# Executive Summary

Our goal is to identify consumers who are likely to leave credit card services, so that Manager may present them with additional benefits to keep them.

The data structure we had supplied for preprocessing has **10,128** rows and **21** columns, with Attrition Flag as our target variable (whether a customer will churn or not). Before we could run the models, we had to clean and preprocess the data. We dealt with missing values, outliers, modifying column distribution, and binning of particular columns, such as age. We were able to better organize our data because of this. We also ran correlation analysis and were able to eliminate certain highly correlated columns. Using stratified sampling, we partitioned the data into training (50 percent), validation (30 percent), and test (20 percent).

We started with the Logistic regression because the data was suitable for modeling, and we ran the model to acquire the best predictions by eliminating certain low-impacting columns. We observed that the transaction amount and total transaction were significant factors in deciding whether or not a customer would churn. For the sake of simplicity, we decided to run a KNN analysis for the best K=7, and the model was able to identify 62 percent of the attritted customers. Because our data is straightforward and most columns can be classified, we reasoned that using a decision tree would improve predictability and interpretability. We used the decision tree, Bootstrap, and boosted tree modeling approaches, and the results were quite similar. However, utilizing the Bootstrap method, we got an accuracy of 82 percent with a lowest misclassification rate of 4.24 percent. We used a neural network to test if we could enhance accuracy for attrition customers, and we achieved an 81 percent value.

The Bootstrap method deliver the best predictability and interpretability. The manager should look for giving more offers to customers to improve their total transaction amount and total transaction count.

# Problem Statement

A bank manager is concerned that more consumers are canceling their credit card services. After we have completed our analysis, we will be able to quickly identify clients who are likely to churn in the next six months. Using these forecasts, the manager may focus their efforts on providing better products that meet the needs of the clients and keeping them in the system.

# Dataset Information

The Data structure of table – 10,128 Rows and 21 Columns. This dataset consists of **10,128** customer data and the features of the customers are described in the remaining columns. Each row depicts information about one customer.  
**à Target Variable:**

The target variable is ***Attrition\_Flag***. It will tell us if the customer is an existing or attritted customer. We are planning to predict which customers will get churned to proactively go to the customer to provide them with better services and turn customers' decisions in the opposite direction.

# Data Preprocessing

## Missing Values

1. After going through the data, we observed missing values in three columns Education\_Level, Marital\_Status and Income\_Category which were marked as Unknown. From data and business sense, Education \_level and Marital\_Status was not of much significance thus we kept unknown values in these columns and removed Unknown category from Income\_Category as Income of a person is a crucial factor in predicting customer attrition. (Refer Appendix - [Missing Values](#_Missing_Values))

## Outliers

* Customer\_Age - The range of customer\_age is between 73 and 26. Both 73 and 26 are acceptable values as there can be customers aged between 73 and 26 and thus, we did not remove any outliers. (Refer Appendix -
* The values in all the remaining columns were within acceptable range and thus no outliers were removed from any of the remaining columns. (Refer Appendix – [Outliers](#_Outliers))

## Column Distribution and Transformation

We observed that credit limit, Avg\_Open\_To\_Buy, Total\_Trans\_Amt, Avg\_Utilization\_Ratio were not normally distributed. All these columns were then transformed to their logarithmic form using JMP transformations. After performing log transformation, all the above columns now show more normality compared to their original form. (Refer Appendix – [Column Distribution and Transformation](#_Column_Distribution_and))

## Binning of Customer Age Variable

We have divided customer age into different age groups. This will help us to analyze different segments of customers and device customized marketing strategies across all age groups as per their needs. (Refer Appendix - [Binning](#_Binning_of_Customer_Age))

## Correlation

We performed correlation analysis to identify predictors which are related to each other.

We observed the following parameter pairs that are correlated

* Credit\_Limit and Avg\_Open\_To\_Buy are correlated
* Total\_Revolving\_Bal and Avg\_Utilization\_Ratio are correlated
* Total\_Trans\_Amt and Total\_Trans\_Ct are correlated

Based on the model selected, we will drop the columns which are highly correlated based on business requirement. ((Refer Appendix – [Correlation](#_Correlation))

## Splitting data into training/validation/test

In our dataset, we have an unequal distribution of target column. Therefore, we split the data into train (50%), validate (30%) and test (20%) dataset using stratified sampling so that equal proportion of Existing and Attrited customers is maintained in all three datasets. (Refer Appendix – [Splitting](#_Splitting_of_Data))

# Methodology

## Logistics Regression (Refer Appendix for Detailed Information – [Logistic](#_Logistic_Regression))

**Definition:** Logistic regression is used for classifying a new observation, where the class is unknown, into one of the classes based on the values of its predictor variables. The target variable Y is categorical.

* **Running the model with all the parameters, below are the analysis and results**
* **Effect Summary**
* **Total\_Trans\_Ct, Total\_Trans\_Amt and Total\_Relationship\_Count** are top 3 prediction variables.
* **Confusion Matrix**

A picture containing table

Description automatically generated

**Misclassification Percentage**

The misclassification rate for Attrition customers in the Validation data set is

= ((198+82)/2712) \*100 =11.51%

The misclassification rate for Attrition customers in the Test data set is

= ((112+58)/1809) \*100 =9.39%

* **Running the model on removing parameters having a higher p-value**

**Confusion Matrix and Misclassification Percentage**

The misclassification rate for Attrition customers in the Validation data set is

= ((198+93)/2712) \*100 =10.73%

The misclassification rate for Attrition customers in the Test data set is

= ((114+60)/ 1809) \*100 =9.62%

* **Replacing Customer Age with Customer Age Binned variable**

**Confusion Matrix and Misclassification rate**

A picture containing calendar

Description automatically generated

The misclassification rate for Attrition customers in the Validation data set is

= ((189+75)/ 2712) \*100=9.73%

The misclassification rate for Attrition customers in the Test data set is

= ((111+59)/1809) \*100 = 9.4

**Conclusion:** In this sampling**,** the Test data set helps predict the attrition customers better than the validation data set. However, the error rate is 9.73% which is high, so the model did not help make predictions for the attrition customers.

## KNN Analysis (Refer Appendix for Detailed Information – [KNN)](#_KNN_–)

In KNN the new observation is classified based on the neighboring records class labels. In this method we identify k-nearest records in the dataset, we then assign the new record to the predominant class within the nearest neighbors. Best performing model was obtained for K=7

**Confusion Matrix:**

Graphical user interface

Description automatically generated with medium confidence

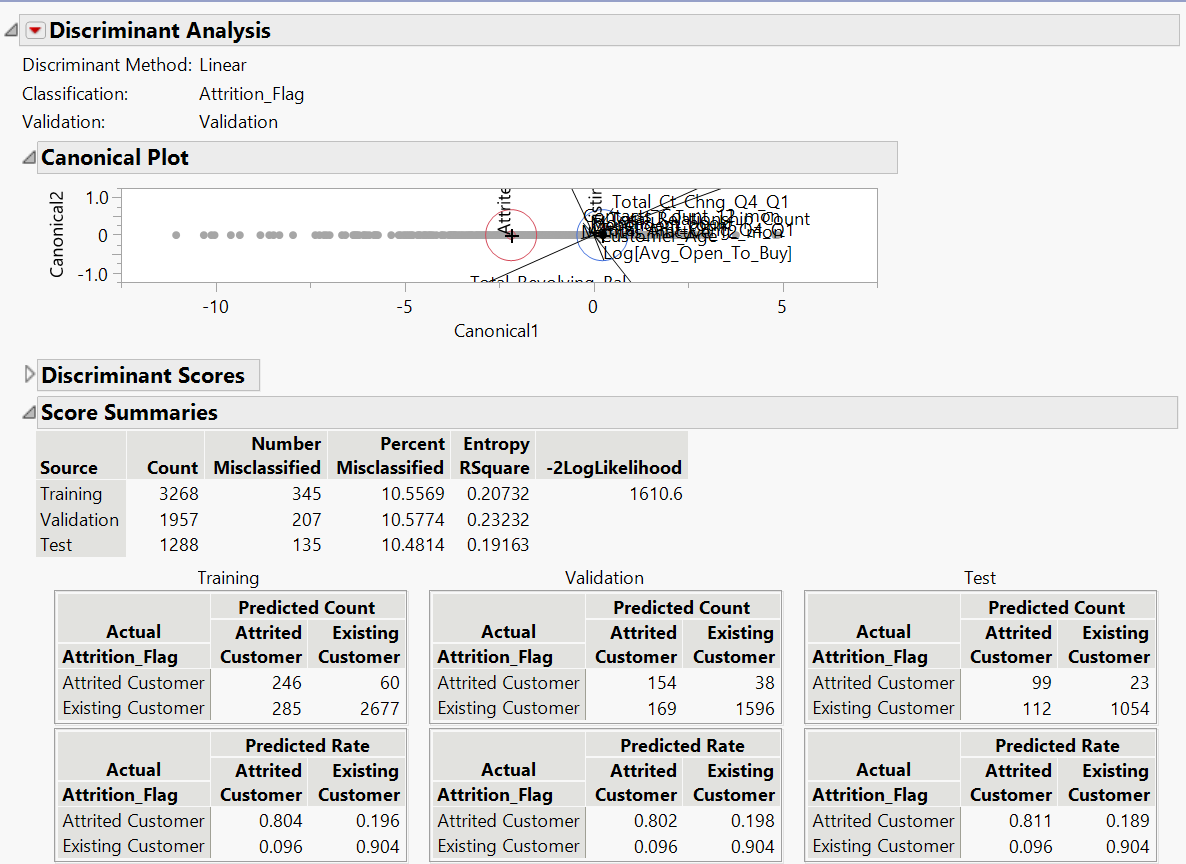
**Misclassification Rate:**

* The misclassification rate for validation dataset is = (163+52) / (269+163+52+2228) = 7.9%
* The misclassification rate for test dataset is = (106+47) / (179+47+106+1477) = 8.4%

**Conclusion**: KNN correctly identified 62% and 63% of Attrited customers. As the Attrited Customers is an important class, we will discard this method.

## Discriminant (*Refer Appendix for Detailed Information –* [*Discriminant*](#_Discriminant)*)*

We performed discriminant analysis by selectively choosing the columns which will lead to the least misclassification rate. As we know that discriminant analysis only takes numerical values as an input and so a lot of columns could not get included in the analysis.



## Decision Tree (Refer Appendix for Detailed Information – [Decision Tree](#_Discriminant))

**Definition:** Decision trees represent a connecting series of tests that branch off further and further down until a specific path matches a class or label.

* **Confusion Matrix and Misclassification Rate**

A picture containing table

Description automatically generated

Misclassification rate for the attrition customers for validation data set is

= ((78+53)/2712) \*100 = 4.83%

Misclassification rate for the attrition customers for testing data set is

= ((45+65)/1809) \*100 = 6.08%

* **Conclusion:** Decision tree is one of the better performing models compared to others as it has low overall misclassification rate, and it can correctly identify approximately 80% of attrited customers.

## Boosted Tree (Refer Appendix for Detailed Information – [Boosted Tree](#_Boosted_Tree))

In Boosted method, series of trees are built so that each tree concentrates on errors from the previous tree. The predictors are randomly selected for each tree. We then sum up all the tree models for the individual tree to produce the boosted model.

Graphical user interface

Description automatically generated with low confidence

**Conclusion** – It is a good classification model because of the low misclassification rate and area under the ROC curve is close to 1. As Attrited customer class is of high importance we will reject this model as other models can identify attrited customers with better accuracy.

## Bootstrap Forest (Refer Appendix for Detailed Information – [Bootstrap Forest](#_Bootstrap_Forest))

**Definition:** To address the shortcomings of a single tree, results are combined from multiple trees. Bootstrapping is a resampling analysis that involves taking columns of characters out of your analysis, rebuilding the tree, and testing if the same nodes are recovered.

**A picture containing graphical user interface

Description automatically generated**

Misclassification rate for attrited customers for Validation data set is

= ((78+39)/2712) \*100 =4.31%

Misclassification rate for attrited customers for Test data set is

= ((24+53)/1809) \*100 = 4.26%

**Conclusion:** It is the best performing model as it has an exceptionally low misclassification rate and AUC value is almost 1, which means it can clearly distinguish between existing and attrited customers. Bootstrap tree can identify attrited customers with more precision than any other model

## Neural Nets (Refer Appendix for Detailed Information – [Neural Nets](#_Neural_network))

This structure supports capturing very complex relationships between predictors and a

response, which is often not possible with other predictive models. The Main idea behind neural networks is to combine the input information in a very flexible way that captures complicated relationships between the input variables and the response.

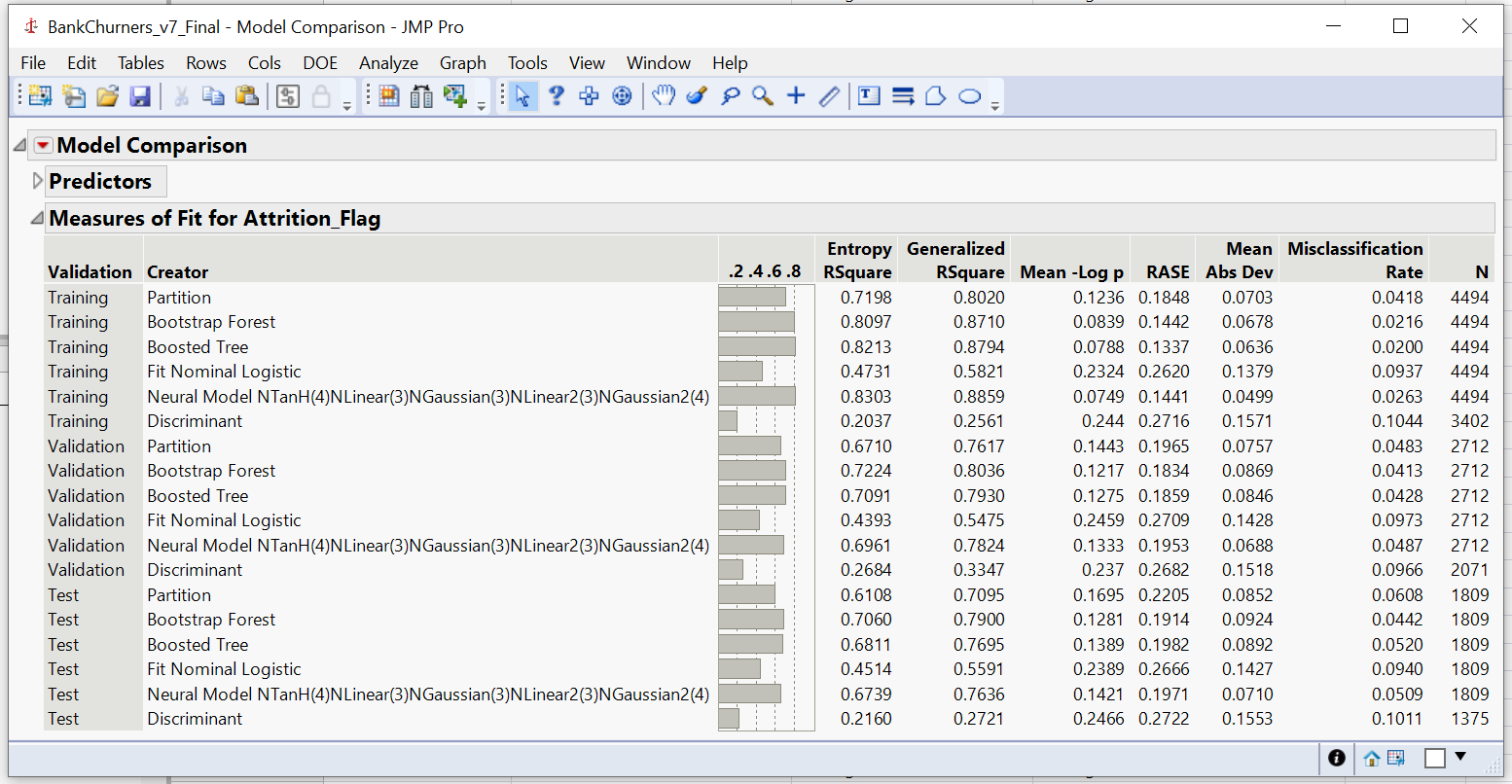
Table

Description automatically generated

**Conclusion –** It is a good classification model because of the low misclassification rate and area under the ROC curve is close to 1. As Attrited customer class is of high importance we will reject this model as Neural Nets are not very helpful since it’s difficult to know about the relationships between Variables and Predictors

# Model Assessment and Selection of Best Model

The top two best performing models are Bootstrap and Neural Network as they have better accuracy and are able to correctly predict more attrited customers (important class) compared to other models. We selected Bootstrap as our final model as Bootstrap tree model is less complex and more interpretable than Neural Network.



# Results and Business Recommendation

We observed that Total\_trans\_Ct, Total\_trans\_Amt and Total\_Revolving\_Bal are the top 3 parameters to differentiate between Existing and Attrited customers. On an average, Existing customers tend to spend more amount, perform more transaction, and have higher revolving balance compared to attrited customers. Thus, our recommendation would be to roll out offers for potential attrited customers to make them spend more and perform more transactions. One example can be providing cashbacks on online and retail shopping to make attrited customers spend more. We can also study the behavior of different age groups using the customer\_age\_binned column to understand which offers would be more susceptible to different age groups based on their transaction history.

# Conclusion:

We performed SEMMA methodology to predict churn customers. We applied several data processing techniques like missing value treatment, handling outliers, column transformation to make the data model ready. Various classification models were built to predict the target variable by selecting the best combination of predictor variables. The best performing model “Bootstrap” was selected as it has the lowest misclassification rate and was able to segregate Attrited customers with more precision.

# References

* We downloaded this dataset from a website with the URL <https://www.kaggle.com/sakshigoyal7/credit-card-customers>.
* Class Presentation Slides and Notes
* Data mining for business analytics book
* We have taken some readings from – <https://www.geeksforgeeks.org/binning-in-data-mining/>
* AUC-ROC curve interpretation - <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>
* The revised JMP is used to build different models – <https://www.jmp.com/en_us/academic/course-materials/regression.html>

# Appendix

#### Column Name and Description –

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Clientnum | Num | Client number. Unique identifier for the customer holding the account |
| Attrition\_Flag | char | Internal event (customer activity) variable - flag to indicate if the customer is an existing customer or has attrited |
| Customer\_Age | Num | Demographic variable - Customer's Age in Years |
| Gender | Char | Demographic variable - M=Male, F=Female |
| Dependent\_count | Num | Demographic variable - Number of dependents |
| Education\_Level | Char | Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.) |
| Marital\_Status | Char | Demographic variable - Married, Single, Unknown |
| Income\_Category | Char | Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown) |
| Card\_Category | Char | Product Variable - Type of Card (Blue, Silver, Gold, Platinum) |
| Months\_on\_book | Num | Months on book (Time of Relationship) |
| Total\_Relationship\_Count | Num | Total no. of products held by the customer |
| Months\_Inactive\_12\_mon | Num | No. of months inactive in the last 12 months |
| Contacts\_Count\_12\_mon | Num | No. of Contacts in the last 12 months |
| Credit\_Limit | Num | Credit Limit on the Credit Card |
| Total\_Revolving\_Bal | Num | Total Revolving Balance on the Credit Card |
| Avg\_Open\_To\_Buy | Num | Open to Buy Credit Line (Average of last 12 months) |
| Total\_Amt\_Chng\_Q4\_Q1 | Num | Change in Transaction Amount (Q4 over Q1) |
| Total\_Trans\_Amt | Num | Total Transaction Amount (Last 12 months) |
| Total\_Trans\_Ct | Num | Total Transaction Count (Last 12 months) |
| Total\_Ct\_Chng\_Q4\_Q1 | Num | Change in Transaction Count (Q4 over Q1) |
| Avg\_Utilization\_Ratio | Num | Average Utilization Ratio |

#### Missing Values

Table

Description automatically generated

#### Outliers

Table

Description automatically generated

#### Column Distribution and Transformation

Box and whisker chart

Description automatically generated with medium confidence Diagram

Description automatically generated A picture containing histogram

Description automatically generatedA picture containing histogram

Description automatically generated

#### Binning of Customer\_Age Variable

Table

Description automatically generated

#### Correlation

Text

Description automatically generated

#### Splitting of Data

Table

Description automatically generated

#### Logistic Regression

Higher p-Values

Table

Description automatically generated

Chart, table

Description automatically generated

**A picture containing graphical user interface

Description automatically generated**

**Effect Summary -**

The Total\_Trans\_Ct, Total\_Trans\_Amt, and Total\_Revolving\_Bal are the top 3 prediction variables.

Chart

Description automatically generated

**Graphical user interface, chart

Description automatically generated**

**Graphical user interface

Description automatically generated**

**ROC**

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy.

**Lift Curve**

In top 30% of the data, the model predicts ~3 times more Attrited customers compared to naïve model.

#### KNN –

We have selected Customer\_Age\_Binned, Total\_Realationship\_Count, Months\_Inactive\_12\_months, Total\_Revolving\_Bal, Total\_Trans\_Amt, Total\_Trans\_Ct predictors for analysis as these columns selection resulted in best performing model.

Chart, line chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

#### Discriminant

Chart, line chart

Description automatically generated

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy.

According to the ROC graph we can see the tradeoff between specificity and sensitivity. Furthermore, it explains how well the model classifies the attrited customers.

A picture containing graphical user interface

Description automatically generated

#### Decision Tree

* **Fit Details**

Text, table

Description automatically generated

* **Leaf Report**

Text

Description automatically generated

* **ROC Curve**

Graphical user interface, chart

Description automatically generated

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy.

* **Lift Curve**

Chart

Description automatically generated

In top 30% of the data, the model predicts ~3 times more attrited customers compared to naïve model.

#### Boosted Tree

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

ROC Curve –

Graphical user interface

Description automatically generated with medium confidence

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy.

Lift Curve –

Graphical user interface, chart, histogram

Description automatically generated

In top 30% of the data, the model predicts ~3 times more attrited customers compared to naïve model.

#### Bootstrap Forest

* **Specifications and Overall Statistics**

**Graphical user interface, text, table

Description automatically generated with medium confidence**

* *Misclassification rate for attrited customers for Validation data set is*

*= ((78+39)/2712) \*100*

*=4.31%*

*Misclassification rate for attrited customers for Test data set is*

*= ((24+53)/1809) \*100*

*= 4.26%*

* **Column Contribution**

Table

Description automatically generated

* Total\_Trans\_Ct, Total\_Trans\_Amt, and Total\_Revolving\_Bal are the top three predictors in predicting the attrited customers.
* **ROC**

**Chart

Description automatically generated**

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy.

* **Lift Curve**

**Graphical user interface, chart

Description automatically generated**In top 30% of the data, the model predicts ~3 times more attrited customers compared to naïve model

#### Neural network

Neural network Diagram

Diagram

Description automatically generated

ROC Curve –

Graphical user interface, chart

Description automatically generated

As the area under the ROC curve is close to 1, the model can segregate Existing and Attrited customers with better accuracy

Lift Curve –

Graphical user interface

Description automatically generated

In top 30% of the data, the model predicts ~3 times more attrited customers compared to naïve model

Other Similar Models –

Table

Description automatically generated

Table

Description automatically generated

#### Model Selection in Details –

The performance of the Logistic model has misclassification rate of 11.51% for the validation dataset. After binning the age variable, the misclassification rate reduced to 9.54%. The top 3 variables with high Log worth are Total\_Trans\_Ct, Total\_Trans\_Amt, Total\_Revolving\_Bal. Logistics Regression Identified 56.3% of the attrited customers correctly.

We performed KNN analysis with K=7 as best, the model predicted 62.3% of the attrited customers correctly for the validation dataset with a misclassification rate of 7.8%, that indicates that KNN model performed better compared to Logistics Regression.

In discriminant analysis, we got a misclassification rate of 10.57% for the validation set and the model predicted 80.2% of the attrited customers correctly.

We used Decision tree in our prediction modeling, and the model predicted 81.9% of the attrited customers correctly with a misclassification rate of 4.83% for validation.