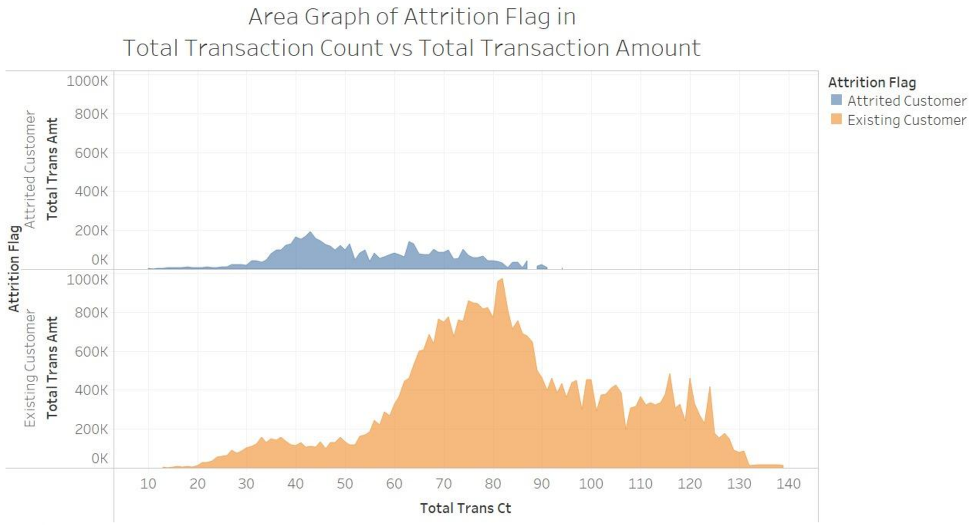


Figure 1: Plot of sum of total transaction amount for total transaction count broken down by attrition flag. Color shows details about attrition flag.

2. Discovery and Data Preparation

2.1 Finding a Data Set

The data set for this investigation was obtained through processing the parent data set. The parent data set is “Credit Card Customers”, which was obtained from https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers. The goal of this dataset is to address the issue of credit attrition, meaning the loss of customers, for a bank. This data set can be analyzed to draw conclusions about what factors are most impactful in the customer's decision to drop the credit card/bank; it will also allow the bank to provide better customer service to these customers in hopes of swaying customers' decisions in the opposite direction.

2.2 Preprocessing of Parent Data Set

Data preprocessing is a necessary step in order to prepare the raw data to be suitable for running models and machine learning. All of the processing of the parent data and data selection described here was performed with RStudio. The R code is included in the Appendix 1. The R code listed reduces the complete data set into a new data set which has been tuned to include only relevant data that minimizes error and or increases model accuracy.

Remove unnecessary features. The first step of preprocessing was to remove unnecessary columns for the modeling. The parent data set consists of 23 columns and 10,127 rows. Of the 23 columns, four of the columns are deemed irrelevant to this investigation. These columns were omitted as these variables contained a correlation higher than 0.80. Some of the variables omitted include months on book (as correlation with credit limit is 0.79), average open to buy (as correlation with credit limit is 0.995), and total transfer count (as correlation with total transfer amount is 0.807). All other variables have a correlation below the threshold of 0.78 (please reference Appendix 1).

Consideration of missing values. By utilizing the "plot missing" function in R, we were able to confirm and create a graph confirming there are zero missing values in all of the variables. See Figure 2 for confirmation of no missing values.

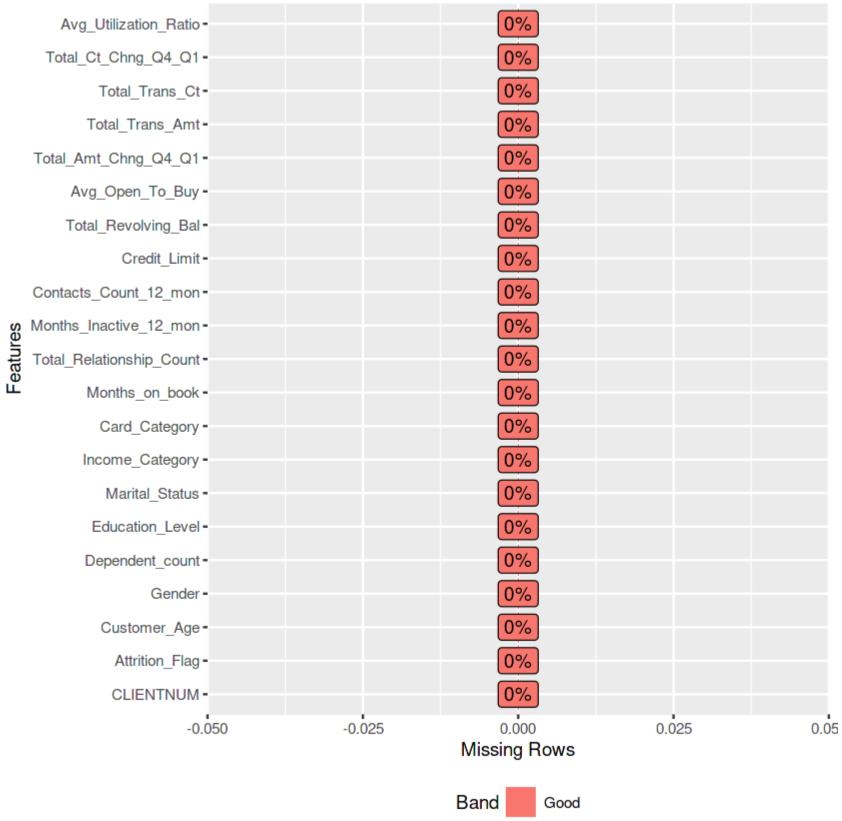


Figure 2: Plot of missing values

Data type conversions. Because some of the variables contained categorical data, the variables needed to be converted into numerical values which was accomplished by fast dummy coding. Some of the categorical variables include gender, marital status, education level, age range, and card category. To study the hypothesis, it is necessary to convert the variables into numerical values in order to determine any correlation and importance of the variables. The data analysis was performed with the R code shown in Appendix 1.

Final Data Set. A review of distributions and correlations is shown in Figure 3. From this figure, it can be concluded that there is no linear relationship between the independent and dependent variables. The accuracy of the linear regression model is 37% and therefore is not an accurate predictor of the hypothesis.

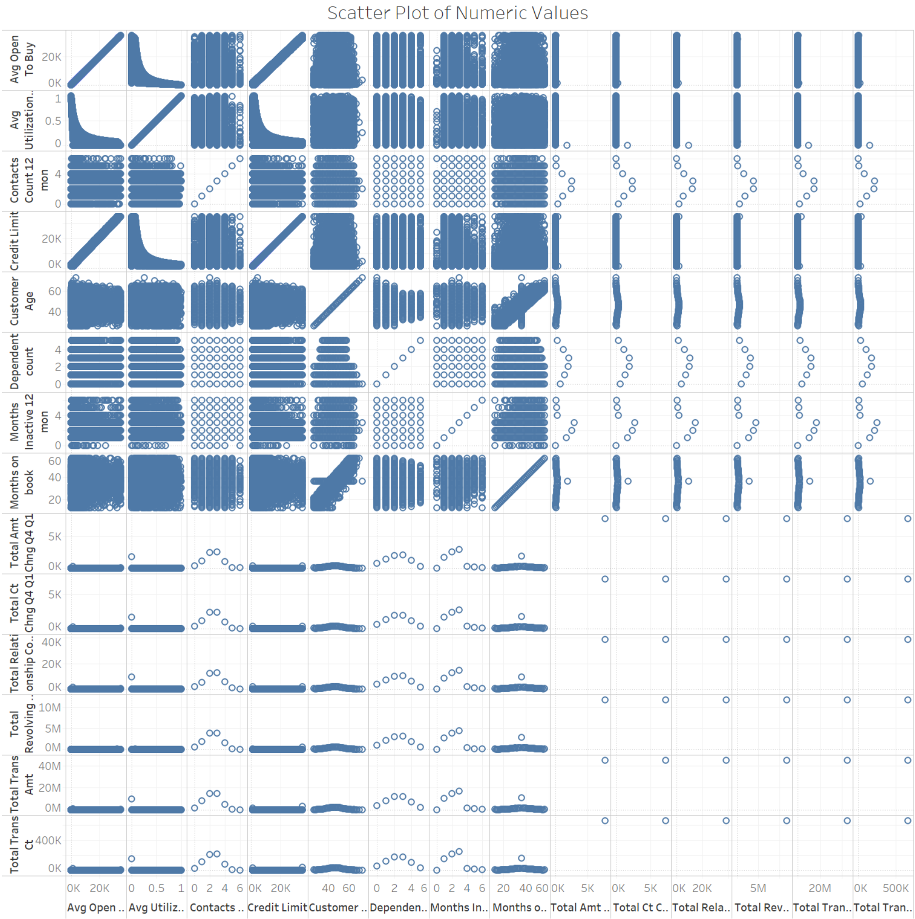


Figure 3: Scatterplot of all numeric columns in dataset

2.3 Exploration of Final Data Set

With the working data set in hand, it is useful to review the basic characteristics and relationships between the variables. The R code related to the visualizations is included in Appendix 2.

Customer demographic relationships. Figure 4 shows the correlation between customer education level and customer attrition. The data set shows 70% of customers have formal education and 35% have a higher level of education. Additionally, customers that have churned are highly educated, as nearly 30% of churned customers have a graduate level education and nearly 19% have a post-graduate education level. While it appears there might be a correlation between education level and customer churn, the bagged decision tree identified this as an insignificant relationship in comparison to the other variables which will be discussed further in Section 3 and Table 1.

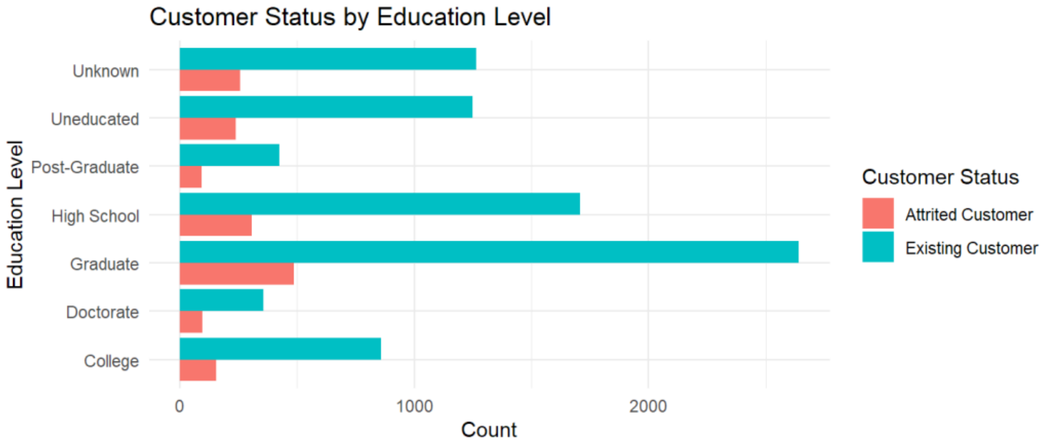


Figure 4: Bar graph showing relationship between education level and customer attrition

3. Model Planning and Building

3.1 Linear Regression

Linear regression is a useful tool to learn about the dataset and determine if the independent and dependent variables are correlated, and if there is a correlation if that is a positive or negative relationship. The results can provide broad insights regarding the hypothesis. Based on the model, we found there are no linear relationships between the independent variables and dependent variables. See Figure 3.

Preprocessing. Data preprocessing is the most important step in preparing the parent data set as this allows the data to be manipulated before the data is used. The prepossing that was conducted consisted of converting all categorical variables into numerical variables in order to be utilized in the regression analysis. The preprocessing was completed correctly in R.

Cross validation. Cross validation is a crucial test in determining the effectiveness of a predictive model and can help to identify overfitting and underfitting. In order to cross validate, the first set is to divide the data set into two sets - the training data set and the test data set which was completed using the caret package in R. Next, the model is trained using the training data set, and the model is validated on the test set.

3.2 Logistic Regression

Logistic regression is an important classification method that can be used to predict the outcome of a dependent variable based on previous observations. The R squared value explains what percent of the intercept can be explained by the predictors; 95% is the ideal value. Based on the model, the R squared value was 42% which is not enough to explain the predictor so this logistic regression cannot be considered an appropriate model for this data set.

3.3 Simple Decision Tree

Decision trees are useful modeling tools as they can organize complex data into more easily understandable categories or subsets of data. A simple decision tree is generated with the rpart package. The cross validation and resulting tree are shown below in Figure 5 and 6. Cross validation was used to validate the model. The three most relevant variables in order are total transaction count, total revolving balance, and total transaction amount. The decision tree uses total transaction count as the first separator and proceeds with revolving balance and total transaction amount as we go further down the tree. Finally, after filtering from many nodes, the customer age is used to make a leaf node of attrited customer among others.Refer to Appendix 3 for R code.

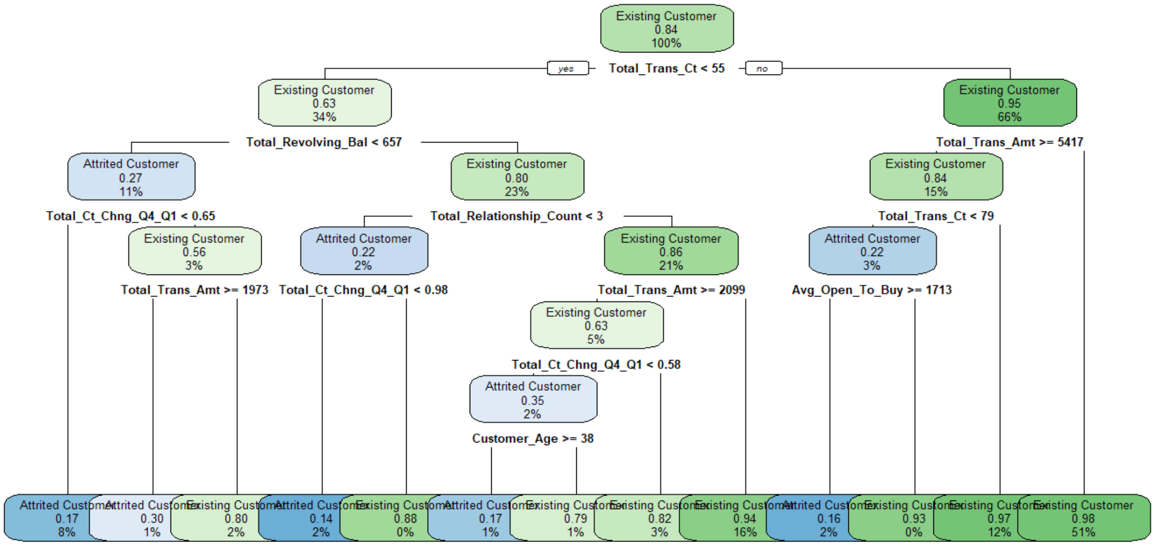

Existing Customer 0.84 100%
yes
Total_Trans_Ct < 55
no
Existing Customer 0.63 34%
Existing Customer 0.95 66%
Total_Revolving_Bal < 657
Total_Trans_Amt >= 5417
Attrited Customer 0.27 11%
Existing Customer 0.80 23%
Existing Customer 0.84 15%
Total_Ct_Chng_Q4_Q1 < 0.65
Total_Relationship_Count < 3
Total_Trans_Ct < 79
Existing Customer 0.56 3%
Attrited Customer 0.22 2%
Existing Customer 0.86 21%
Attrited Customer 0.22 3%
Total_Trans_Amt >= 1973
Total_Ct_Chng_Q4_Q1 < 0.98
Total_Trans_Amt >= 2099
Avg_Open_To_Buy >= 1713
Existing Customer 0.63 5%
Total_Ct_Chng_Q4_Q1 < 0.58
Attrited Customer 0.35 2%
Customer_Age >= 38
Attrited Customer 0.17 8%
Attrited Customer 0.30 1%
Existing Customer 0.80 2%
Attrited Customer 0.14 2%
Existing Customer 0.88 0%
Attrited Customer 0.17 1%
Existing Customer 0.79 1%
Existing Customer 0.82 3%
Existing Customer 0.94 16%
Attrited Customer 0.16 2%
Existing Customer 0.93 0%
Existing Customer 0.97 12%
Existing Customer 0.98 51%

Figure 5: Simple decision tree without pruning

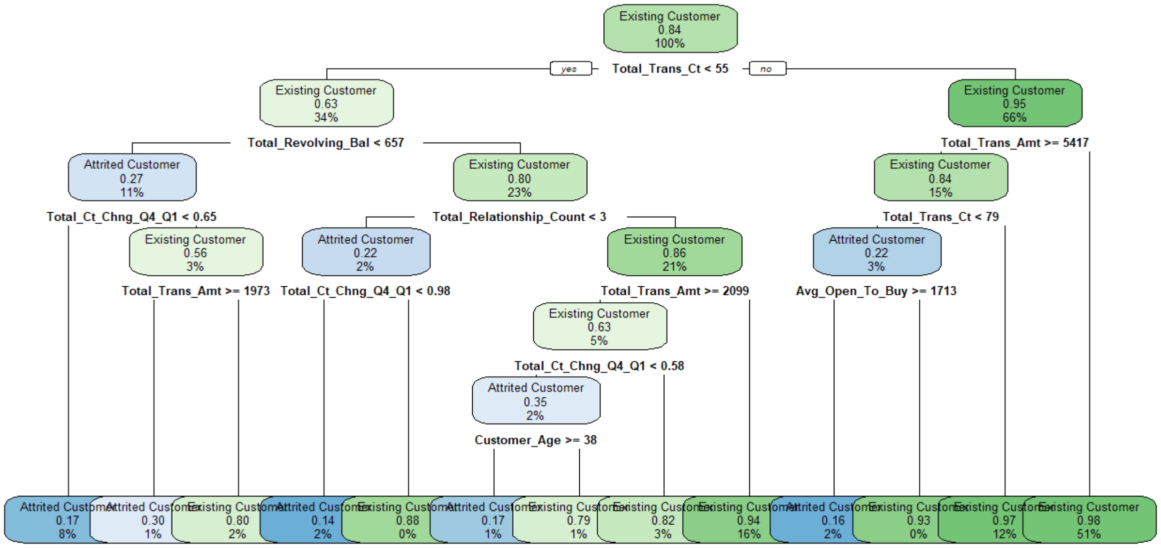

Existing Customer 0.84 100%
yes
Total_Trans_Ct < 55
no
Existing Customer 0.63 34%
Existing Customer 0.95 66%
Total_Revolving_Bal < 657
Total_Trans_Amt >= 5417
Attrited Customer 0.27 11%
Existing Customer 0.80 23%
Existing Customer 0.84 15%
Total_Ct_Chng_Q4_Q1 < 0.65
Total_Relationship_Count < 3
Total_Trans_Ct < 79
Existing Customer 0.56 3%
Attrited Customer 0.22 2%
Existing Customer 0.86 21%
Attrited Customer 0.22 3%
Total_Trans_Amt >= 1973
Total_Ct_Chng_Q4_Q1 < 0.98
Total_Trans_Amt >= 2099
Avg_Open_To_Buy >= 1713
Existing Customer 0.63 5%
Total_Ct_Chng_Q4_Q1 < 0.58
Attrited Customer 0.35 2%
Customer_Age >= 38
Attrited Customer 0.17 8%
Attrited Customer 0.30 1%
Existing Customer 0.80 2%
Attrited Customer 0.14 2%
Existing Customer 0.88 0%
Attrited Customer 0.17 1%
Existing Customer 0.79 1%
Existing Customer 0.82 3%
Existing Customer 0.94 16%
Attrited Customer 0.16 2%
Existing Customer 0.93 0%
Existing Customer 0.97 12%
Existing Customer 0.98 51%

Figure 6: Simple decision tree with pruning - pruning has no effect on the tree

3.4 Bagged Decision Tree

In the tree bagging process, randomly selected subsets of the complete data set are extracted and fit to a decision tree model. This process is repeated to generate a series of tree results, which are then averaged together in order to generate the final model. This model is generated through the treebag method in caret. There are no tuning parameters available for this model. Table 1 below shows the overall importance of each variable which was derived from the bagged decision tree. As shown, the most significant variable in determining customer churn is the total transaction amount from the customer, followed by the total revolving balance. Other important variables to consider are the transaction count change from quarter four to quarter one and the overall transaction count per customer. Additionally, the model shows that gender, marital status, number of dependents, and the card category have the least significant among the variables within the decision tree model. Table 1 shows the remaining 18 columns after removing the omitted columns. The variables are listed from most significant to least significant based on the simple decision tree in Table 1.

VARIABLE NAME	DATA TYPE	OVERALL IMPORTANCE
Total_Trans_Amt	Numerical	1039.32
Total_Revoling_Bal	Numerical	894.72
Total_Ct_Change_Q4_Q1	Numerical	839.67
Total_Trans_Ct	Numerical	711.71
Aver_Utilization_Ratio	Numerical	627.34
Total_Amt_Chng_Q4_Q1	Numerical	412.18
Customer_Age	Numerical	214.10
Avg_Open_To_Buy	Numerical	172.62
Credit_Limit	Numerical	170.83
Months_On_book	Numerical	166.15
Months_Inactive_12_mon	Numerical	146.16
Education_Level	String	140.20
Contacts_Count_12_mon	Numerical	128.89
Income_Category	String	123.27
Gender	String	62.87
Marital_Status	String	78.15
Dependent_count	Numerical	72.42
Card_Catergory	String	15.09

Table 1: Variable descriptions from the extracted data of the credit card customers data set.

4. Results and Performance

4.1 Comparison of Non-Optimized Models

This comparison of the four models is used using the outputs of the data set which were preprocessed from the parent data set. The bagged decision tree model outperformed all of the other models with an overall accuracy of 96%. The overall accuracy of the other models is considered poor. In comparison, the linear regression had an accuracy of 37%, the logistic regression obtained an accuracy of 42%, and the simple decision tree had a 93% accuracy. See Appendix 4 for R code. After completing the linear regression model, it was clear that there were no linear relationships between the variables in the data set. Therefore, the linear and logistic regressions are not helpful in the decision making process, and a decision tree was used to make further investigation. Building a decision proved to be much more effective as the attrition flag was clearly stated using the most significant factors.The classification decision tree utilized both numeric and categorical variables to deliver over 92% accuracy. Ultimately, the bagged decision tree model was used to provide the overall level of significance of the independent variables as referenced in Table 1. Furthermore, the p-value was less than 2.2e^{-16}, which is statistically significant. Therefore, the data supports the hypothesis that the total transaction amount and the total number of transactions have an impact on customer attrition.

4.2 Optimized Decision Tree

Simple decision tree. The results of the simple decision tree model, described in Section 3.3, are shown in Figure 5 and Figure 6. Figure 6 demonstrates that pruning does not affect the function tree complexity. Furthermore, Figure 7 shows that the cp value and optimal cp value intersects only tree sizes 8,9 and 10. This indicates that pruning these branches won’t have much of an impact on the decision tree. Thus, following this inference from the cp value, we choose to not prune the decision tree as it would not improve the tree.

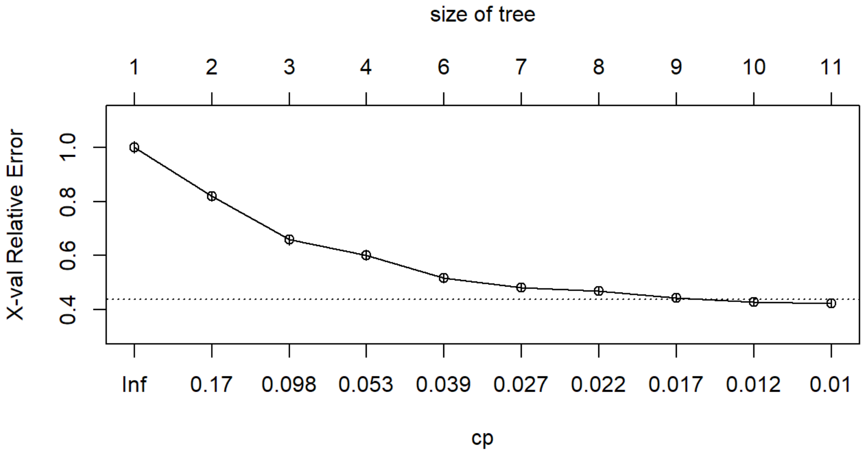


Figure 7: Complexity parameters vs relative error of decision tree

Bagged decision tree. The best performing model is the bagged decision tree. The accuracy was 96%. See appendix 4 for R code. ADD chart of RMSE performance of the bagged decision tree model

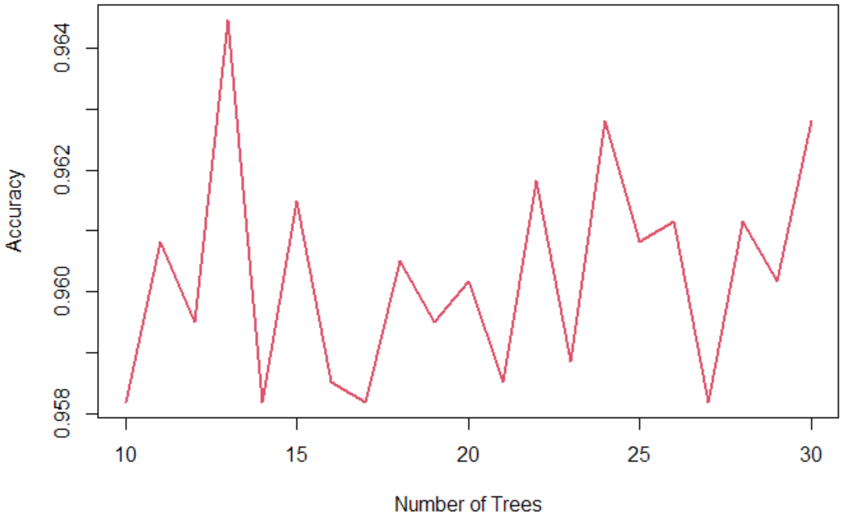
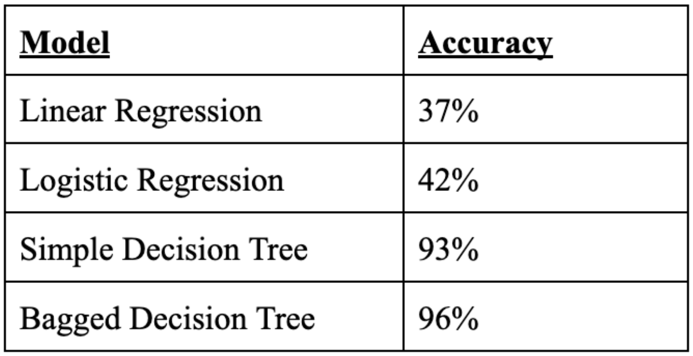


Figure 8: By fitting the bagging multiple times and predicting the testing sample, we can draw the following boxplot to show the variance of the prediction error at different number of trees

4.3 Results and Interpretation Related to Hypothesis

The initial hypothesis for this investigation is that total transaction amount and the total number of transactions will have a large impact on customer attrition. Customer attrition is measured by the number of customers that are churned at the end of the 12-month period. Shown in Figure 1, the bagged decision tree model confirmed that the most significant variable in determining attrited customers is total transaction amount. Based on this analysis, the initial hypothesis is supported. However, the bagged decision tree model identified the second most significant variable to be the customer's total revolving balance, the third most significant to be the customer's total transaction count change from quarter four to quarter one, and the fourth most significant to be total transaction count. Therefore, the initial hypothesis is partially supported and should be adjusted to reflect the other significantly important variables.



Model	**Accuracy**
Linear Regression	37%
Logistic Regression	42%
Simple Decision Tree	93%
Bagged Decision Tree	96%

Table 2: Performance accuracy of optimized models

4.4 Evaluation against Success Criteria

To summarize, four different models have been tested to better understand what is causing customer churn. Based on our analysis, the success criteria was defined by applying the bias-variance trade off and testing different models to decrease error. As such, both of these objectives have been met. The bias-variance tradeoff was achieved through the bagged decision tree. Bagging has been proven to achieve a higher accuracy since it uses independent variables from a random subset. The sample is then replaced to get an ensemble of different models. As the model repeats the sampling process, it simplifies the model and enhances total model accuracy. As described in Section 3.4, the model achieved accuracy of 96%.

To achieve the second criterion, different models were analyzed and tested. In accordance with Table 2, you can see the accuracy rates from each model. As we tested each model, we learned that logistic regression was the worst performing model. Further, a model that was attempted but was not used, was random forest. Though both of the models are great tools to fit and test, it was proven to be not effective for this effort. Aside from having limitations in our modeling methodology, the bagging decision tree supported our hypothesis.

Customer	Attrited	Existing
Attrited	410	46
Existing	78	2504

Accuracy: 0.9592

Table 3: Confusion matrix of bagged decision tree

5. Discussion and Recommendations

As shown previously, the variable which had the greatest impact when determining if a customer will churn or not was the total transaction amount. Another variable to consider when looking at a customer who may churn, would be the total transactions count. Customers who had a high number of total transactions were less likely to leave. This may be due to the high quality of customer service being offered to these customers when it comes to customers having questions or issues with disputing transactions. However, the overall ease and quality of service provided by the banks can increase the customers likelihood of staying with the credit card company.

Recommendations. Based on the bagged decision tree model, total number of transactions and the transaction amount are considered important. That being said, banks should consider ways to increase the number of transactions customers make as well as ways to increase the total transaction amount. Some suggestions may include offering customers additional rewards/points based on the customers spending history. This can encourage customers to use their credit card to make purchases they were not previously putting on their card in order to increase points to maximize potential benefits. As this is a quality service that customers may value and lead to an increase in retaining customers for a longer period of time.

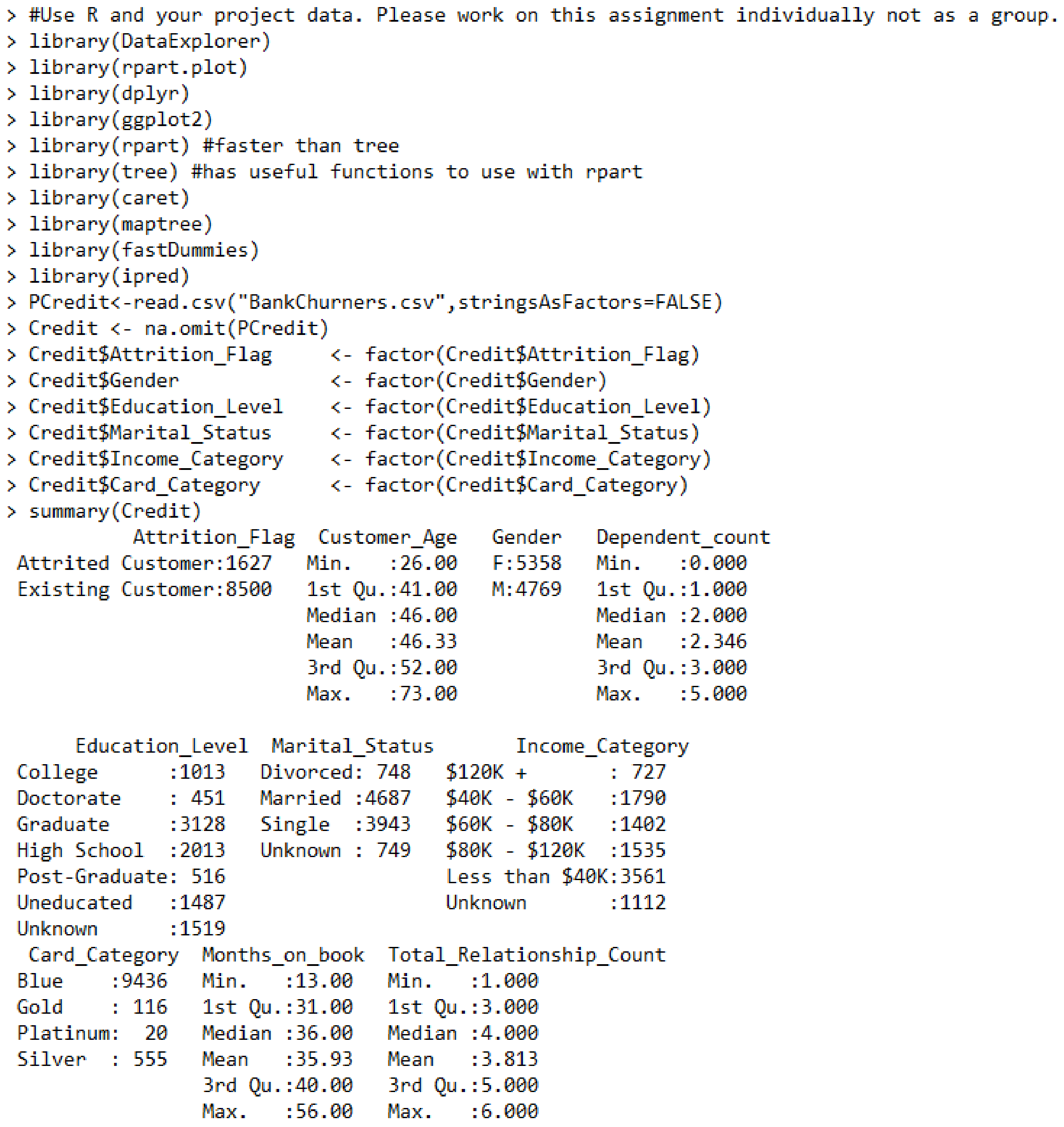
Another recommendation based on the bagged decision tree model for banks to consider is to regularly look at which customers have a low total transaction amount and proactively offer higher touch customer service in hopes of keeping customers likely to churn. Additionally, based on the simple decision tree, customers with a total transaction count of less than 55 are the most likely to churn, and the bank should pay special attention to these customers.

References

Kaggle, https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Appendices

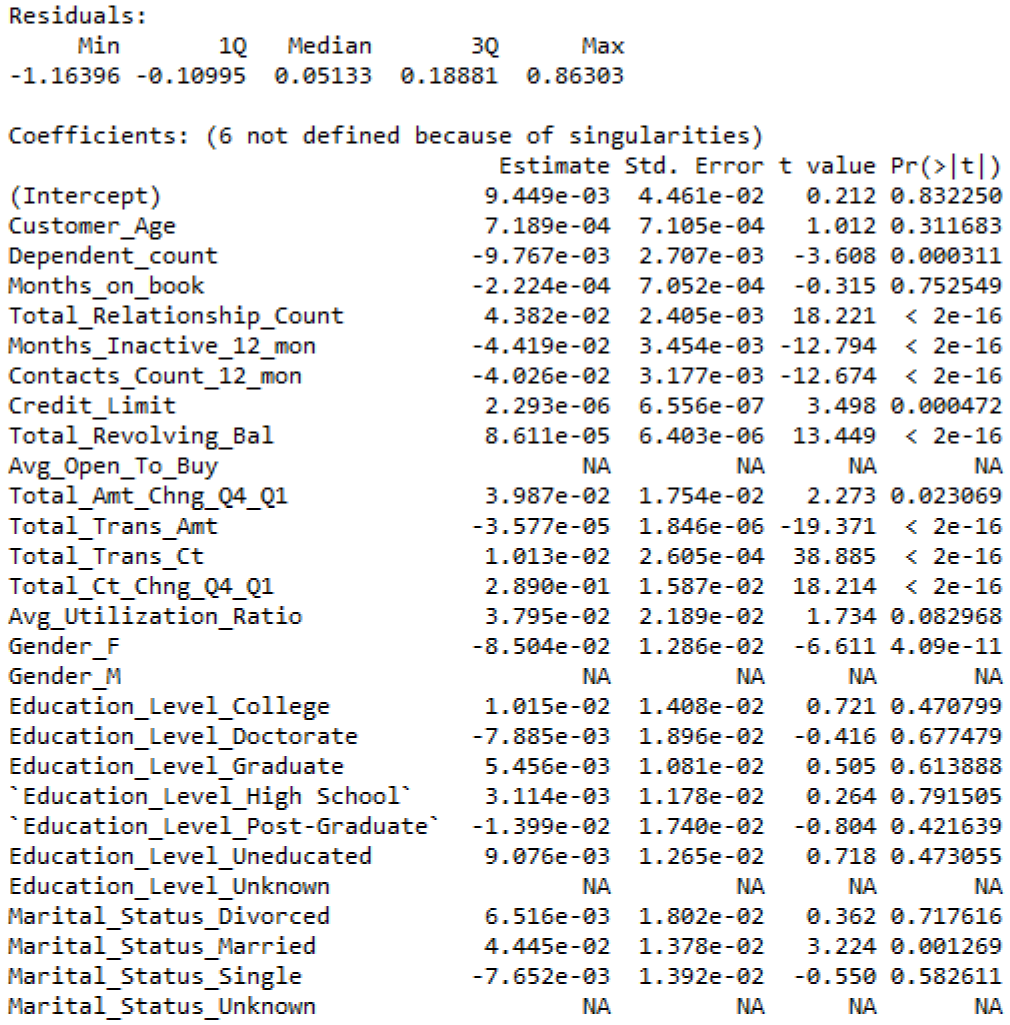
1. Preprocessing and Data Selection Code



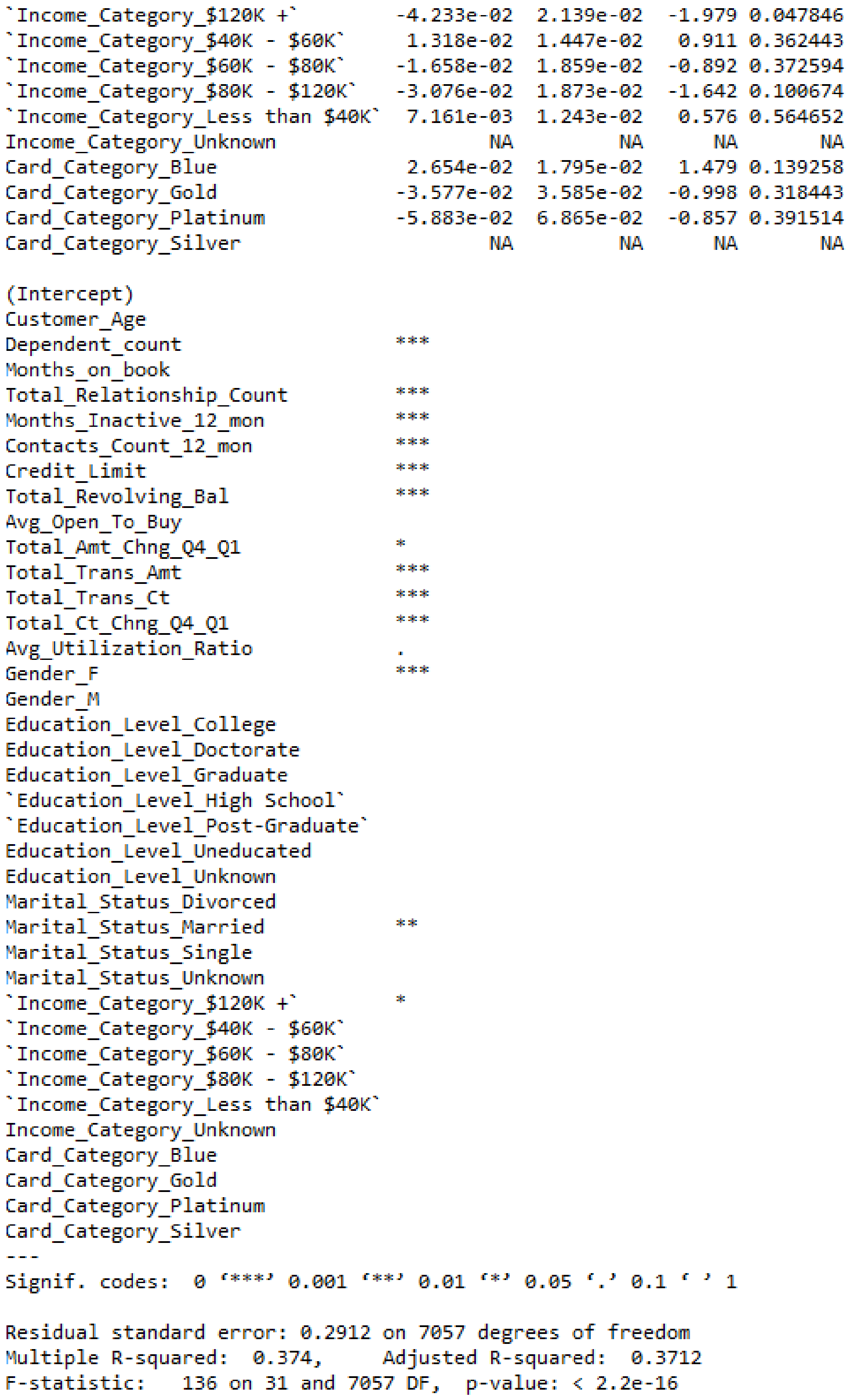
```
> #Use R and your project data. Please work on this assignment individually not as a group.
> library(DataExplorer)
> library(rpart.plot)
> library(dplyr)
> library(ggplot2)
> library(rpart) #faster than tree
> library(tree) #has useful functions to use with rpart
> library(caret)
> library(maptree)
> library(fastDummies)
> library(ipred)
> PCredit<-read.csv("BankChurners.csv",stringsAsFactors=FALSE)
> Credit <- na.omit(PCredit)
> Credit$Attrition_Flag     <- factor(Credit$Attrition_Flag)
> Credit$Gender             <- factor(Credit$Gender)
> Credit$Education_Level    <- factor(Credit$Education_Level)
> Credit$Marital_Status     <- factor(Credit$Marital_Status)
> Credit$Income_Category    <- factor(Credit$Income_Category)
> Credit$Card_Category      <- factor(Credit$Card_Category)
> summary(Credit)
           Attrition_Flag  Customer_Age   Gender    Dependent_count
 Attrited Customer:1627   Min.   :26.00   F:5358   Min.   :0.000  
 Existing Customer:8500   1st Qu.:41.00   M:4769   1st Qu.:1.000  
                          Median :46.00            Median :2.000  
                          Mean   :46.33            Mean   :2.346  
                          3rd Qu.:52.00            3rd Qu.:3.000  
                          Max.   :73.00            Max.   :5.000  

      Education_Level  Marital_Status       Income_Category
 College      :1013   Divorced: 748   $120K +        : 727  
 Doctorate    : 451   Married :4687   $40K - $60K    :1790  
 Graduate     :3128   Single  :3943   $60K - $80K    :1402  
 High School  :2013   Unknown : 749   $80K - $120K   :1535  
 Post-Graduate: 516                   Less than $40K:3561  
 Uneducated   :1487                   Unknown        :1112  
 Unknown      :1519                                         
  Card_Category  Months_on_book  Total_Relationship_Count
 Blue    :9436   Min.   :13.00   Min.   :1.000           
 Gold    : 116   1st Qu.:31.00   1st Qu.:3.000           
 Platinum:  20   Median :36.00   Median :4.000           
 Silver  : 555   Mean   :35.93   Mean   :3.813           
                 3rd Qu.:40.00   3rd Qu.:5.000           
                 Max.   :56.00   Max.   :6.000           
 Months_Inactive_12_mon Contacts_Count_12_mon  Credit_Limit
 Min.   :0.000          Min.   :0.000         Min.   : 1438
 1st Qu.:2.000          1st Qu.:2.000         1st Qu.: 2555
 Median :2.000          Median :2.000         Median : 4549
 Mean   :2.341          Mean   :2.455         Mean   : 8632
 3rd Qu.:3.000          3rd Qu.:3.000         3rd Qu.:11068
 Max.   :6.000          Max.   :6.000         Max.   :34516

 Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1
 Min.   :   0        Min.   :    3   Min.   :0.0000
 1st Qu.: 359        1st Qu.: 1324   1st Qu.:0.6310
 Median :1276        Median : 3474   Median :0.7360
 Mean   :1163        Mean   : 7469   Mean   :0.7599
 3rd Qu.:1784        3rd Qu.: 9859   3rd Qu.:0.8590
 Max.   :2517        Max.   :34516   Max.   :3.3970

 Total_Trans_Amt Total_Trans_Ct   Total_Ct_Chng_Q4_Q1
 Min.   :  510   Min.   : 10.00   Min.   :0.0000
 1st Qu.: 2156   1st Qu.: 45.00   1st Qu.:0.5820
 Median : 3899   Median : 67.00   Median :0.7020
 Mean   : 4404   Mean   : 64.86   Mean   :0.7122
 3rd Qu.: 4741   3rd Qu.: 81.00   3rd Qu.:0.8180
 Max.   :18484   Max.   :139.00   Max.   :3.7140

 Avg_Utilization_Ratio
 Min.   :0.0000
 1st Qu.:0.0230
 Median :0.1760
 Mean   :0.2749
 3rd Qu.:0.5030
 Max.   :0.9990
```

2. *Treating Data and Visual Overview*

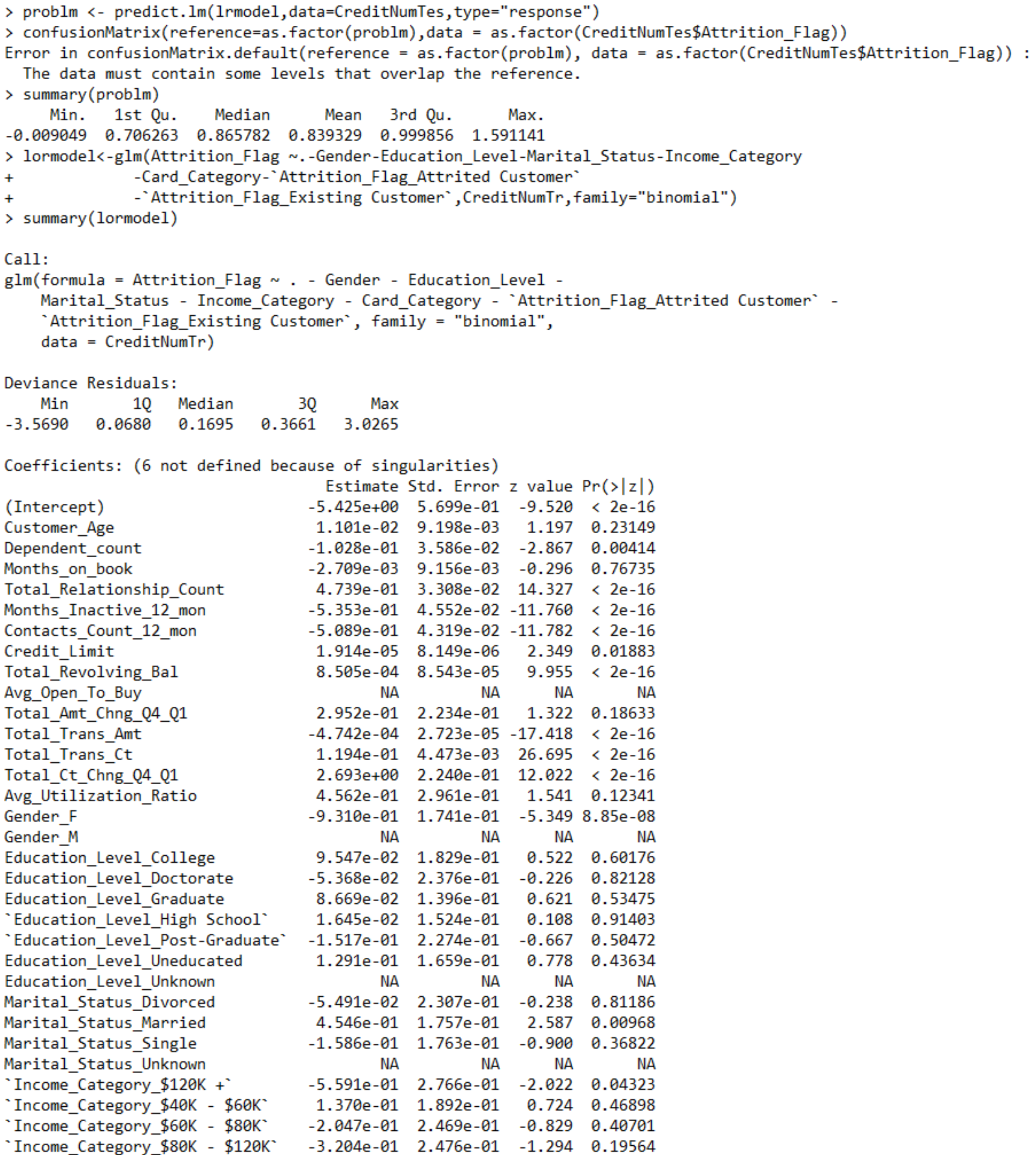
```
> #separating numeric data into another variable
> numericData<-(Credit ~-Credit$Attrition_Flag-Credit$Gender-Credit$Education_Level
+             -Credit$Marital_Status-Credit$Income_Category-Credit$Card_Category)
> #eda charts
> hist(Credit$Customer_Age)
> hist(Credit$Dependent_count)
> ggplot(Credit, aes(Attrition_Flag, fill = Attrition_Flag)) +geom_bar()
+theme(legend.position = 'none')
> ggplot(Credit, aes(y=Education_Level))+geom_bar(aes(fill =Attrition_Flag),position =
"dodge")+xlab("Count") + ylab("Education Level") + ggtitle("Customer Status by Education
Level" )+  labs(fill = "Customer Status") + theme_minimal()
> ggplot(Credit, aes(y=Marital_Status))+geom_bar(aes(fill =Attrition_Flag),position =
"dodge")+xlab("Count") + ylab("Marital Status") + ggtitle("Customer Status by Marital Status"
)+  labs(fill = "Customer Status") + theme_minimal()
> ggplot(Credit, aes(x=Credit_Limit, fill = Attrition_Flag)) + geom_histogram() +
theme_minimal() + scale_x_continuous(breaks = seq(0,50000, by=10000))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> plot_missing(Credit)
> #using caret to spilt data into testing and training
> set.seed(1234)
> trainIndex <- createDataPartition(Credit$Attrition_Flag, p = .7,list=FALSE)
> training <- Credit[trainIndex,]
> testing <- Credit[-trainIndex,]
> #traning data
> CreditNumTr <- dummy_cols(training)
> CreditNumTr$Attrition_Flag<-ifelse(CreditNumTr$Attrition_Flag == "Existing Customer",1,0)
> #testdata
> CreditNumTe <- dummy_cols(testing)
> CreditNumTe$Attrition_Flag<-ifelse(CreditNumTe$Attrition_Flag == "Existing Customer",1,0)
> CreditNumTes<-subset(CreditNumTe,select=-c(CreditNumTe$Gender,CreditNumTe$Education_Level,
+
CreditNumTe$Marital_Status,CreditNumTe$Income_Category,CreditNumTe$Card_Category
+                      ,CreditNumTe$`Attrition_Flag_Attrited Customer`
+                      ,CreditNumTe$`Attrition_Flag_Existing Customer`))
> #creating model removing categorical columns
> lrmodel<-lm(Attrition_Flag ~.-Gender-Education_Level-Marital_Status-Income_Category
+              -Card_Category-`Attrition_Flag_Attrited Customer`
+              -`Attrition_Flag_Existing Customer`,CreditNumTr)
> summary(lrmodel)

Call:
lm(formula = Attrition_Flag ~ . - Gender - Education_Level -
    Marital_Status - Income_Category - Card_Category - `Attrition_Flag_Attrited Customer` -
    `Attrition_Flag_Existing Customer`, data = CreditNumTr)
```



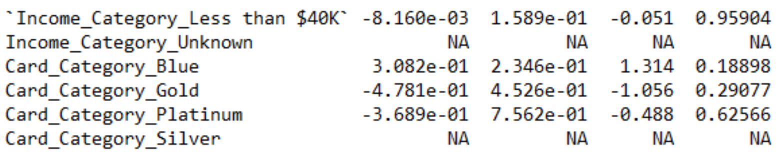
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.16396 -0.10995  0.05133  0.18881  0.86303 

Coefficients: (6 not defined because of singularities)
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       9.449e-03  4.461e-02   0.212 0.832250
Customer_Age                      7.189e-04  7.105e-04   1.012 0.311683
Dependent_count                  -9.767e-03  2.707e-03  -3.608 0.000311
Months_on_book                   -2.224e-04  7.052e-04  -0.315 0.752549
Total_Relationship_Count          4.382e-02  2.405e-03  18.221  < 2e-16
Months_Inactive_12_mon           -4.419e-02  3.454e-03 -12.794  < 2e-16
Contacts_Count_12_mon            -4.026e-02  3.177e-03 -12.674  < 2e-16
Credit_Limit                      2.293e-06  6.556e-07   3.498 0.000472
Total_Revolving_Bal               8.611e-05  6.403e-06  13.449  < 2e-16
Avg_Open_To_Buy                          NA         NA      NA       NA
Total_Amt_Chng_Q4_Q1              3.987e-02  1.754e-02   2.273 0.023069
Total_Trans_Amt                  -3.577e-05  1.846e-06 -19.371  < 2e-16
Total_Trans_Ct                    1.013e-02  2.605e-04  38.885  < 2e-16
Total_Ct_Chng_Q4_Q1               2.890e-01  1.587e-02  18.214  < 2e-16
Avg_Utilization_Ratio             3.795e-02  2.189e-02   1.734 0.082968
Gender_F                         -8.504e-02  1.286e-02  -6.611 4.09e-11
Gender_M                                 NA         NA      NA       NA
Education_Level_College           1.015e-02  1.408e-02   0.721 0.470799
Education_Level_Doctorate        -7.885e-03  1.896e-02  -0.416 0.677479
Education_Level_Graduate          5.456e-03  1.081e-02   0.505 0.613888
`Education_Level_High School`     3.114e-03  1.178e-02   0.264 0.791505
`Education_Level_Post-Graduate`  -1.399e-02  1.740e-02  -0.804 0.421639
Education_Level_Uneducated        9.076e-03  1.265e-02   0.718 0.473055
Education_Level_Unknown                  NA         NA      NA       NA
Marital_Status_Divorced           6.516e-03  1.802e-02   0.362 0.717616
Marital_Status_Married            4.445e-02  1.378e-02   3.224 0.001269
Marital_Status_Single            -7.652e-03  1.392e-02  -0.550 0.582611
Marital_Status_Unknown                   NA         NA      NA       NA
```



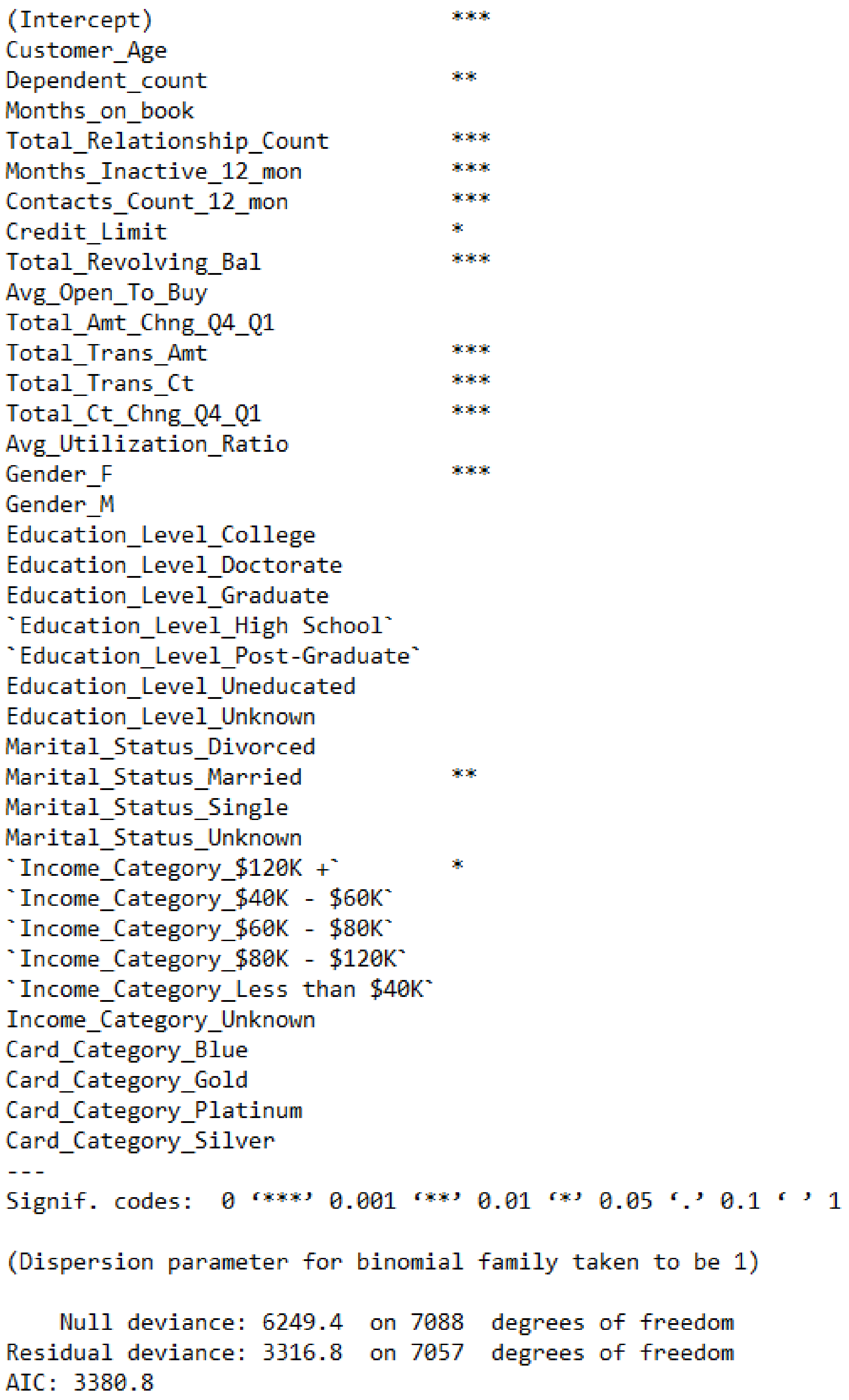
```
`Income_Category_$120K +`        -4.233e-02  2.139e-02  -1.979 0.047846
`Income_Category_$40K - $60K`     1.318e-02  1.447e-02   0.911 0.362443
`Income_Category_$60K - $80K`    -1.658e-02  1.859e-02  -0.892 0.372594
`Income_Category_$80K - $120K`   -3.076e-02  1.873e-02  -1.642 0.100674
`Income_Category_Less than $40K`  7.161e-03  1.243e-02   0.576 0.564652
Income_Category_Unknown                  NA         NA      NA       NA
Card_Category_Blue                2.654e-02  1.795e-02   1.479 0.139258
Card_Category_Gold               -3.577e-02  3.585e-02  -0.998 0.318443
Card_Category_Platinum           -5.883e-02  6.865e-02  -0.857 0.391514
Card_Category_Silver                     NA         NA      NA       NA

(Intercept)
Customer_Age
Dependent_count                  ***
Months_on_book
Total_Relationship_Count         ***
Months_Inactive_12_mon           ***
Contacts_Count_12_mon            ***
Credit_Limit                     ***
Total_Revolving_Bal              ***
Avg_Open_To_Buy
Total_Amt_Chng_Q4_Q1             *
Total_Trans_Amt                  ***
Total_Trans_Ct                   ***
Total_Ct_Chng_Q4_Q1              ***
Avg_Utilization_Ratio            .
Gender_F                         ***
Gender_M
Education_Level_College
Education_Level_Doctorate
Education_Level_Graduate
`Education_Level_High School`
`Education_Level_Post-Graduate`
Education_Level_Uneducated
Education_Level_Unknown
Marital_Status_Divorced
Marital_Status_Married           **
Marital_Status_Single
Marital_Status_Unknown
`Income_Category_$120K +`        *
`Income_Category_$40K - $60K`
`Income_Category_$60K - $80K`
`Income_Category_$80K - $120K`
`Income_Category_Less than $40K`
Income_Category_Unknown
Card_Category_Blue
Card_Category_Gold
Card_Category_Platinum
Card_Category_Silver
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2912 on 7057 degrees of freedom
Multiple R-squared:  0.374,     Adjusted R-squared:  0.3712
F-statistic:   136 on 31 and 7057 DF,  p-value: < 2.2e-16
```



```
> problm <- predict.lm(lrmodel,data=CreditNumTes,type="response")
> confusionMatrix(reference=as.factor(problm),data = as.factor(CreditNumTes$Attrition_Flag))
Error in confusionMatrix.default(reference = as.factor(problm), data = as.factor(CreditNumTes$Attrition_Flag)) :
  The data must contain some levels that overlap the reference.
> summary(problm)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.009049  0.706263  0.865782  0.839329  0.999856  1.591141
> lormodel<-glm(Attrition_Flag ~.-Gender-Education_Level-Marital_Status-Income_Category
+             -Card_Category-`Attrition_Flag_Attrited Customer`
+             -`Attrition_Flag_Existing Customer`,CreditNumTr,family="binomial")
> summary(lormodel)

Call:
glm(formula = Attrition_Flag ~ . - Gender - Education_Level -
    Marital_Status - Income_Category - Card_Category - `Attrition_Flag_Attrited Customer` -
    `Attrition_Flag_Existing Customer`, family = "binomial",
    data = CreditNumTr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5690   0.0680   0.1695   0.3661   3.0265

Coefficients: (6 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.425e+00  5.699e-01  -9.520  < 2e-16
Customer_Age                     1.101e-02  9.198e-03   1.197  0.23149
Dependent_count                 -1.028e-01  3.586e-02  -2.867  0.00414
Months_on_book                  -2.709e-03  9.156e-03  -0.296  0.76735
Total_Relationship_Count         4.739e-01  3.308e-02  14.327  < 2e-16
Months_Inactive_12_mon          -5.353e-01  4.552e-02 -11.760  < 2e-16
Contacts_Count_12_mon           -5.089e-01  4.319e-02 -11.782  < 2e-16
Credit_Limit                     1.914e-05  8.149e-06   2.349  0.01883
Total_Revolving_Bal              8.505e-04  8.543e-05   9.955  < 2e-16
Avg_Open_To_Buy                         NA         NA      NA       NA
Total_Amt_Chng_Q4_Q1             2.952e-01  2.234e-01   1.322  0.18633
Total_Trans_Amt                 -4.742e-04  2.723e-05 -17.418  < 2e-16
Total_Trans_Ct                   1.194e-01  4.473e-03  26.695  < 2e-16
Total_Ct_Chng_Q4_Q1              2.693e+00  2.240e-01  12.022  < 2e-16
Avg_Utilization_Ratio            4.562e-01  2.961e-01   1.541  0.12341
Gender_F                        -9.310e-01  1.741e-01  -5.349 8.85e-08
Gender_M                                NA         NA      NA       NA
Education_Level_College          9.547e-02  1.829e-01   0.522  0.60176
Education_Level_Doctorate       -5.368e-02  2.376e-01  -0.226  0.82128
Education_Level_Graduate         8.669e-02  1.396e-01   0.621  0.53475
`Education_Level_High School`    1.645e-02  1.524e-01   0.108  0.91403
`Education_Level_Post-Graduate` -1.517e-01  2.274e-01  -0.667  0.50472
Education_Level_Uneducated       1.291e-01  1.659e-01   0.778  0.43634
Education_Level_Unknown                 NA         NA      NA       NA
Marital_Status_Divorced         -5.491e-02  2.307e-01  -0.238  0.81186
Marital_Status_Married           4.546e-01  1.757e-01   2.587  0.00968
Marital_Status_Single           -1.586e-01  1.763e-01  -0.900  0.36822
Marital_Status_Unknown                  NA         NA      NA       NA
`Income_Category_$120K +`       -5.591e-01  2.766e-01  -2.022  0.04323
`Income_Category_$40K - $60K`    1.370e-01  1.892e-01   0.724  0.46898
`Income_Category_$60K - $80K`   -2.047e-01  2.469e-01  -0.829  0.40701
`Income_Category_$80K - $120K`  -3.204e-01  2.476e-01  -1.294  0.19564
```



```
`Income_Category_Less than $40K` -8.160e-03  1.589e-01  -0.051  0.95904
Income_Category_Unknown                  NA         NA      NA       NA
Card_Category_Blue                3.082e-01  2.346e-01   1.314  0.18898
Card_Category_Gold               -4.781e-01  4.526e-01  -1.056  0.29077
Card_Category_Platinum           -3.689e-01  7.562e-01  -0.488  0.62566
Card_Category_Silver                     NA         NA      NA       NA
```



```
(Intercept)                       ***
Customer_Age
Dependent_count                   **
Months_on_book
Total_Relationship_Count          ***
Months_Inactive_12_mon            ***
Contacts_Count_12_mon             ***
Credit_Limit                      *
Total_Revolving_Bal               ***
Avg_Open_To_Buy
Total_Amt_Chng_Q4_Q1
Total_Trans_Amt                   ***
Total_Trans_Ct                    ***
Total_Ct_Chng_Q4_Q1               ***
Avg_Utilization_Ratio
Gender_F                          ***
Gender_M
Education_Level_College
Education_Level_Doctorate
Education_Level_Graduate
`Education_Level_High School`
`Education_Level_Post-Graduate`
Education_Level_Uneducated
Education_Level_Unknown
Marital_Status_Divorced
Marital_Status_Married            **
Marital_Status_Single
Marital_Status_Unknown
`Income_Category_$120K +`         *
`Income_Category_$40K - $60K`
`Income_Category_$60K - $80K`
`Income_Category_$80K - $120K`
`Income_Category_Less than $40K`
Income_Category_Unknown
Card_Category_Blue
Card_Category_Gold
Card_Category_Platinum
Card_Category_Silver
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6249.4  on 7088  degrees of freedom
Residual deviance: 3316.8  on 7057  degrees of freedom
AIC: 3380.8
```

```
Number of Fisher Scoring iterations: 6
```

3. General Visualization Code

```
> problom <- predict.glm(lormodel,data=CreditNumTes,type="response")
> confusionMatrix(problom,as.factor(CreditNumTes$Attrition_Flag))
Error: `data` and `reference` should be factors with the same levels.
> summary(problom)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.002234 0.824025 0.963634 0.839329 0.991722 0.999960 
> #create tree
> hit.rtree<-rpart(Attrition_Flag ~., data=training, method="class")
> #summarize full tree (no pruning)
> summary(hit.rtree)
Call:
rpart(formula = Attrition_Flag ~ ., data = training, method = "class")
  n= 7089 

           CP nsplit rel error    xerror       xstd
1  0.16681299      0 1.0000000 1.0000000 0.02714589
2  0.07199298      2 0.6663740 0.6742757 0.02297509
3  0.03994732      3 0.5943810 0.6022827 0.02185434
4  0.02897278      5 0.5144864 0.5267779 0.02057540
5  0.01843723      6 0.4855136 0.5039508 0.02016494
6  0.01492537      9 0.4214223 0.4407375 0.01896180
7  0.01141352     10 0.4064969 0.4258121 0.01866200
8  0.01097454     11 0.3950834 0.4266901 0.01867982
9  0.01053556     13 0.3731343 0.4266901 0.01867982
10 0.01000000     15 0.3520632 0.4214223 0.01857258

Variable importance
          Total_Trans_Ct          Total_Trans_Amt 
                      25                       19 
     Total_Revolving_Bal    Avg_Utilization_Ratio 
                      14                       13 
     Total_Ct_Chng_Q4_Q1 Total_Relationship_Count 
                       9                        8 
            Credit_Limit     Total_Amt_Chng_Q4_Q1 
                       4                        3 
         Avg_Open_To_Buy             Customer_Age 
                       2                        1 
          Months_on_book 
                       1 

Node number 1: 7089 observations,    complexity param=0.166813
  predicted class=Existing Customer  expected loss=0.1606715  P(node) =1
    class counts:  1139  5950
   probabilities: 0.161 0.839 
  left son=2 (2379 obs) right son=3 (4710 obs)
  Primary splits:
      Total_Trans_Ct        < 54.5    to the left,  improve=339.2040, (0 missing)
      Total_Ct_Chng_Q4_Q1   < 0.504   to the left,  improve=274.4460, (0 missing)
      Total_Trans_Amt       < 2936.5  to the left,  improve=259.3219, (0 missing)
      Total_Revolving_Bal   < 581.5   to the left,  improve=258.9396, (0 missing)
      Avg_Utilization_Ratio < 0.0205  to the left,  improve=194.1999, (0 missing)
  Surrogate splits:
      Total_Trans_Amt       < 2871    to the left,  agree=0.927, adj=0.781, (0 split)
      Total_Ct_Chng_Q4_Q1   < 0.504   to the left,  agree=0.759, adj=0.281, (0 split)
      Total_Amt_Chng_Q4_Q1  < 0.5075  to the left,  agree=0.695, adj=0.091, (0 split)
      Customer_Age          < 26.5    to the left,  agree=0.667, adj=0.008, (0 split)
      Contacts_Count_12_mon < 5.5     to the right, agree=0.666, adj=0.006, (0 split)

Node number 2: 2379 observations,    complexity param=0.166813
```

```
  predicted class=Existing Customer  expected loss=0.3783102  P(node) =0.3355904
    class counts:   900  1479
   probabilities: 0.378 0.622 
  left son=4 (778 obs) right son=5 (1601 obs)
  Primary splits:
      Total_Revolving_Bal      < 606.5   to the left,  improve=309.5645, (0 missing)
      Avg_Utilization_Ratio    < 0.0265  to the left,  improve=238.5497, (0 missing)
      Total_Relationship_Count < 2.5     to the left,  improve=204.2594, (0 missing)
      Total_Trans_Amt          < 2009    to the right, improve=197.5695, (0 missing)
      Total_Ct_Chng_Q4_Q1      < 0.631   to the left,  improve=144.1076, (0 missing)
  Surrogate splits:
      Avg_Utilization_Ratio    < 0.022   to the left,  agree=0.964, adj=0.889, (0 split)
      Credit_Limit             < 1836.5  to the left,  agree=0.731, adj=0.176, (0 split)
      Avg_Open_To_Buy          < 33763   to the right, agree=0.687, adj=0.044, (0 split)
      Total_Relationship_Count < 2.5     to the left,  agree=0.687, adj=0.042, (0 split)
      Total_Trans_Amt          < 922     to the left,  agree=0.686, adj=0.041, (0 split)

Node number 3: 4710 observations,    complexity param=0.03994732
  predicted class=Existing Customer  expected loss=0.0507431  P(node) =0.6644096
    class counts:   239  4471
   probabilities: 0.051 0.949 
  left son=6 (1074 obs) right son=7 (3636 obs)
  Primary splits:
      Total_Trans_Amt       < 5417    to the right, improve=29.99075, (0 missing)
      Contacts_Count_12_mon < 5.5     to the right, improve=23.49315, (0 missing)
      Total_Revolving_Bal   < 581.5   to the left,  improve=22.44493, (0 missing)
      Avg_Utilization_Ratio < 0.0205  to the left,  improve=18.35761, (0 missing)
      Total_Trans_Ct        < 64.5    to the left,  improve=13.71180, (0 missing)
  Surrogate splits:
      Total_Trans_Ct           < 92.5    to the right, agree=0.896, adj=0.546, (0 split)
      Total_Relationship_Count < 2.5     to the left,  agree=0.836, adj=0.282, (0 split)
      Card_Category            splits as  RLLL,        agree=0.788, adj=0.070, (0 split)
      Credit_Limit             < 33543   to the right, agree=0.782, adj=0.043, (0 split)
      Avg_Open_To_Buy          < 31674.5 to the right, agree=0.782, adj=0.043, (0 split)

Node number 4: 778 observations,    complexity param=0.02897278
  predicted class=Attrited Customer  expected loss=0.2557841  P(node) =0.1097475
    class counts:   579   199
   probabilities: 0.744 0.256 
  left son=8 (663 obs) right son=9 (115 obs)
  Primary splits:
      Total_Ct_Chng_Q4_Q1      < 0.7805  to the left,  improve=40.56696, (0 missing)
      Total_Trans_Amt          < 1946.5  to the right, improve=40.26052, (0 missing)
      Total_Relationship_Count < 2.5     to the left,  improve=28.69208, (0 missing)
      Total_Amt_Chng_Q4_Q1     < 1.06    to the left,  improve=24.73586, (0 missing)
      Months_Inactive_12_mon   < 1.5     to the right, improve=20.29139, (0 missing)
  Surrogate splits:
      Total_Amt_Chng_Q4_Q1  < 1.0635  to the left,  agree=0.868, adj=0.104, (0 split)
      Contacts_Count_12_mon < 0.5     to the right, agree=0.855, adj=0.017, (0 split)
      Avg_Utilization_Ratio < 0.402   to the left,  agree=0.853, adj=0.009, (0 split)

Node number 5: 1601 observations,    complexity param=0.07199298
  predicted class=Existing Customer  expected loss=0.2004997  P(node) =0.2258429
    class counts:   321  1280
   probabilities: 0.200 0.800 
  left son=10 (150 obs) right son=11 (1451 obs)
  Primary splits:
      Total_Relationship_Count < 2.5     to the left,  improve=108.61810, (0 missing)
      Total_Trans_Amt          < 2102    to the right, improve= 74.26230, (0 missing)
      Total_Revolving_Bal      < 2381.5  to the right, improve= 43.11666, (0 missing)
      Total_Ct_Chng_Q4_Q1      < 0.5845  to the left,  improve= 40.46930, (0 missing)
      Marital_Status           splits as  LRLL,        improve= 20.73835, (0 missing)
  Surrogate splits:
      Total_Trans_Amt      < 756.5   to the left,  agree=0.909, adj=0.033, (0 split)
      Total_Amt_Chng_Q4_Q1 < 0.1505  to the left,  agree=0.908, adj=0.013, (0 split)
      Total_Trans_Ct       < 15.5    to the left,  agree=0.908, adj=0.013, (0 split)

Node number 6: 1074 observations,    complexity param=0.03994732
  predicted class=Existing Customer  expected loss=0.1545624  P(node) =0.1515023
    class counts:   166   908
   probabilities: 0.155 0.845 
  left son=12 (187 obs) right son=13 (887 obs)
  Primary splits:
      Total_Trans_Ct        < 79.5    to the left,  improve=156.97070, (0 missing)
      Total_Revolving_Bal   < 579     to the left,  improve= 60.88729, (0 missing)
      Total_Trans_Amt       < 11035.5 to the left,  improve= 51.12395, (0 missing)
      Avg_Utilization_Ratio < 0.0165  to the left,  improve= 47.99255, (0 missing)
      Total_Amt_Chng_Q4_Q1  < 0.8885  to the right, improve= 41.95393, (0 missing)
  Surrogate splits:
      Total_Amt_Chng_Q4_Q1  < 0.535   to the left,  agree=0.837, adj=0.064, (0 split)
      Total_Trans_Amt       < 6738    to the left,  agree=0.834, adj=0.048, (0 split)
      Total_Ct_Chng_Q4_Q1   < 1.024   to the right, agree=0.831, adj=0.027, (0 split)
      Contacts_Count_12_mon < 3.5     to the right, agree=0.828, adj=0.011, (0 split)

Node number 7: 3636 observations
  predicted class=Existing Customer  expected loss=0.02007701  P(node) =0.5129073
    class counts:    73  3563
   probabilities: 0.020 0.980 

Node number 8: 663 observations,    complexity param=0.01097454
  predicted class=Attrited Customer  expected loss=0.188537  P(node) =0.09352518
    class counts:   538   125
   probabilities: 0.811 0.189 
  left son=16 (400 obs) right son=17 (263 obs)
  Primary splits:
      Total_Trans_Amt          < 1946.5  to the right, improve=29.544980, (0 missing)
      Months_Inactive_12_mon   < 1.5     to the right, improve=18.591090, (0 missing)
      Total_Relationship_Count < 2.5     to the left,  improve=14.505210, (0 missing)
      Customer_Age             < 31.5    to the right, improve= 9.487436, (0 missing)
      Total_Ct_Chng_Q4_Q1      < 0.5735  to the left,  improve= 9.252941, (0 missing)
  Surrogate splits:
      Total_Trans_Ct       < 34.5    to the right, agree=0.783, adj=0.452, (0 split)
      Total_Amt_Chng_Q4_Q1 < 0.5005  to the right, agree=0.655, adj=0.129, (0 split)
      Total_Ct_Chng_Q4_Q1  < 0.288   to the right, agree=0.650, adj=0.118, (0 split)
      Credit_Limit         < 10965.5 to the left,  agree=0.649, adj=0.114, (0 split)
      Avg_Open_To_Buy      < 10965.5 to the left,  agree=0.649, adj=0.114, (0 split)

Node number 9: 115 observations,    complexity param=0.01492537
  predicted class=Existing Customer  expected loss=0.3565217  P(node) =0.01622232
    class counts:    41    74
   probabilities: 0.357 0.643 
  left son=18 (19 obs) right son=19 (96 obs)
  Primary splits:
      Total_Relationship_Count < 2.5     to the left,  improve=15.891310, (0 missing)
      Total_Amt_Chng_Q4_Q1     < 1.0605  to the left,  improve= 6.154691, (0 missing)
      Total_Trans_Amt          < 1975    to the right, improve= 5.025901, (0 missing)
      Total_Ct_Chng_Q4_Q1      < 0.9245  to the left,  improve= 3.032997, (0 missing)
      Contacts_Count_12_mon    < 2.5     to the right, improve= 2.672556, (0 missing)
  Surrogate splits:
      Total_Trans_Amt       < 813.5   to the left,  agree=0.852, adj=0.105, (0 split)
      Total_Trans_Ct        < 16      to the left,  agree=0.852, adj=0.105, (0 split)
      Card_Category         splits as  R--L,        agree=0.843, adj=0.053, (0 split)
      Contacts_Count_12_mon < 4.5     to the right, agree=0.843, adj=0.053, (0 split)

Node number 10: 150 observations,    complexity param=0.01053556
  predicted class=Attrited Customer  expected loss=0.2266667  P(node) =0.02115954
    class counts:   116    34
   probabilities: 0.773 0.227 
  left son=20 (124 obs) right son=21 (26 obs)
  Primary splits:
      Total_Ct_Chng_Q4_Q1    < 0.856   to the left,  improve=15.984930, (0 missing)
      Total_Amt_Chng_Q4_Q1   < 1.173   to the left,  improve=10.107790, (0 missing)
      Total_Trans_Amt        < 1673    to the right, improve= 5.933201, (0 missing)
      Contacts_Count_12_mon  < 2.5     to the right, improve= 4.712712, (0 missing)
      Months_Inactive_12_mon < 1.5     to the right, improve= 4.425051, (0 missing)
  Surrogate splits:
      Total_Amt_Chng_Q4_Q1 < 1.173   to the left,  agree=0.867, adj=0.231, (0 split)
      Customer_Age         < 62      to the left,  agree=0.847, adj=0.115, (0 split)
      Credit_Limit         < 1547.5  to the right, agree=0.833, adj=0.038, (0 split)

Node number 11: 1451 observations,    complexity param=0.01843723
  predicted class=Existing Customer  expected loss=0.1412819  P(node) =0.2046833
    class counts:   205  1246
   probabilities: 0.141 0.859 
  left son=22 (362 obs) right son=23 (1089 obs)
  Primary splits:
      Total_Trans_Amt       < 2102    to the right, improve=54.26276, (0 missing)
      Total_Ct_Chng_Q4_Q1   < 0.587   to the left,  improve=26.21504, (0 missing)
      Total_Revolving_Bal   < 2381.5  to the right, improve=25.21203, (0 missing)
      Gender                splits as  LR,          improve=16.81874, (0 missing)
      Avg_Utilization_Ratio < 0.8225  to the right, improve=15.05920, (0 missing)
  Surrogate splits:
      Total_Trans_Ct        < 48.5    to the right, agree=0.778, adj=0.110, (0 split)
      Customer_Age          < 26.5    to the left,  agree=0.759, adj=0.036, (0 split)
      Months_on_book        < 13.5    to the left,  agree=0.756, adj=0.022, (0 split)
      Avg_Utilization_Ratio < 0.9545  to the right, agree=0.753, adj=0.008, (0 split)
      Contacts_Count_12_mon < 5.5     to the right, agree=0.752, adj=0.006, (0 split)

Node number 12: 187 observations,    complexity param=0.01141352
  predicted class=Attrited Customer  expected loss=0.2566845  P(node) =0.0263789
    class counts:   139    48
   probabilities: 0.743 0.257 
  left son=24 (170 obs) right son=25 (17 obs)
  Primary splits:
      Credit_Limit         < 3484    to the right, improve=14.64064, (0 missing)
      Total_Amt_Chng_Q4_Q1 < 0.833   to the right, improve=14.26025, (0 missing)
      Avg_Open_To_Buy      < 1838    to the right, improve=13.66300, (0 missing)
      Total_Revolving_Bal  < 825     to the left,  improve=13.56658, (0 missing)
      Total_Trans_Ct       < 72.5    to the left,  improve=13.56478, (0 missing)
  Surrogate splits:
      Avg_Open_To_Buy       < 1691    to the right, agree=0.968, adj=0.647, (0 split)
      Avg_Utilization_Ratio < 0.5475  to the left,  agree=0.957, adj=0.529, (0 split)
      Months_on_book        < 55      to the left,  agree=0.920, adj=0.118, (0 split)
      Total_Trans_Amt       < 5451.5  to the right, agree=0.914, adj=0.059, (0 split)

Node number 13: 887 observations
  predicted class=Existing Customer  expected loss=0.03043968  P(node) =0.1251234
    class counts:    27   860
   probabilities: 0.030 0.970 

Node number 16: 400 observations
  predicted class=Attrited Customer  expected loss=0.0675  P(node) =0.05642545
    class counts:   373    27
   probabilities: 0.932 0.067

Node number 17: 263 observations,    complexity param=0.01097454
  predicted class=Attrited Customer  expected loss=0.3726236  P(node) =0.03709973
    class counts:   165    98
   probabilities: 0.627 0.373
  left son=34 (142 obs) right son=35 (121 obs)
  Primary splits:
      Total_Relationship_Count < 3.5     to the left,  improve=23.851240, (0 missing)
      Total_Trans_Amt          < 1097.5  to the left,  improve=19.252760, (0 missing)
      Total_Ct_Chng_Q4_Q1      < 0.5845  to the left,  improve=14.192290, (0 missing)
      Total_Amt_Chng_Q4_Q1     < 0.422   to the left,  improve=11.678920, (0 missing)
      Total_Trans_Ct           < 22.5    to the left,  improve= 9.461445, (0 missing)
  Surrogate splits:
      Total_Ct_Chng_Q4_Q1    < 0.54    to the left,  agree=0.627, adj=0.190, (0 split)
      Total_Trans_Amt        < 1202.5  to the left,  agree=0.620, adj=0.174, (0 split)
      Months_Inactive_12_mon < 2.5     to the right, agree=0.605, adj=0.140, (0 split)
      Total_Trans_Ct         < 36.5    to the left,  agree=0.601, adj=0.132, (0 split)
      Credit_Limit           < 7235.5  to the left,  agree=0.578, adj=0.083, (0 split)

Node number 18: 19 observations
  predicted class=Attrited Customer  expected loss=0.05263158  P(node) =0.002680209
    class counts:    18     1
   probabilities: 0.947 0.053

Node number 19: 96 observations
  predicted class=Existing Customer  expected loss=0.2395833  P(node) =0.01354211
    class counts:    23    73
   probabilities: 0.240 0.760

Node number 20: 124 observations
  predicted class=Attrited Customer  expected loss=0.1209677  P(node) =0.01749189
    class counts:   109    15
   probabilities: 0.879 0.121

Node number 21: 26 observations
  predicted class=Existing Customer  expected loss=0.2692308  P(node) =0.003667654
    class counts:     7    19
   probabilities: 0.269 0.731

Node number 22: 362 observations,    complexity param=0.01843723
  predicted class=Existing Customer  expected loss=0.378453  P(node) =0.05106503
    class counts:   137   225
   probabilities: 0.378 0.622
  left son=44 (174 obs) right son=45 (188 obs)
  Primary splits:
      Total_Ct_Chng_Q4_Q1  < 0.6315  to the left,  improve=39.31964, (0 missing)
      Total_Amt_Chng_Q4_Q1 < 0.9045  to the left,  improve=27.03317, (0 missing)
      Customer_Age         < 39.5    to the right, improve=24.35209, (0 missing)
      Total_Revolving_Bal  < 2384.5  to the right, improve=21.81566, (0 missing)
      Total_Trans_Ct       < 47.5    to the left,  improve=20.42265, (0 missing)
  Surrogate splits:
      Total_Amt_Chng_Q4_Q1 < 0.8465  to the left,  agree=0.699, adj=0.374, (0 split)
      Total_Trans_Ct       < 47.5    to the left,  agree=0.666, adj=0.305, (0 split)
      Credit_Limit         < 4602    to the left,  agree=0.622, adj=0.213, (0 split)
      Avg_Open_To_Buy      < 2519.5  to the left,  agree=0.616, adj=0.201, (0 split)
      Total_Trans_Amt      < 2833.5  to the left,  agree=0.616, adj=0.201, (0 split)

Node number 23: 1089 observations
  predicted class=Existing Customer  expected loss=0.06244261  P(node) =0.1536183
    class counts:    68  1021
```

```
probabilities: 0.062 0.938

Node number 24: 170 observations
  predicted class=Attrited Customer  expected loss=0.1941176  P(node) =0.02398082
    class counts:   137    33
   probabilities: 0.806 0.194

Node number 25: 17 observations
  predicted class=Existing Customer  expected loss=0.1176471  P(node) =0.002398082
    class counts:     2    15
   probabilities: 0.118 0.882

Node number 34: 142 observations
  predicted class=Attrited Customer  expected loss=0.1760563  P(node) =0.02003103
    class counts:   117    25
   probabilities: 0.824 0.176

Node number 35: 121 observations,    complexity param=0.01053556
  predicted class=Existing Customer  expected loss=0.3966942  P(node) =0.0170687
    class counts:    48    73
   probabilities: 0.397 0.603
  left son=70 (12 obs) right son=71 (109 obs)
  Primary splits:
      Total_Trans_Amt       < 995     to the left,  improve=9.697172, (0 missing)
      Total_Amt_Chng_Q4_Q1  < 0.397   to the left,  improve=7.203288, (0 missing)
      Total_Ct_Chng_Q4_Q1   < 0.5695  to the left,  improve=6.110099, (0 missing)
      Avg_Utilization_Ratio < 0.03    to the right, improve=3.546184, (0 missing)
      Education_Level       splits as  LLLLRRR,     improve=3.424205, (0 missing)
  Surrogate splits:
      Total_Trans_Ct < 20.5    to the left,  agree=0.934, adj=0.333, (0 split)

Node number 44: 174 observations,    complexity param=0.01843723
  predicted class=Attrited Customer  expected loss=0.3793103  P(node) =0.02454507
    class counts:   108    66
   probabilities: 0.621 0.379
  left son=88 (127 obs) right son=89 (47 obs)
  Primary splits:
      Customer_Age           < 37.5    to the right, improve=26.13475, (0 missing)
      Months_on_book         < 23.5    to the right, improve=23.55266, (0 missing)
      Total_Revolving_Bal    < 2092    to the right, improve=11.74433, (0 missing)
      Months_Inactive_12_mon < 1.5     to the right, improve=11.33273, (0 missing)
      Total_Trans_Ct         < 44.5    to the left,  improve=10.78010, (0 missing)
  Surrogate splits:
      Months_on_book       < 27.5    to the right, agree=0.897, adj=0.617, (0 split)
      Total_Amt_Chng_Q4_Q1 < 1.1325  to the left,  agree=0.759, adj=0.106, (0 split)
      Dependent_count      < 0.5     to the right, agree=0.747, adj=0.064, (0 split)

Node number 45: 188 observations
  predicted class=Existing Customer  expected loss=0.1542553  P(node) =0.02651996
    class counts:    29   159
   probabilities: 0.154 0.846

Node number 70: 12 observations
  predicted class=Attrited Customer  expected loss=0  P(node) =0.001692763
    class counts:    12     0
   probabilities: 1.000 0.000

Node number 71: 109 observations
  predicted class=Existing Customer  expected loss=0.3302752  P(node) =0.01537593
    class counts:    36    73
   probabilities: 0.330 0.670
```

```
Node number 88: 127 observations
  predicted class=Attrited Customer  expected loss=0.2125984  P(node) =0.01791508
    class counts:   100    27
   probabilities: 0.787 0.213

Node number 89: 47 observations
  predicted class=Existing Customer  expected loss=0.1702128  P(node) =0.00662999
    class counts:     8    39
   probabilities: 0.170 0.830

> #Readable Plot using rpart.plot
> rpart.plot(hit.rtree,tweak=1.5,box.palette="auto")
Warning message:
labs do not fit even at cex 0.15, there may be some overplotting
> #Attempting Pruning
> printcp(hit.rtree) #display crossvalidated error for each tree size

Classification tree:
rpart(formula = Attrition_Flag ~ ., data = training, method = "class")

Variables actually used in tree construction:
[1] Credit_Limit             Customer_Age
[3] Total_Ct_Chng_Q4_Q1      Total_Relationship_Count
[5] Total_Revolving_Bal      Total_Trans_Amt
[7] Total_Trans_Ct

Root node error: 1139/7089 = 0.16067

n= 7089

         CP nsplit rel error  xerror     xstd
1  0.166813      0   1.00000 1.00000 0.027146
2  0.071993      2   0.66637 0.67428 0.022975
3  0.039947      3   0.59438 0.60228 0.021854
4  0.028973      5   0.51449 0.52678 0.020575
5  0.018437      6   0.48551 0.50395 0.020165
6  0.014925      9   0.42142 0.44074 0.018962
7  0.011414     10   0.40650 0.42581 0.018662
8  0.010975     11   0.39508 0.42669 0.018680
9  0.010536     13   0.37313 0.42669 0.018680
10 0.010000     15   0.35206 0.42142 0.018573
> plotcp(hit.rtree) #plot cv error
> #select CP with lowest crossvalidated error
> #we can grab this from the plotcp table automatically with
> opt.cp <- hit.rtree$cptable[which.min(hit.rtree$cptable[,"xerror"]),"CP"]
> hit.rtree.pruned <- prune(hit.rtree, cp=opt.cp)
> rpart.plot(hit.rtree.pruned,tweak=1.5,box.palette="auto")
Warning message:
labs do not fit even at cex 0.15, there may be some overplotting
> #list out variables in order of importance
> varImp(hit.rtree)
                           Overall
Avg_Open_To_Buy          13.662996
Avg_Utilization_Ratio   517.705148
Contacts_Count_12_mon    30.878416
Credit_Limit             14.640642
Customer_Age             59.974279
Education_Level           3.424205
Gender                   16.818737
Marital_Status           20.738352
```

```
Months_Inactive_12_mon    54.640255
Months_on_book            23.552656
Total_Amt_Chng_Q4_Q1     143.127900
Total_Ct_Chng_Q4_Q1      613.697790
Total_Relationship_Count 395.817364
Total_Revolving_Bal      767.291567
Total_Trans_Amt          776.245669
Total_Trans_Ct           564.115521
Dependent_count            0.000000
Income_Category            0.000000
Card_Category              0.000000
> varImp(hit.rtree.pruned)
                            Overall
Avg_Open_To_Buy           13.662996
Avg_Utilization_Ratio    517.705148
Contacts_Count_12_mon     30.878416
Credit_Limit              14.640642
Customer_Age              59.974279
Education_Level            3.424205
Gender                    16.818737
Marital_Status            20.738352
Months_Inactive_12_mon    54.640255
Months_on_book            23.552656
Total_Amt_Chng_Q4_Q1     143.127900
Total_Ct_Chng_Q4_Q1      613.697790
Total_Relationship_Count 395.817364
Total_Revolving_Bal      767.291567
Total_Trans_Amt          776.245669
Total_Trans_Ct           564.115521
Dependent_count            0.000000
Income_Category            0.000000
Card_Category              0.000000
```

4. *Model Testing, Evaluation, and Comparison*

```
> #using testing data to measure accuracy
> probt <- predict(hit.rtree,data=testing,type="class")
> confusionMatrix(data=probt,testing$Attrition_Flag)
Error in table(data, reference, dnn = dnn, ...) :
  all arguments must have the same length
> summary(prob)
Attrited Customer Existing Customer
             1129              5960
> gbag <- bagging(Attrition_Flag ~ ., data = training, coob=TRUE)
> varImp((gbag))
                            Overall
Avg_Open_To_Buy           172.62035
Avg_Utilization_Ratio     627.33617
Card_Category              15.09024
Contacts_Count_12_mon     128.88843
Credit_Limit              170.82539
Customer_Age              214.07951
Dependent_count            72.41552
Education_Level           140.20238
Gender                     62.86606
Income_Category           123.27461
Marital_Status             78.15476
Months_Inactive_12_mon    146.16365
Months_on_book            166.15496
Total_Amt_Chng_Q4_Q1      412.17961
Total_Ct_Chng_Q4_Q1       839.67051
```

```
Total_Relationship_Count  498.15266
Total_Revolving_Bal       894.71615
Total_Trans_Amt          1039.31688
Total_Trans_Ct            711.71256
> probbag <- predict(gbag,testing,type="class")
> confusionMatrix(probbag,testing$Attrition_Flag)
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               410                46
  Existing Customer                78              2504

               Accuracy : 0.9592
                 95% CI : (0.9515, 0.9659)
    No Information Rate : 0.8394
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8445

 Mcnemar's Test P-Value : 0.005371

            Sensitivity : 0.8402
            Specificity : 0.9820
         Pos Pred Value : 0.8991
         Neg Pred Value : 0.9698
             Prevalence : 0.1606
         Detection Rate : 0.1350
   Detection Prevalence : 0.1501
      Balanced Accuracy : 0.9111

       'Positive' Class : Attrited Customer
```