Table of Contents

1.0 Problem Analysis

This chapter presents the problem statements corresponding to the study, the objectives, scope, and the methodology adopted in the study. It sets the stage for the subsequent chapter where the predictive solutions implemented are discussed and documented.

1.1 Problem Statements

The growing number of customer churns for credit card services presents itself as a concerning problem to bank managers. While customer churn is no stranger to most sectors even before two decades ago (Saradhi and Palshikar, 2011), its implications however, are much more prominent in markets that face intense competition, and in cases where the acquisition of new customers is much costlier than the retention of those that are existing. According to Kumar (2022), industry studies demonstrated that efforts aimed at acquiring new customers are five to seven times costlier than that of retaining current customers. The banking sector is one case of such sectors. The most upfront repercussions of customer churn for banks would be the loss in revenue that could have been gained had the customer stayed. The reputation and perception of the bank in the public's eyes would also suffer as a consequence of customers leaving because increasing churn rates would signal to the public that there exist intrinsic problems with the bank, for instance, in terms of its banking processes, customers' experiences which might be sub-optimal, or a lack of competitive offerings by the banks to name a few (Pahul Preet Singh et al., 2024).

Following the abovementioned discussions, it is thereby crucial for bank managers to be able to monitor and manage the customer churn rates in their banks, both of which could be done through the comprehensive understanding of the preferences, needs, and concerns of their customers. In the case of the credit card services in the banking context, these valuable insights obtained would then serve to help banks predict which customers is likely to discontinue their credit card services in the near future. Following that, intervening measures such as direct engagement with the customers could then be introduced and implemented by the banks in the attempt to prevent the churning of those customers (Wu & Li, 2021). While manual analysis has traditionally been the norm with its downside of human errors and the inability to accommodate for scalable solutions, the advancement in advanced analytics and machine learning techniques on the other hand, provides for far more accurate and timely predictions (Fatema Akbar Mohamed & Ali Khalifa Al-Khalifa, 2023). Having said that, the development of models for the purpose of credit card churn predictions will be documented and discussed in this report. Implications of the models' results in the context of bank's credit card services will also be addressed.

1.2 Objectives of the Study

In accordance with the problems identified, the objectives of this study are as follows:

1. To develop a machine learning model for the prediction of customer churn in the credit card domain.
2. To identify factors that significantly influence customer churn rates in the said domain.
3. To equip bank managers with actionable insights and recommendations in better managing credit card churn rates.

1.3 Scope of the Study

The boundary of this study is confined within the area of prediction of customer churn rates in the credit card service domain within the banking sector. The dataset used in this study is from Goyal (2021). The steps involved in the analysis of this study are exploratory data analysis, data preprocessing, predictive modelling, model evaluation, and lastly model interpretation. The preprocessing done on the dataset includes, undersampling, logarithmic transformation, feature selection, data sampling, and data partitioning. All of the preprocessing steps were performed on SAS Enterprise Miner (version 15.2). Hyperparameter tuning was performed on each of the models selected and the details of the tuned parameters will be further elaborated upon in *Section 2.5*. Modelling techniques are limited to tree-based models (Decision Tree, Extreme Gradient Boosting, High-Performance Tree, and High-Performance Forest) and neural networks models (standard Neural Network model and High-Performance Neural Network models). Lastly, the evaluation measures of this study are namely, F1 Score, Precision, Recall, Specificity, Sensitivity, Misclassification Rates (MISC) and the Receiver Operating Characteristic (ROC) curve.

1.4 Methodology of the Study

In this study, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was employed. Due to its systematic approach to data mining that aligns closely with the processes of data analytics, the framework is often adopted not just in the technology domain, but also in many business contexts (Plotnikova et al., 2020). While there are six stages to the framework, this study will only employ the first five stages of the methodology as the deployment stage is out of the study's scope. The five stages are as follows:

1. Business Understanding:

   The initial phase of the framework involves a deep dive into the area of credit card services, with the focus being primarily on the retention of customers. The problems to be addressed, as outlined in the section above, were identified and the objectives and scope of the study are set to correspond to the problems stated.
2. Data Understanding:

   In this phase, exploratory data analysis is performed on the dataset. The purpose is to examine the structure of the dataset, uncover any problematic entries, as well as to explore how the variables relate to one another, especially in relation to the target variable (customer attrition). Descriptive statistics namely the measures of central tendencies (mean, median, and mode), the measure of dispersion (standard deviation), and skewness will be examined in this study. The detection of missing values and inconsistent entries will also be performed in this stage.
3. Data Preparation:

   Once exploratory data analysis has been performed, the dataset will then be cleansed accordingly to prepare it for the next step of modelling. Imputation will be done should there be any missing values detected. If the distributions of the continuous variables are found to be highly skewed, transformation will then be performed. Finally, the dataset will be portioned into training and validation sets, according to the desired train-test ratio, for the subsequent modelling and evaluation stages.
4. Modelling:

   Various predictive models will be generated at this stage using the in-built modelling tools of SAS Enterprise Miner. The models generated can be classified into two general model types namely, tree-based models and neural network models. Each model will then undergo hyperparameter optimization.
5. Evaluation:

   At this last stage of the methodology in the case of this study, each model within the two groups will be evaluated against the chosen performance metrics namely, the F1 score, precision, recall, specificity, sensitivity, the ROC curve, and the AUC statistics to ensure that there is a balanced view of the models' performances. The results of the best models among each group will then be interpreted from a business standpoint. Subsequently, recommendations will be made, highlighting crucial factors influencing customer attrition prediction.

2.0 Solution Development

This chapter presents the process flow of the data analysis done in SAS Enterprise Miner, beginning with the initial exploratory data analysis of the dataset, and ending with the interpretation of the models' results in a business context, alongside the corresponding recommendations based on the model outcomes.

2.1 Data Pipeline in SAS Enterprise Miner

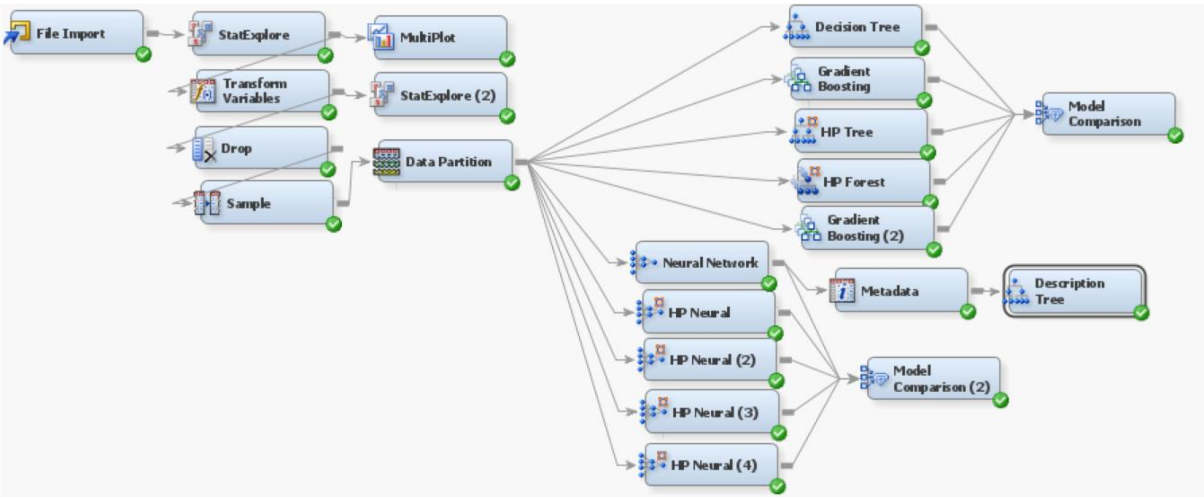


Figure 1: Data Pipeline

The figure above shows the workflow of the data analysis of this study. The process begins with the importing of the Comma-Separated Value (CSV) file into the SAS Enterprise Miner's environment. Following importation, an initial exploratory data analysis is performed using the StatExplore and MultiPlot built-in functions. Preprocessing will the follow where the relevant variables are transformed, dropped, undersampled, and subsequently partitioned into training and validation datasets. Following that, the each of the models shown in the figure will be modelled. Finally, the workflow ends with the Model Comparison node where performances of each model within their respective groups are compared to one another.

2.2 Metadata of Dataset

The dataset contains 16998 records across 21 variables, categorized into 1 ID variable, 19 input variables, and 1 binary target variable. Among the input variables, 5 are nominal, while the rest are interval. The descriptions of what each variable represents are detailed in the table below (*Table 1*). SAS Enterprise Miner displays this similar metadata as well in *Figure 2*, providing a comprehensive overview of the dataset for analysis.

Table 1: Metadata

No.	Name of Variable	Role	Type of Data	Description
1	CLIENTNUM	ID	Nominal	The identification number for each customer that owns a credit card account.
2	Attrition_Flag	Target	Binary	‘1’ represents that the account has been closed while ‘0’ represents that the account is still active.
3	Customer_Age	Input	Interval	The age of the customers (in years).
4	Gender	Input	Nominal	M=Male, F=Female
5	Dependent_count	Input	Ordinal	The total number of dependents of the account holder.
6	Education_Level	Input	Ordinal	Academic qualification of the account holder (College, Doctorate, Graduate, ‘High School’, Post-Graduate, Uneducated, and Unknown).
7	Marital_Status	Input	Nominal	Marital status of the account holder (Married, Single, Divorced, or Unknown).
8	Income_Category	Input	Ordinal	Yearly earnings bracket for the account holder (‘$120K +’, ‘$40K - $60K’, ‘$60K - $80K’, ‘$80K - $120K’, ‘Less than $40K’, and Unknown).
9	Card_Category	Input	Nominal	Category that the credit card belongs to (Blue, Silver, Gold, or Platinum).
10	Months_on_book	Input	Ordinal	Duration of relationship with banks (in months).

11	Total_Relationship_Count	Input	Ordinal	Amount of the bank’s products owned by the customer.
12	Months_Inactive_12_mon	Input	Ordinal	Number of months that the account holder is inactive in the last 12 months.
13	Contacts_Count_12_mon	Input	Ordinal	The number of contacts the account holder had with the bank in the last 12 months.
14	Credit_Limit	Input	Interval	Maximum spending limit available on the card.
15	Total_Revolving_Bal	Input	Interval	Total balance carried over on the credit card from month to month.
16	Avg_Open_To_Buy	Input	Interval	Available credit for purchases (average of last 12 months).
17	Total_Amt_Chng_Q4_Q1	Input	Interval	Difference in transaction amount (Q4 over Q1).
18	Total_Trans_Amt	Input	Interval	Sum of transaction amount (past 12 months).
19	Total_Trans_Ct	Input	Interval	Number of transactions made by the account holder (past 12 months).
20	Total_Ct_Chng_Q4_Q1	Input	Interval	Difference in transaction count (Q4 over Q1).
21	Avg_Utilization_Ratio	Input	Interval	Card utilization ratio on average.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Attrition_Flag	Target	Binary	No		No	.	.
Avg_Open_To_Buy	Input	Interval	No		No	.	.
Avg_Utilization_Ratio	Input	Interval	No		No	.	.
CLIENTNUM	ID	Nominal	No		No	.	.
Card_Category	Input	Nominal	No		No	.	.
Contacts_Count_12_mon	Input	Ordinal	No		No	.	.
Credit_Limit	Input	Interval	No		No	.	.
Customer_Age	Input	Interval	No		No	.	.
Dependent_count	Input	Ordinal	No		No	.	.
Education_Level	Input	Ordinal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Income_Category	Input	Ordinal	No		No	.	.
Marital_Status	Input	Nominal	No		No	.	.
Months_Inactive_12_mon	Input	Ordinal	No		No	.	.
Months_on_book	Input	Ordinal	No		No	.	.
Total_Amt_Chng_Q4_Q1	Input	Interval	No		No	.	.
Total_Ct_Chng_Q4_Q1	Input	Interval	No		No	.	.
Total_Relationship_Count	Input	Ordinal	No		No	.	.
Total_Revolving_Bal	Input	Interval	No		No	.	.
Total_Trans_Amt	Input	Interval	No		No	.	.
Total_Trans_Ct	Input	Interval	No		No	.	.

Figure 2: Metadata Displayed in SAS Enterprise Miner's Environment

2.3 Exploratory Data Analysis (EDA)

2.3.1 EDA using StatExplore

This section documents the steps performed for the exploratory data analysis of the dataset.

Table 2: EDA using StatExplore

<table>
<tr><th>Exploration</th><th>Results of Exploration</th><th>Findings</th></tr>
<tr>
<td>Summary Statistics for Class Variables</td>
<td>
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN
<table>
<tr><th>Data Role</th><th>Variable Name</th><th>Role</th><th>Number of Levels</th><th>Missing</th><th>Mode</th><th>Mode Percentage</th><th>Mode2</th><th>Mode2 Percentage</th></tr>
<tr><td>TRAIN</td><td>Card_Category</td><td>INPUT</td><td>4</td><td>0</td><td>Blue</td><td>93.18</td><td>Silver</td><td>5.48</td></tr>
<tr><td>TRAIN</td><td>Contacts_Count_12_mon</td><td>INPUT</td><td>7</td><td>0</td><td>3</td><td>33.38</td><td>2</td><td>31.87</td></tr>
<tr><td>TRAIN</td><td>Dependent_count</td><td>INPUT</td><td>6</td><td>0</td><td>3</td><td>26.98</td><td>2</td><td>26.22</td></tr>
<tr><td>TRAIN</td><td>Education_Level</td><td>INPUT</td><td>7</td><td>0</td><td>Graduate</td><td>30.89</td><td>High School</td><td>19.88</td></tr>
<tr><td>TRAIN</td><td>Gender</td><td>INPUT</td><td>2</td><td>0</td><td>F</td><td>52.91</td><td>M</td><td>47.09</td></tr>
<tr><td>TRAIN</td><td>Income_Category</td><td>INPUT</td><td>6</td><td>0</td><td>Less than $40K</td><td>35.16</td><td>$40K - $60K</td><td>17.68</td></tr>
<tr><td>TRAIN</td><td>Marital_Status</td><td>INPUT</td><td>4</td><td>0</td><td>Married</td><td>46.28</td><td>Single</td><td>38.94</td></tr>
<tr><td>TRAIN</td><td>Months_Inactive_12_mon</td><td>INPUT</td><td>7</td><td>0</td><td>3</td><td>37.98</td><td>2</td><td>32.41</td></tr>
<tr><td>TRAIN</td><td>Months_on_book</td><td>INPUT</td><td>44</td><td>0</td><td>36</td><td>24.32</td><td>37</td><td>3.54</td></tr>
<tr><td>TRAIN</td><td>Total_Relationship_Count</td><td>INPUT</td><td>6</td><td>0</td><td>3</td><td>22.76</td><td>4</td><td>18.88</td></tr>
<tr><td>TRAIN</td><td>Attrition_Flag</td><td>TARGET</td><td>2</td><td>0</td><td>0</td><td>83.93</td><td>1</td><td>16.07</td></tr>
</table>
</td>
<td>
▪ No missing values were detected and the roles for each variable were correctly assigned.
▪ The most common card category is ‘Blue’ (93.18%).
▪ Most customers contacted the bank 3 times in the last 12 months and most of them have a total of 3 dependents.
▪ ‘Graduate’ (30.89%) is the most prevalent academic attainment in the pool of credit card account holders, followed by ‘High School’ (19.88%).
</td>
</tr>
</table>

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Card_Category	INPUT	4	0	Blue	93.18	Silver	5.48
TRAIN	Contacts_Count_12_mon	INPUT	7	0	3	33.38	2	31.87
TRAIN	Dependent_count	INPUT	6	0	3	26.98	2	26.22
TRAIN	Education_Level	INPUT	7	0	Graduate	30.89	High School	19.88
TRAIN	Gender	INPUT	2	0	F	52.91	M	47.09
TRAIN	Income_Category	INPUT	6	0	Less than $40K	35.16	$40K - $60K	17.68
TRAIN	Marital_Status	INPUT	4	0	Married	46.28	Single	38.94
TRAIN	Months_Inactive_12_mon	INPUT	7	0	3	37.98	2	32.41
TRAIN	Months_on_book	INPUT	44	0	36	24.32	37	3.54
TRAIN	Total_Relationship_Count	INPUT	6	0	3	22.76	4	18.88
TRAIN	Attrition_Flag	TARGET	2	0	0	83.93	1	16.07

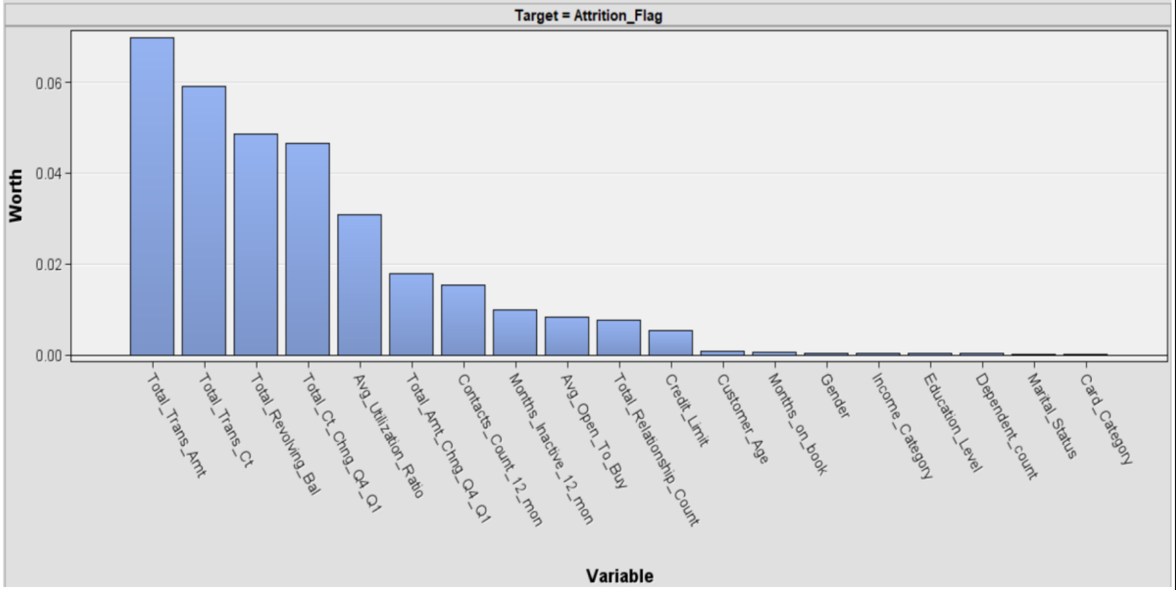
- The gender distribution is almost equal with females showing slightly higher percentage (52.91%) than males (47.09%).
- Most customers (35.16%) have yearly earnings lesser than $ 40K.
- Most customers are married (51.79%) and most of them remained inactive for a duration of 3 months within the past year.
- Most of the customers maintained their relationships with the bank for a period of 36 months and the majority of them own 3 products with the bank.
- There is a significant class imbalance among the classes of the target variable, which is 83.93% of those who have not churned against 16.07% of those

<table>
<tr><td></td><td></td><td>who have churned. Class balancing through undersampling will be performed in Section 2.4.</td></tr>
<tr><td>Summary Statistics for Interval Variables</td><td>
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN
<table>
<tr><th>Variable</th><th>Role</th><th>Mean</th><th>Standard Deviation</th><th>Non Missing</th><th>Missing</th><th>Minimum</th><th>Median</th><th>Maximum</th><th>Skewness</th><th>Kurtosis</th></tr>
<tr><td>Avg_Open_To_Buy</td><td>INPUT</td><td>7469.14</td><td>9090.685</td><td>10127</td><td>0</td><td>3</td><td>3474</td><td>34516</td><td>1.661697</td><td>1.798617</td></tr>
<tr><td>Avg_Utilization_Ratio</td><td>INPUT</td><td>0.274894</td><td>0.275691</td><td>10127</td><td>0</td><td>0</td><td>0.176</td><td>0.999</td><td>0.718008</td><td>-0.79497</td></tr>
<tr><td>Credit_Limit</td><td>INPUT</td><td>8631.954</td><td>9088.777</td><td>10127</td><td>0</td><td>1438.3</td><td>4549</td><td>34516</td><td>1.666726</td><td>1.808989</td></tr>
<tr><td>Customer_Age</td><td>INPUT</td><td>46.32596</td><td>8.016814</td><td>10127</td><td>0</td><td>26</td><td>46</td><td>73</td><td>-0.03361</td><td>-0.28862</td></tr>
<tr><td>Total_Amt_Chng_Q4_Q1</td><td>INPUT</td><td>0.759941</td><td>0.219207</td><td>10127</td><td>0</td><td>0</td><td>0.736</td><td>3.397</td><td>1.732063</td><td>9.993501</td></tr>
<tr><td>Total_Ct_Chng_Q4_Q1</td><td>INPUT</td><td>0.712222</td><td>0.238086</td><td>10127</td><td>0</td><td>0</td><td>0.702</td><td>3.714</td><td>2.064031</td><td>15.68929</td></tr>
<tr><td>Total_Revolving_Bal</td><td>INPUT</td><td>1162.814</td><td>814.9873</td><td>10127</td><td>0</td><td>0</td><td>1276</td><td>2517</td><td>-0.14884</td><td>-1.14599</td></tr>
<tr><td>Total_Trans_Amt</td><td>INPUT</td><td>4404.086</td><td>3397.129</td><td>10127</td><td>0</td><td>510</td><td>3899</td><td>18484</td><td>2.041003</td><td>3.894023</td></tr>
<tr><td>Total_Trans_Ct</td><td>INPUT</td><td>64.85869</td><td>23.47257</td><td>10127</td><td>0</td><td>10</td><td>67</td><td>139</td><td>0.153673</td><td>-0.36716</td></tr>
</table>
</td><td>

No missing values were detected and the roles for each variable were correctly assigned.
Most variables follow an almost normal distribution.
Variables with skewed distributions are namely, Avg_Open_To_Buy, Credit_Limit, Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1, and Total_Trans_Amt. These variables will be treated using logarithmic transformation. The threshold set for identifying skewness in the distribution is between -1 and 1.

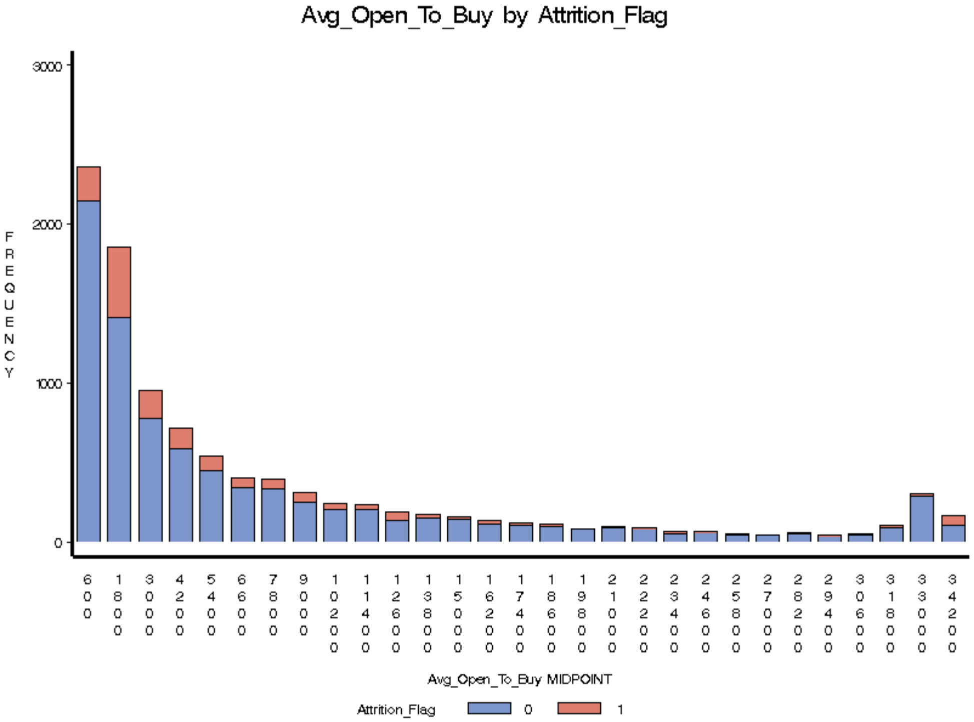
</td></tr>
</table>

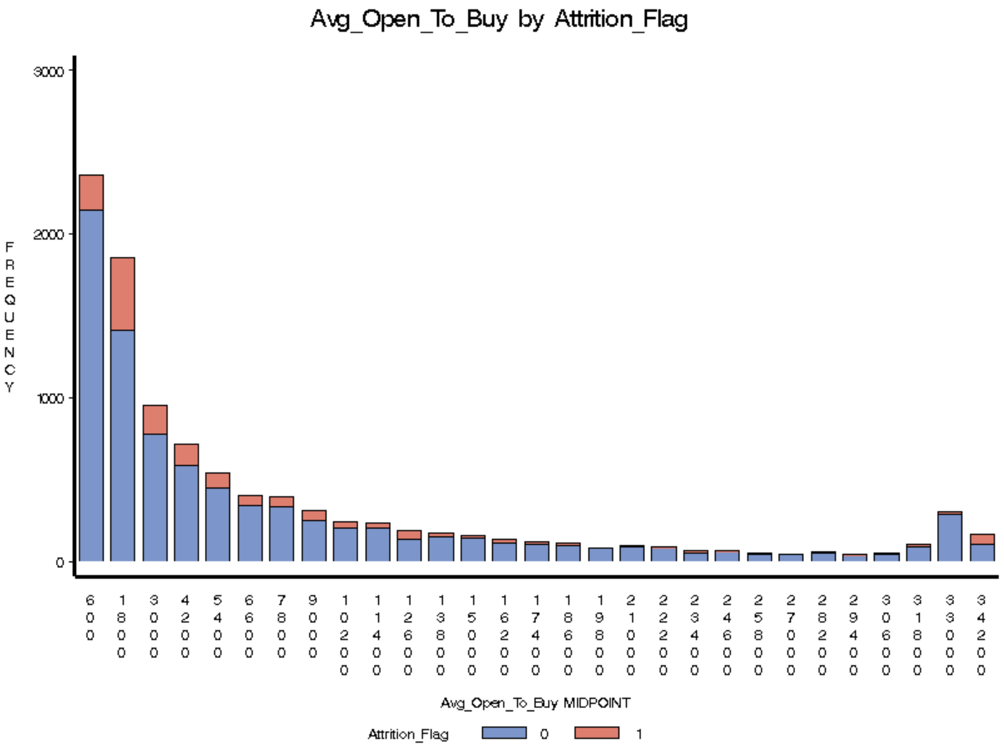
Variable Worth		▪ The Variable Worth figure shows the contributions of each variable in predicting customer attrition. ▪ Customer_Age, Months_on_book, Gender, Income_Category, Education_Level, Dependent_Count, Marital_Status, and Card_Category appear to have the minimal influence on the prediction. ▪ Total_Trans_Amt has the highest worth with a value of 0.070012, followed by Total_Trans_Ct, and Total_Revolving_Bal.

2.3.2 EDA using MultiPlot

In this subsection, the MultiPlot built-in function of SAS Enterprise Miner is employed to perform the necessary EDA. All the graphs generated below represents the distribution of each variable grouped by the target variable of customer attrition. The blue bars in the histogram represent customers who have not churned while the red bars represent customers who have churned.

Table 3: EDA using MultiPlot

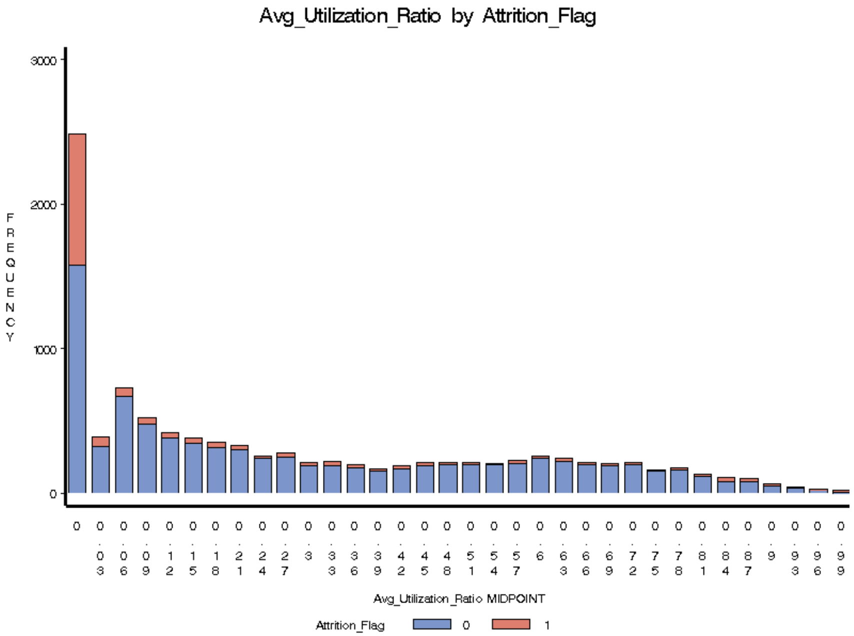
Name of Variable	Results of Exploration	Findings
Avg_Open_To_Buy		▪ The distribution is skewed to the right, with most customers being concentrated at the lower Avg_Open_To_Buy range. ▪ The frequencies generally decrease as Avg_Open_To_Buy values increase for both attrition categories. ▪ Individuals with no attrition (attrition flag 0) are more frequent in lower Avg_Open_To_Buy values than those with attrition

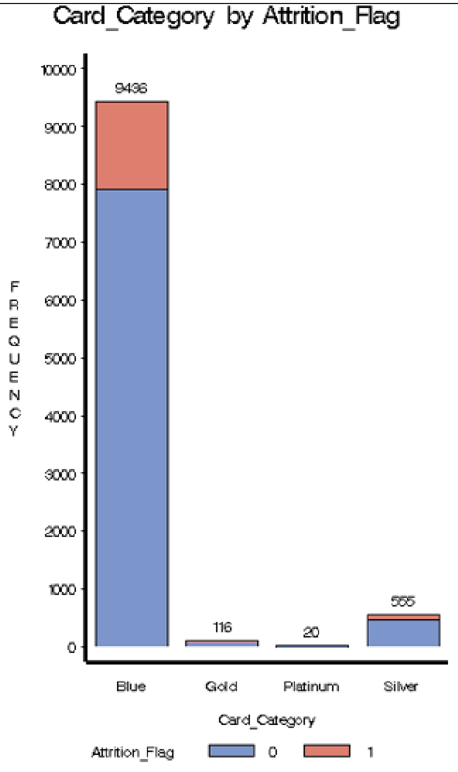
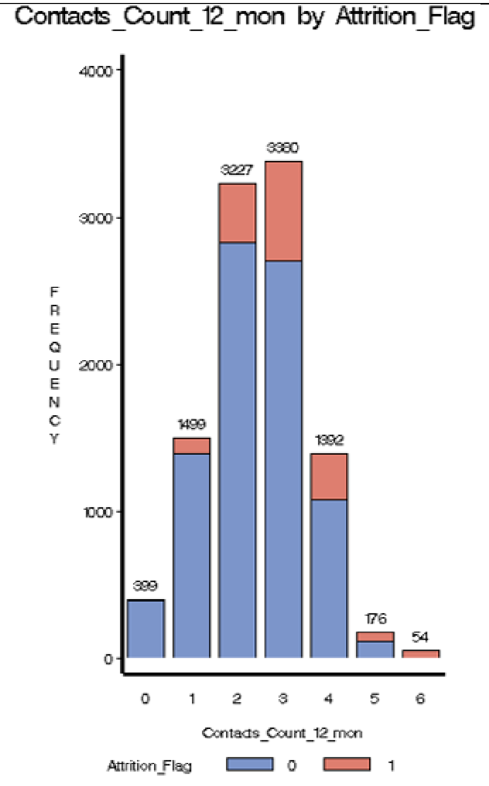
<table>
<tr>
<td></td>
<td>


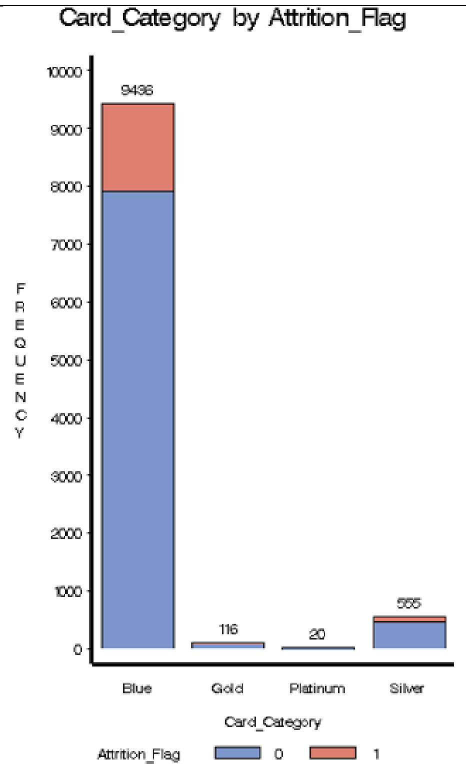
</td>
<td>
(attrition flag 1), implying that customers who stayed with the bank tend to have lower average available credit in comparison to those who have left.

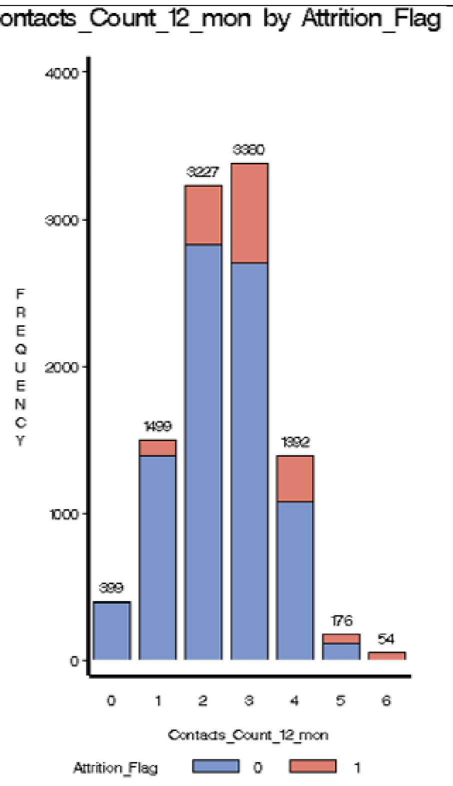
Higher Avg_Open_To_Buy values are uncommon among both attrition groups.
The general trend is homogenous for individuals regardless of attrition status, with most customers having lower Avg_Open_To_Buy capacity.

</td>
</tr>
</table>

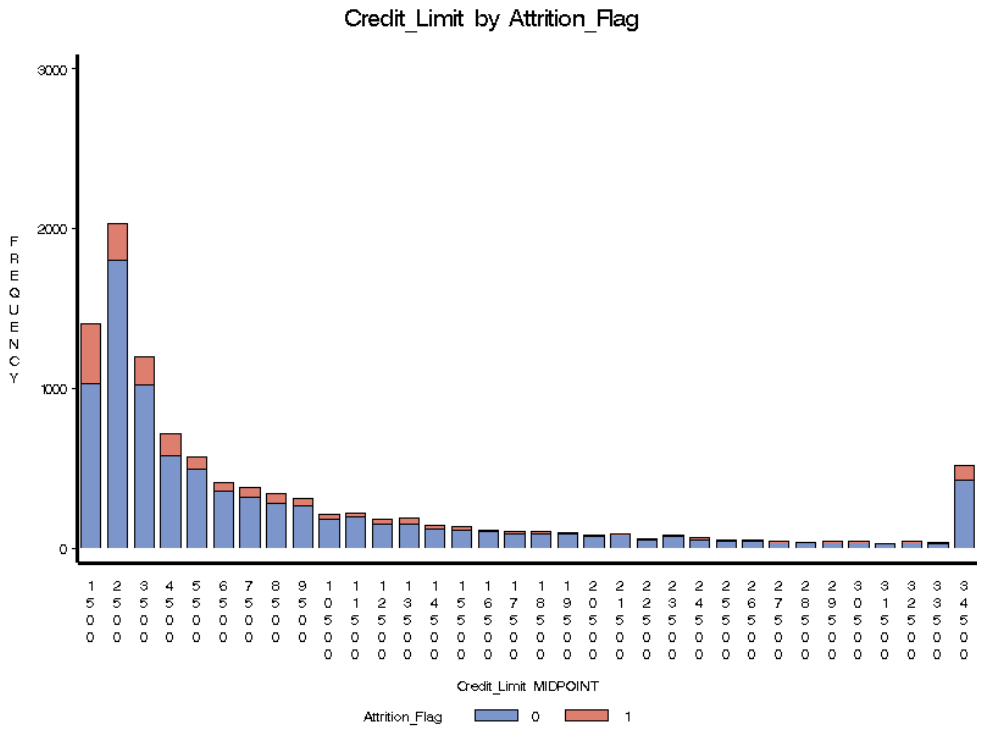
Avg_Utilization_Ratio		▪ Even though the distribution seems like it is skewed to the right, its skewness value of 0.718008 (*Section 2.3.1*) does however, fall within the acceptable threshold of -1 to 1. ▪ Majority of the customers with low Avg_Utilization_Ratio values consisted of customers who did not churn. ▪ As the Avg_Utilization_Ratio increases, frequencies for both groups of customers decrease as well. ▪ Overall, the trend is that majority of the customers, regardless of their attrition status, have Avg_Utilization_Ratio values that are low.

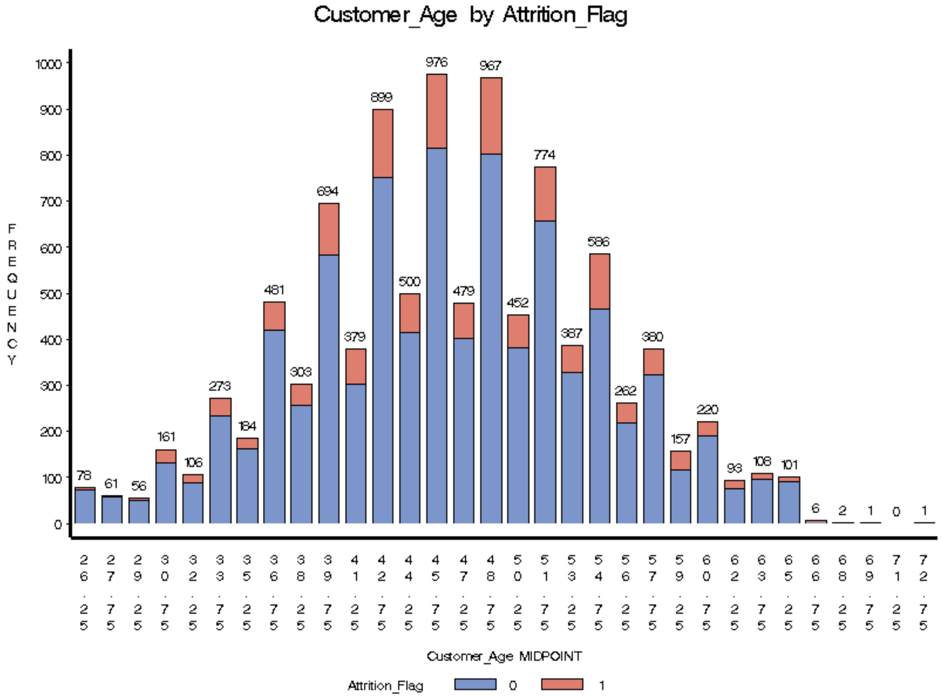
Card_Category and Contact_Count_12_mon	  	▪ Discussion on the skewness of the distributions will not be relevant for the Card_Category and Contact_Count_12_mon variables because they are non-continuous variables. ▪ The Blue card category has he highest frequency of customers with majority of them not experiencing attrition. ▪ Gold, Platinum, and Silver categories has significantly fewer customers, with attrition happening but to a considerably lesser extent. ▪ The highest frequency of customer contacts in the past one-year falls at 3 contacts for both groups of customer attrition.

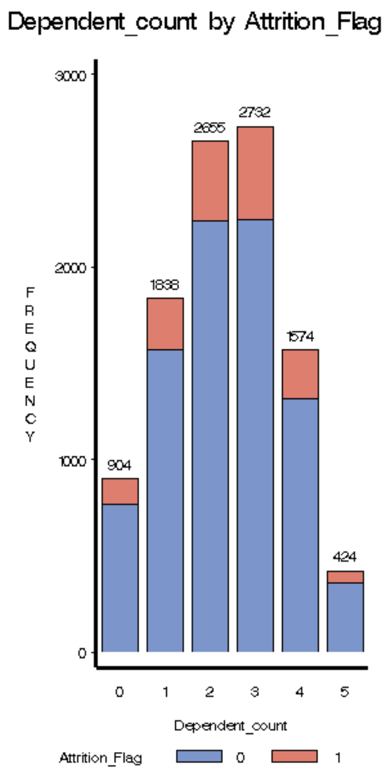


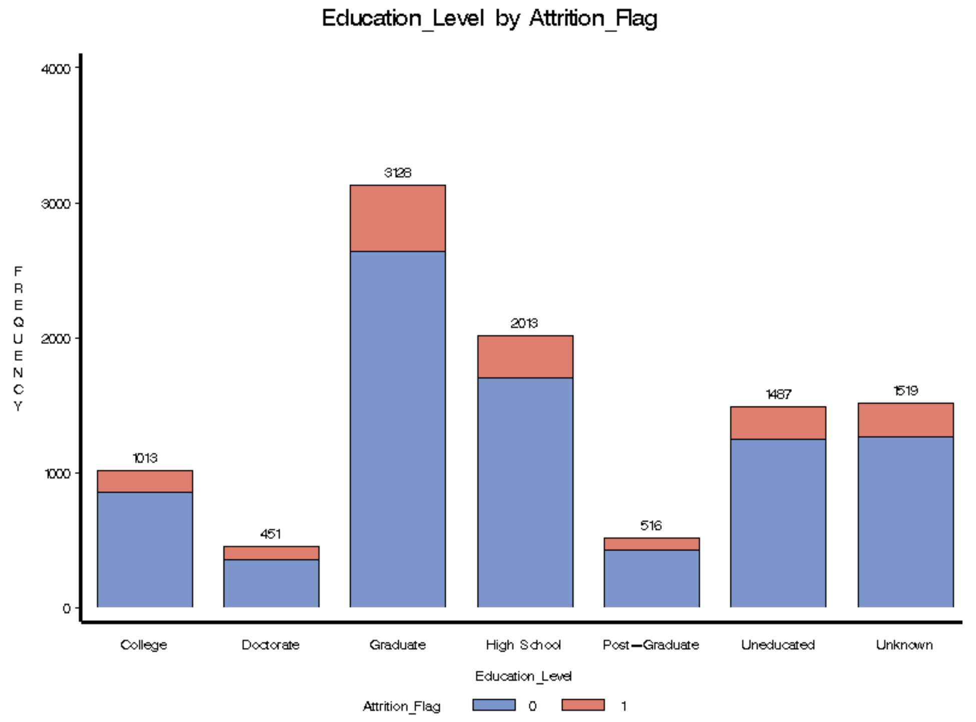


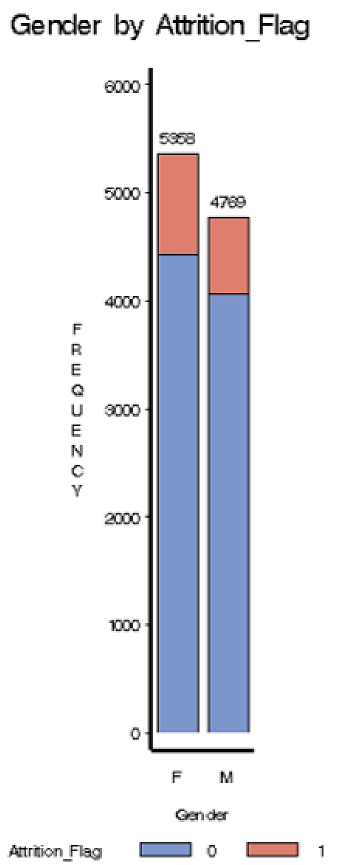
- The frequency of contacts drastically decreases for higher contact counts (5 and 6) in the last 12 months for both groups of customers.
- All customers who contacted the bank 6 times in the past year have churned.

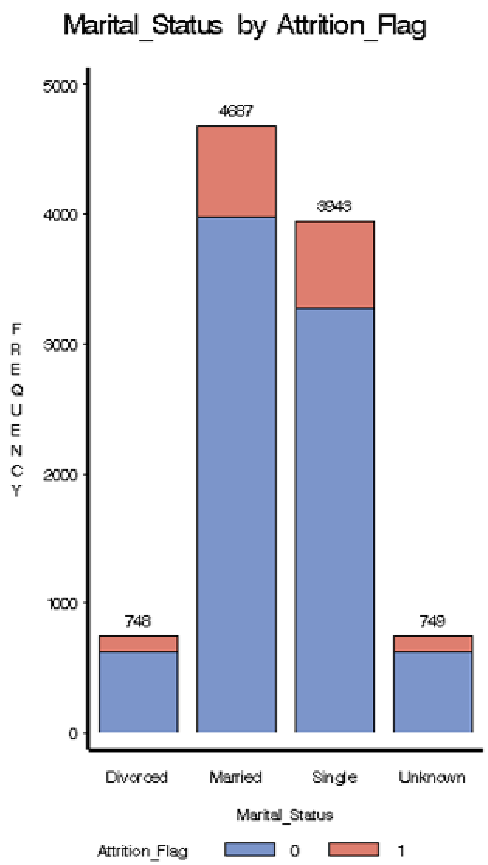
Credit_Limit		▪ The distribution is skewed to the right, with most customers being concentrated at the lower Credit_Limit range. ▪ The frequencies decrease as the maximum spending limit available on the card increases, regardless of the customer attrition status. ▪ The sudden peak at the end suggests that a considerable number of customers have credit limits near the upper limit value, possibly due to the bank's policy of setting a maximum credit limit for certain types of accounts. ▪ The overall distribution suggests that most customers have lower Credit_Limit values

		regardless of their attrition status.
Customer_Age		▪ The distribution follows an almost normal distribution. ▪ Both the frequencies for churned and non-churned customers are the highest in the middle age ranges while younger and older age groups show lower frequencies of both churn and non-churn rates. ▪ The overall patterns suggests that middle-age are the bank's primary demographic, with relatively stable retention rates across these ages.

Dependent_count		▪ Discussion on the skewness of the distribution will not be relevant for the Dependent_count variable because it is a non-continuous variable. ▪ There is an obvious decrease in the frequency of customers with 4 dependents and beyond, more so for those who did not churn. ▪ The largest group of customers, for both attrition status, have 2 to 3 dependents. ▪ The frequency of those who have churned (attrition flag 1) is constantly less than those who did not churn (attrition flag 0).

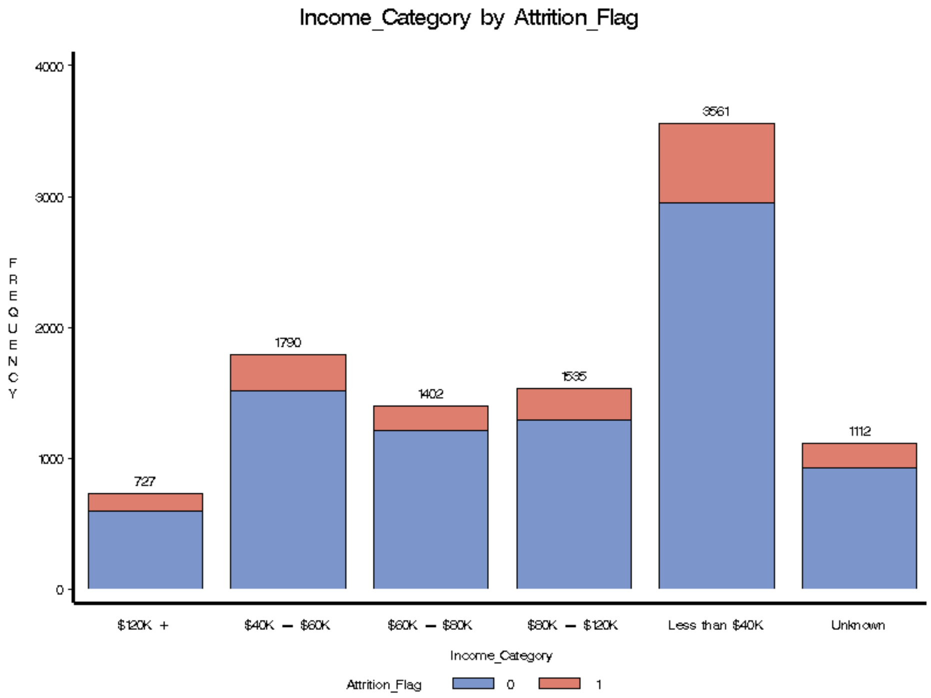
Education_Level		▪ Discussion on the skewness of the distribution will not be relevant for the Education_Level variable because it is a non-continuous variable. ▪ Across all of the academic qualifications, the number of customers who have not churned is consistently higher than those that have churned. ▪ The ‘Unknown’ category which has a considerable number of customers demonstrates that academics data might not be available for the bank’s customers.

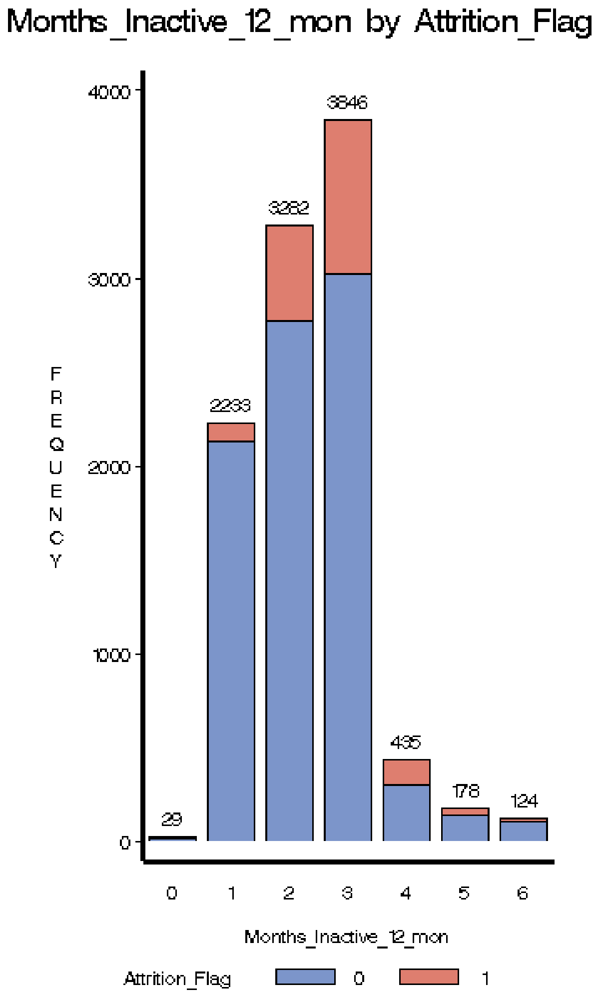
<table>
<tr>
<td>Gender and Marital_Status</td>
<td>

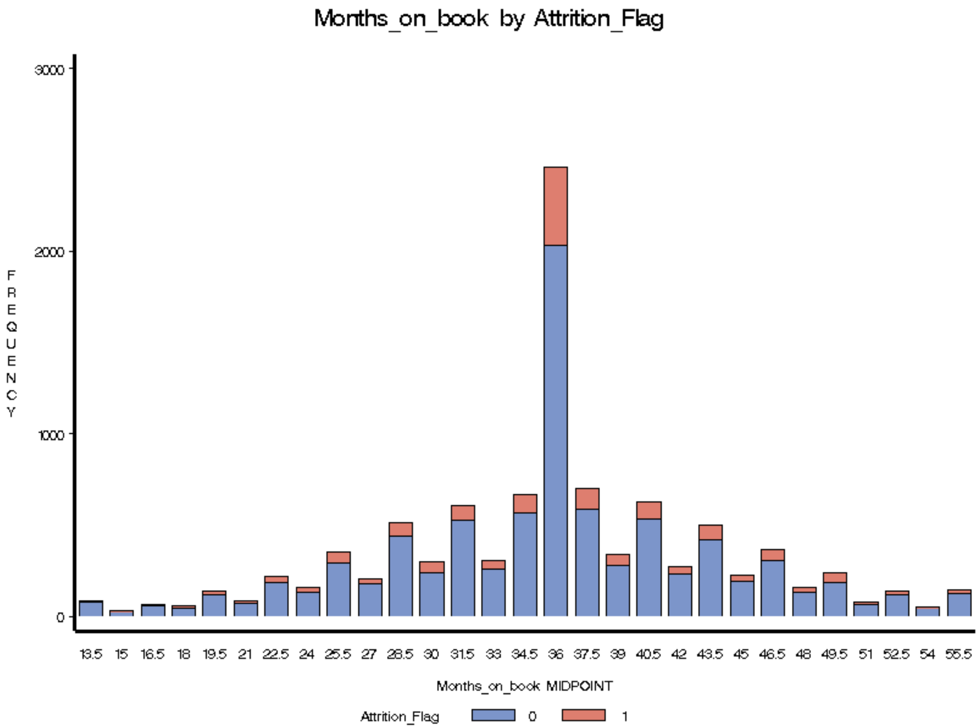

</td>
<td>

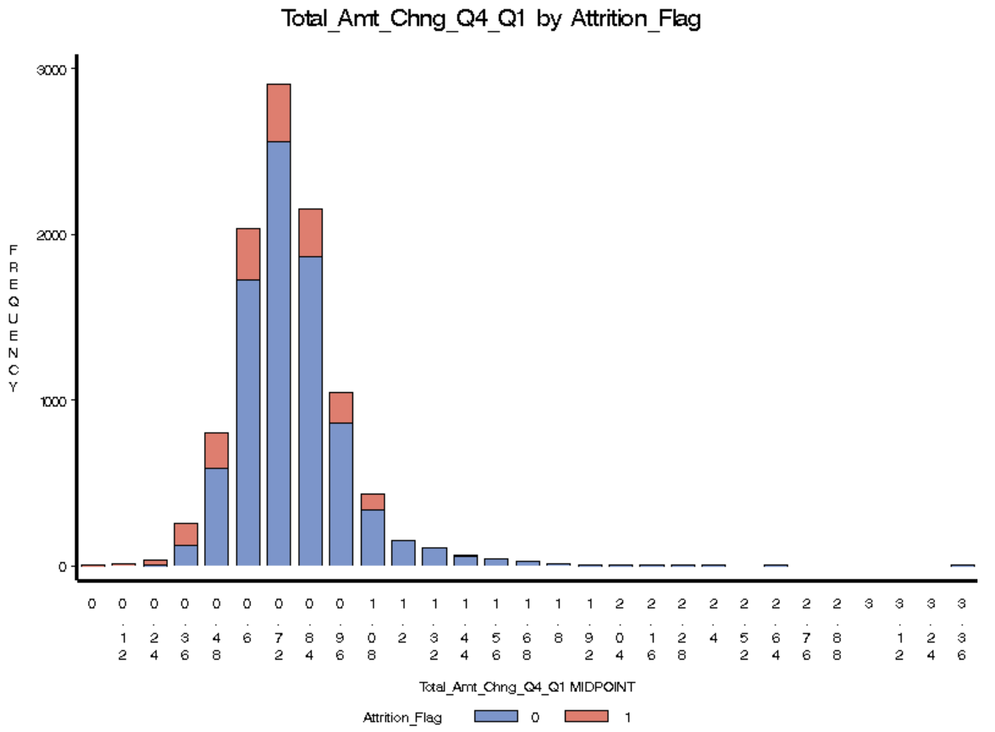
Discussion on the skewness of the distributions will not be relevant for the Gender and Marital_Status variables because they are non-continuous variables.
The proportion of those who did not churn is consistently higher than those who did churn across the respective groups within the Gender and Marital_Status variables.
Churn rates are relatively higher in married and single customers as opposed to those who are divorced or have indicated ‘Unknown’ as their marital status.

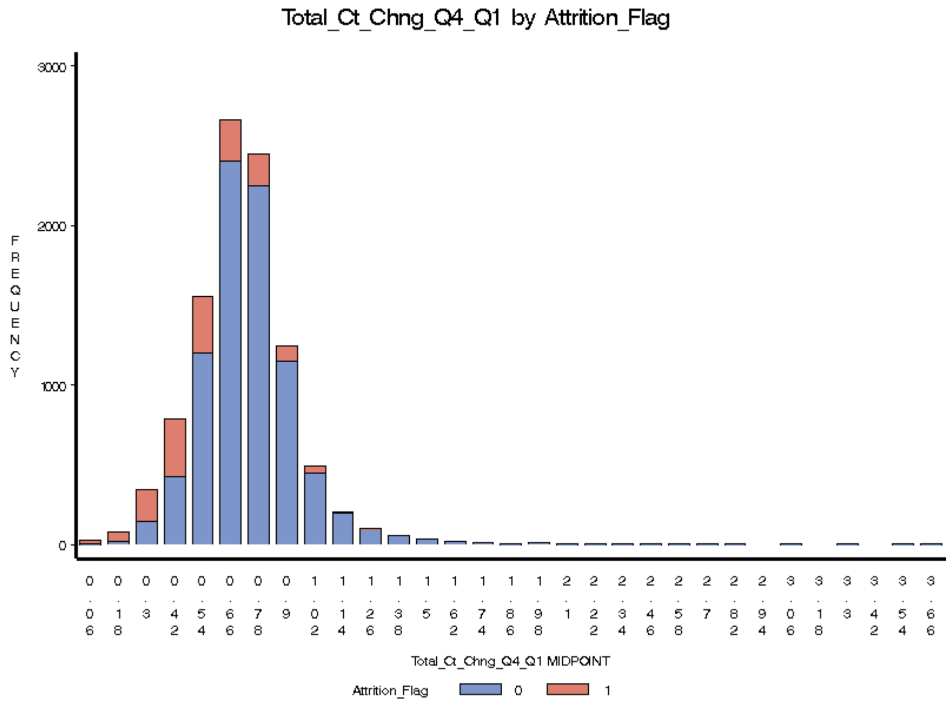
</td>
</tr>
</table>

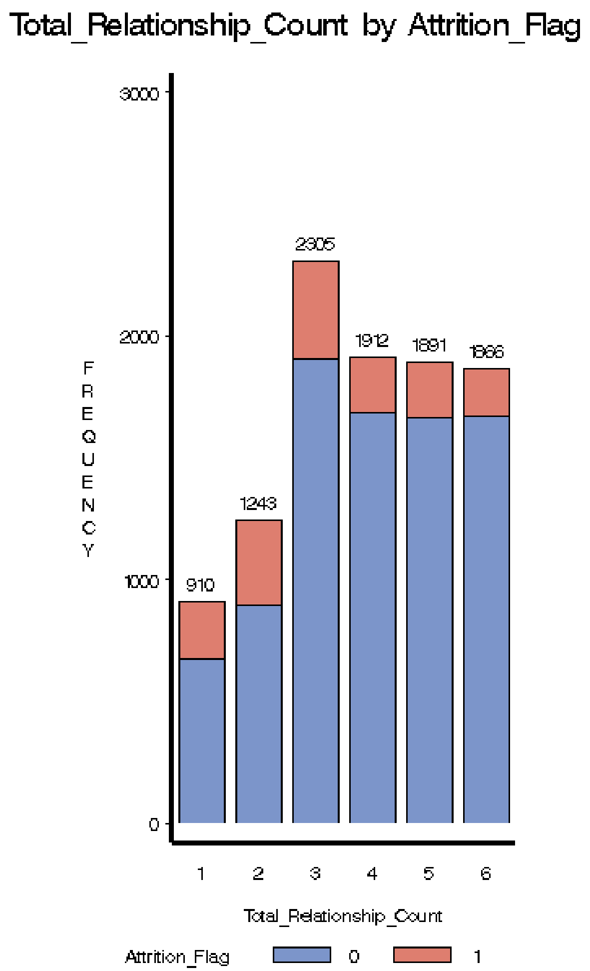
Income_Category	 	▪ Discussion on the skewness of the distribution will not be relevant for the Income_Category variable because it is a non-continuous variable. ▪ Across all income brackets, the number of customers who have not churned outnumbered those who have churned, indicating higher customer retention rates across all income levels. ▪ While the income bracket of ‘Less than $40 K’ makes up the majority of the bank’s customers, this income bracket also had the largest number of customers who have churned compared to the rest of the brackets.

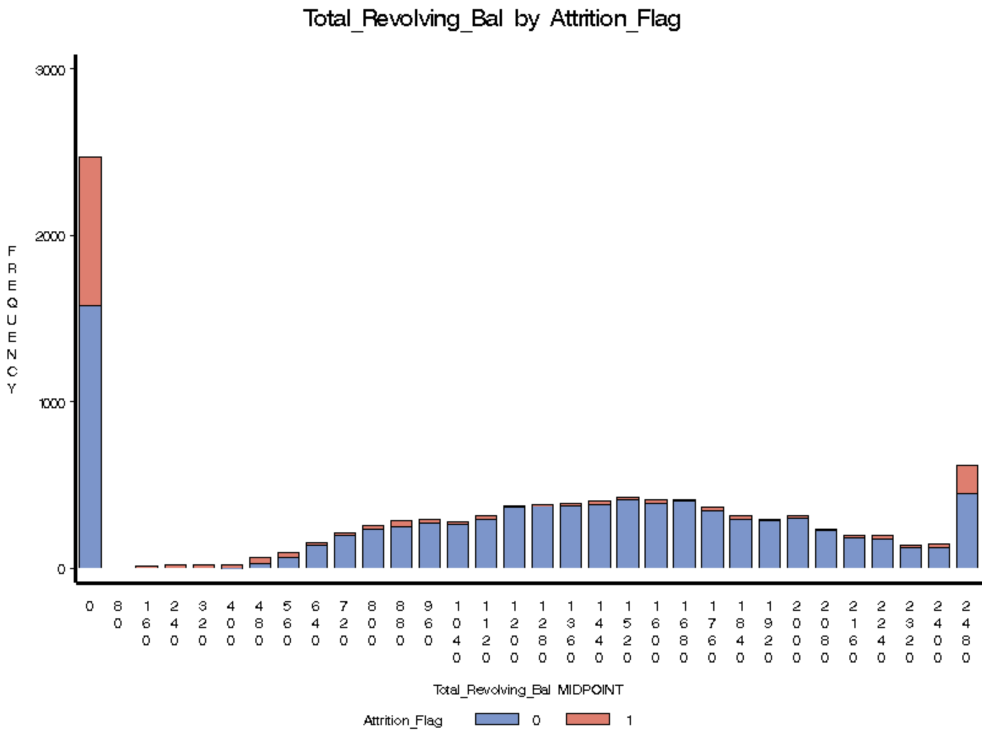
Months_Inactive_12_mon		▪ Discussion on the skewness of the distribution will not be relevant for the Months_Inactive_12_mon variable because it is a non-continuous variable. ▪ The two most frequent inactivity period is 2 and 3 months for both groups of customer attrition status. ▪ Very few customers showed 5 or 6 months of inactivity, but for those who do, churn rates are lower than non-churn rates. ▪ A sharp drop in total customer frequency is observed from 4 months of inactivity onwards for both attrition groups.

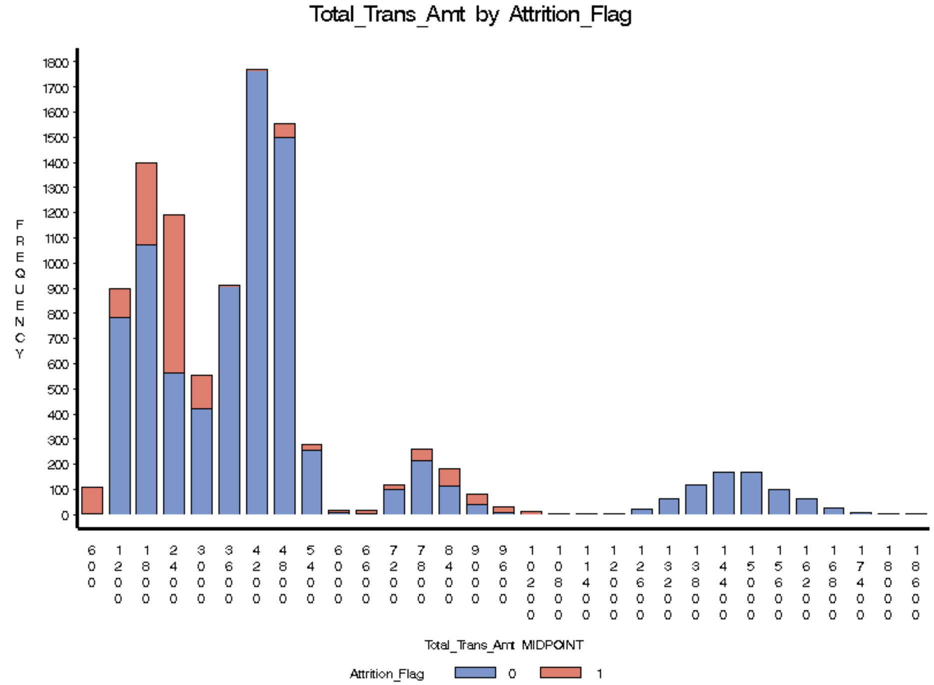
Months_on_book		▪ Discussion on the skewness of the distribution will not be relevant for the Months_on_book variable because it is a non-continuous variable. ▪ Across all of the midpoint bins of the Months_on_book variable, the number of customers who have not churned is constantly higher than those who have churned. ▪ Churn rates was the highest for customers who have maintained a 36-month long relationship with the bank.

Total_Amt_Chng_Q4_Q1		▪ The distribution is skewed to the right, with most customers being concentrated at the lower Total_Amt_Chng_Q4_Q1 range. ▪ Across all midpoint bins, the number of customers who have retained are considerably higher than their churned counterparts, implying that consistent customer engagement and transaction activity may play a significant role in customer loyalty and retention.

Total_Ct_Chng_Q4_Q1		▪ The distribution is skewed to the right, with most customers being concentrated at the lower Total_Ct_Chng_Q4_Q1 range. ▪ Across all midpoint bins, except for the lower range of Total_Ct_Chng_Q4_Q1, the number of customers who have retained are consistently higher than their churned counterparts, implying that higher engagement through transaction frequency is potentially linked to customer retention.

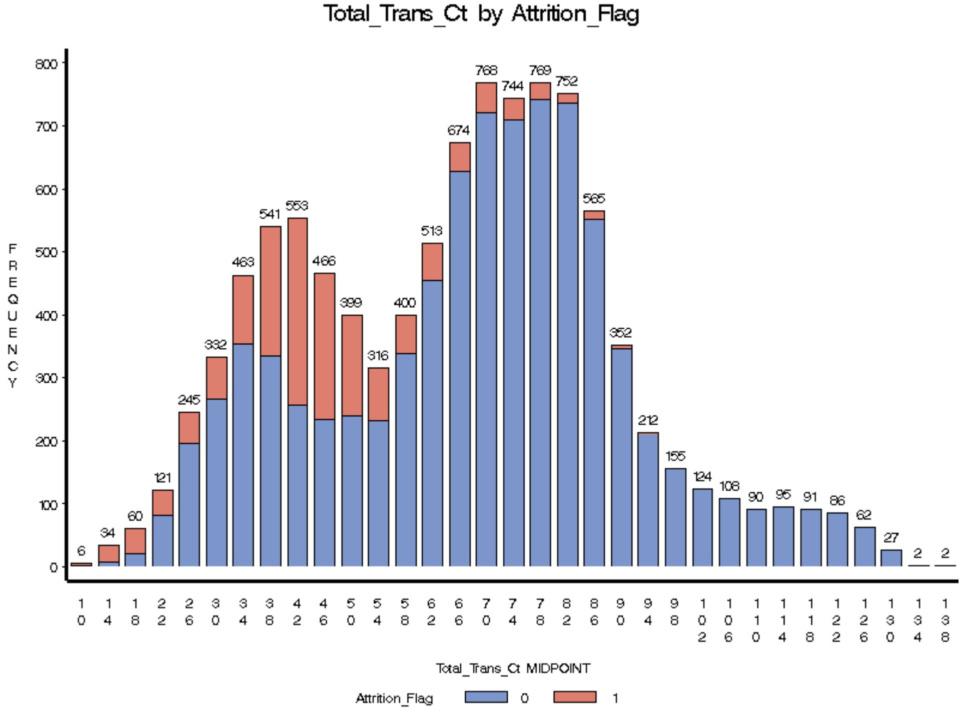
Total_Relationship_Count		▪ The distribution follows an almost normal distribution. ▪ Majority of the customers who have churned and not churned have three total relationships with the bank. ▪ The number of customers who have churned has constantly outnumbered those who have not churned, across all relationship count categories, suggesting that a greater number of relationships with the bank might correlate with lower attrition rates.

Total_Revolving_Bal		▪ Even though the distribution seems like it is skewed to the right, its skewness value of -0.14884 (*Section 2.3.1*) does however, fall within the acceptable threshold of -1 to 1. ▪ The number of customers who have churned has constantly outnumbered those who have not churned, across all midpoint bins of total revolving balance, implying that a greater amount of revolving balance with the bank might correlate with lower attrition rates. ▪ The highest frequency of churned customers occurred at the end of the distribution, suggesting that despite having substantial credit activity, these

<table>
<tr>
<td></td>
<td></td>
<td>customers are still at risk of leaving the bank.</td>
</tr>
<tr>
<td>Total_Trans_Amt</td>
<td>


</td>
<td>n is skewed to the right, with most ng concentrated at the lower mt range.
ith lower transaction amounts igher frequencies of both retention

uency of customer churn occurs at action amount midpoints, possibly ower spending is linked to customer</td>
</tr>
</table>

Total_Trans_Ct		▪ The distribution follows an almost normal distribution. ▪ The majority of the churned customers are found in the lower ▪ Customers with higher number of transaction counts tend to have lower attrition tendencies, as seen by the larger blue bars in the higher transaction count midpoints. ▪ There are considerably lesser customers with very low or very high number of transactions. ▪ Churn rates are considerably higher in the lower range of transaction count values as compared to higher ranges. ▪ The overall pattern suggests that an increased in transaction

		count may lead to an increased in customer retention.

In summary, consistent with the findings in the Variable Worth diagram, results from the Multiplot analysis demonstrates that variables namely, Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Avg_Utilization_Ratio, Total_Amt_Chng_Q4_Q1 and Contacts_Count_12_mon were found to have considerable predictive power on the outcomes of customer attrition.

2.4 Data Preparation

Considering that there were no missing values detected in the dataset, imputation is hence not needed. Instead, transformation of the variables with skewed distributions,, followed by feature selection, the undersampling of the dataset, and finally the portioning of the dataset, all of which will be performed in the sequence that they are listed.

Table 4: Preprocessing Steps

<table>
<tr><th>Type of Data Preparation</th><th>Output</th><th>Explanation</th></tr>
<tr><td>Transform Variables</td><td>
<table>
<tr><th>Source</th><th>Method</th><th>Variable Name</th><th>Skewness</th></tr>
<tr><td>Input</td><td>Original</td><td>Avg_Open_To_Buy</td><td>1.661697</td></tr>
<tr><td>Input</td><td>Original</td><td>Credit_Limit</td><td>1.666726</td></tr>
<tr><td>Input</td><td>Original</td><td>Total_Amt_Chng_Q4_Q1</td><td>1.732063</td></tr>
<tr><td>Input</td><td>Original</td><td>Total_Ct_Chng_Q4_Q1</td><td>2.064031</td></tr>
<tr><td>Input</td><td>Original</td><td>Total_Trans_Amt</td><td>2.041003</td></tr>
<tr><td>Output</td><td>Computed</td><td>LOG_Avg_Open_To_Buy</td><td>-0.0953</td></tr>
<tr><td>Output</td><td>Computed</td><td>LOG_Credit_Limit</td><td>0.457303</td></tr>
<tr><td>Output</td><td>Computed</td><td>LOG_Total_Amt_Chng_Q4_Q1</td><td>0.64844</td></tr>
<tr><td>Output</td><td>Computed</td><td>LOG_Total_Ct_Chng_Q4_Q1</td><td>0.510172</td></tr>
<tr><td>Output</td><td>Computed</td><td>LOG_Total_Trans_Amt</td><td>0.26278</td></tr>
</table>
</td><td>

Logarithmic transformation is applied to the 5 variables with skewed distribution which were previously identified in Section 2.3.1.
Outputs of the transformation procedure shows that skewness for all of the problematic variables has been reduced and are within the acceptable range of -1 to 1.

</td></tr>
<tr><td>Drop</td><td>
Dropped Variables Summary
<table>
<tr><th>ROLE</th><th>LEVEL</th><th>COUNT</th></tr>
<tr><td>HIDDEN</td><td></td><td>5</td></tr>
<tr><td>INPUT</td><td>INTERVAL</td><td>1</td></tr>
<tr><td>INPUT</td><td>NOMINAL</td><td>3</td></tr>
<tr><td>INPUT</td><td>ORDINAL</td><td>4</td></tr>
</table>
</td><td>

In this step, and according to the Variable Worth diagram in Section 2.3.1, variables which are found to have minimal contributions in predicting the attrition of the bank’s customers will be dropped from the dataset. This is to ensure that model performance could be optimized by removing any redundant or irrelevant data.
8 variables were removed, and they are namely, Customer_Age, Months_on_book, Gender, Income_Category, Education_Level,

</td></tr>
</table>

<table>
<tr>
<td></td>
<td></td>
<td>Dependent_count, Marital_Status, and Card_Category.</td>
</tr>
<tr>
<td>Sampling</td>
<td>
Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA
<table>
<tr><th>Variable</th><th>Numeric Value</th><th>Formatted Value</th><th>Frequency Count</th><th>Percent</th><th>Label</th></tr>
<tr><td>Attrition_Flag</td><td>0</td><td>0</td><td>8500</td><td>83.9340</td><td></td></tr>
<tr><td>Attrition_Flag</td><td>1</td><td>1</td><td>1627</td><td>16.0660</td><td></td></tr>
</table>
Data=SAMPLE
<table>
<tr><th>Variable</th><th>Numeric Value</th><th>Formatted Value</th><th>Frequency Count</th><th>Percent</th><th>Label</th></tr>
<tr><td>Attrition_Flag</td><td>0</td><td>0</td><td>1627</td><td>50</td><td></td></tr>
<tr><td>Attrition_Flag</td><td>1</td><td>1</td><td>1627</td><td>50</td><td></td></tr>
</table>
</td>
<td>

Undersampling technique has been applied to the dataset to solve the issue of class imbalance within the target variable.
Initially, number of observations was 8500 for retained customers and 1627 for churned customers. After the undersampling procedure, the number of observations for both groups are now equal, with 1627 observations in each group.

</td>
</tr>
<tr>
<td>Data Partitioning</td>
<td>
Data=TRAIN
<table>
<tr><th>Variable</th><th>Numeric Value</th><th>Formatted Value</th><th>Frequency Count</th><th>Percent</th><th>Label</th></tr>
<tr><td>Attrition_Flag</td><td>0</td><td>0</td><td>1138</td><td>49.9780</td><td></td></tr>
<tr><td>Attrition_Flag</td><td>1</td><td>1</td><td>1139</td><td>50.0220</td><td></td></tr>
</table>
Data=VALIDATE
<table>
<tr><th>Variable</th><th>Numeric Value</th><th>Formatted Value</th><th>Frequency Count</th><th>Percent</th><th>Label</th></tr>
<tr><td>Attrition_Flag</td><td>0</td><td>0</td><td>489</td><td>50.0512</td><td></td></tr>
<tr><td>Attrition_Flag</td><td>1</td><td>1</td><td>488</td><td>49.9488</td><td></td></tr>
</table>
</td>
<td>

Finally, the dataset is split into a 70-30 training-validation ratio.
The purpose of the split is to prepare the data for modelling, after which the validation dataset will be used to assess the model’s performance and to assess if overfitting is present.

</td>
</tr>
</table>

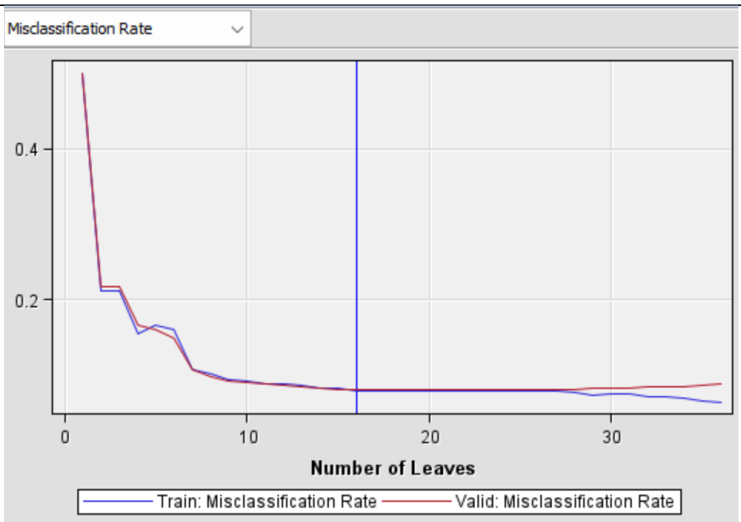
2.5 Predictive Modelling

The models included in this study are namely tree-based and neural network models. There will be a total of 5 variations done for each group of models, as could be seen in the table below. The tree-based models consisted of Decision Tree, two variations of Extreme Gradient Boosting, HP Tree, HP Forest. The neural network category however consisted of 1 conventional neural network model, and 4 other variations of HP neural network models. The following subsections will present the models mentioned, accompanied by their respective optimization properties and the validation outputs of the results.

2.5.1 Tree-Based Models

The 5 tree-based models are as follows.

Table 5: Tree-Based Modelling

<table>
<tr><th>Model</th><th>Optimization Properties</th><th>Validation Results</th></tr>
<tr>
<td>Decision Tree</td>
<td>Predefined Settings:
▪ Significance Level (0.2)
▪ Maximum Branch (2)
▪ Maximum Depth (6)
▪ Leaf Size (5)</td>
<td>


<table>
<tr><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr>
<tr><td>_NOBS_</td><td>Sum of Frequencies</td><td>2277</td><td>977</td></tr>
<tr><td>_MISC_</td><td>Misclassification Rate</td><td>0.07993</td><td>0.08086</td></tr>
<tr><td>_MAX_</td><td>Maximum Absolute Error</td><td>0.988115</td><td>0.988115</td></tr>
<tr><td>_SSE_</td><td>Sum of Squared Errors</td><td>311.1837</td><td>139.3546</td></tr>
<tr><td>_ASE_</td><td>Average Squared Error</td><td>0.068332</td><td>0.071318</td></tr>
<tr><td>_RASE_</td><td>Root Average Squared Error</td><td>0.261404</td><td>0.267054</td></tr>
<tr><td>_DIV_</td><td>Divisor for ASE</td><td>4554</td><td>1954</td></tr>
<tr><td>_DFT_</td><td>Total Degrees of Freedom</td><td>2277</td><td>.</td></tr>
</table>
Variable Importance
<table>
<tr><th>Variable Name</th><th>Label</th><th>Number of Splitting Rules</th><th>Importance</th><th>Validation Importance</th><th>Ratio of Validation to Training Importance</th></tr>
<tr><td>Total_Trans_Ct</td><td></td><td>4</td><td>1.0000</td><td>1.0000</td><td>1.0000</td></tr>
<tr><td>LOG_Total_Trans_Amt</td><td>Transformed Total_Trans_Amt</td><td>5</td><td>0.5426</td><td>0.5862</td><td>1.0804</td></tr>
<tr><td>Total_Revolving_Bal</td><td></td><td>1</td><td>0.3842</td><td>0.4360</td><td>1.1348</td></tr>
<tr><td>Total_Relationship_Count</td><td></td><td>1</td><td>0.2992</td><td>0.2761</td><td>0.9227</td></tr>
<tr><td>LOG_Total_Ct_Chng_Q4_Q1</td><td>Transformed Total_Ct_Chng_Q4_Q1</td><td>2</td><td>0.2366</td><td>0.2163</td><td>0.9143</td></tr>
<tr><td>Months_Inactive_12_mon</td><td></td><td>1</td><td>0.1460</td><td>0.0960</td><td>0.6577</td></tr>
<tr><td>Avg_Utilization_Ratio</td><td></td><td>1</td><td>0.1235</td><td>0.0826</td><td>0.6688</td></tr>
</table>
</td>
</tr>
</table>

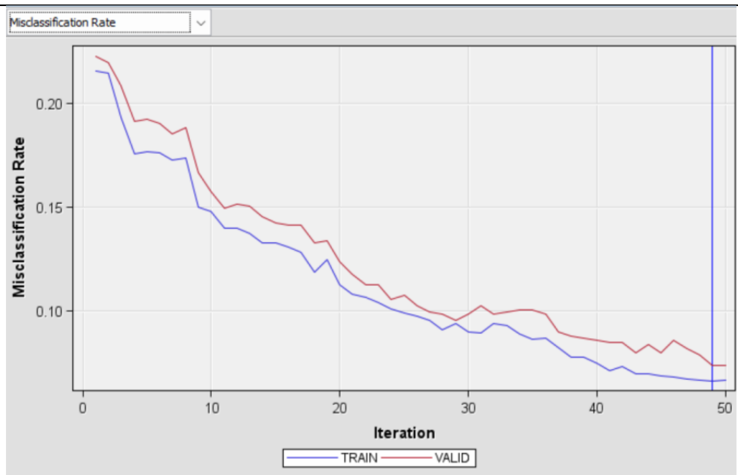
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	92.1811	91.6155	448	45.8547
1	0	7.8189	7.7869	38	3.8895
0	1	8.3503	8.3845	41	4.1965
1	1	91.6497	92.2131	450	46.0594

Gradient Boosting

Predefined Settings:

- N Iterations (5)
- Maximum Branch (2)
- Maximum Depth (2)
- Leaf Fraction (0.001)



Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2277	977
SUMW	Sum of Case Weights Times Freq	4554	1954
MISC	Misclassification Rate	0.066315	0.073695
MAX	Maximum Absolute Error	0.924925	0.91524
SSE	Sum of Squared Errors	287.894	134.2106
ASE	Average Squared Error	0.063218	0.068685
RASE	Root Average Squared Error	0.251432	0.262078
DIV	Divisor for ASE	4554	1954
DFT	Total Degrees of Freedom	2277	.

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	Total_Trans_Ct		40	1.00000	1.00000	1.00000
2	Total_Revolving_Bal		21	0.61483	0.60574	0.98522
3	LOG_Total_Trans_Amt	Transformed Total_Trans_Amt	41	0.59548	0.65794	1.10490
4	LOG_Total_Ct_Chng_Q4_Q1	Transformed Total_Ct_Chng_Q4_Q1	10	0.33987	0.32090	0.94420
5	Total_Relationship_Count		10	0.31029	0.23923	0.77098
6	LOG_Total_Amt_Chng_Q4_Q1	Transformed Total_Amt_Chng_Q4_Q1	9	0.22358	0.21173	0.94698
7	Months_Inactive_12_mon		7	0.21686	0.20258	0.93417
8	Contacts_Count_12_mon		7	0.14657	0.08032	0.54798
9	LOG_Credit_Limit	Transformed Credit_Limit	1	0.03450	0.00000	0.00000
10	Avg_Utilization_Ratio		1	0.02126	0.00000	0.00000

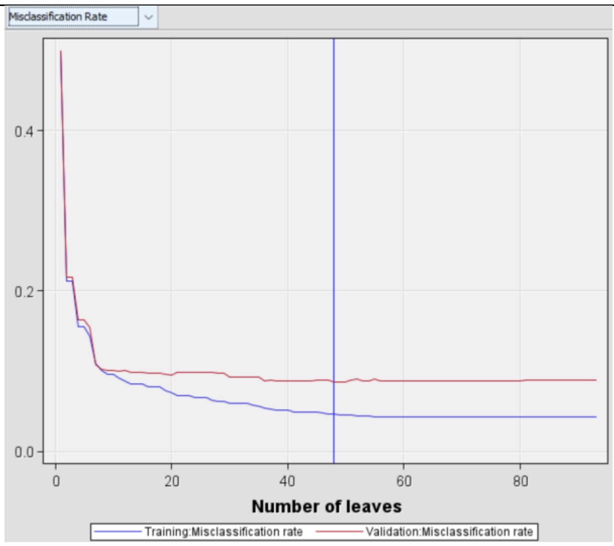
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	93.3472	91.8200	449	45.9570
1	0	6.6528	6.5574	32	3.2753
0	1	8.0645	8.1800	40	4.0942
1	1	91.9355	93.4426	456	46.6735

HP Tree

Predefined Settings:

- Significance Level (0.2)
- Maximum Branch (2)
- Maximum Depth (10)
- Leaf Size (5)



Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.038345	0.075054
DIV	Divisor for ASE	4554	1954
MAX	Maximum Absolute Error	0.994413	1
NOBS	Sum of Frequencies	2277	977
RASE	Root Average Squared Error	0.195819	0.27396
SSE	Sum of Squared Errors	174.6229	146.6561
DISF	Frequency of Classified Cases	2277	977
MISC	Misclassification Rate	0.046113	0.087001
WRONG	Number of Wrong Classifications	105	85

Variable Name	Importance	Validation Importance ▲
LOG_Avg_Open_To_Buy	0.104126	0
LOG_Credit_Limit	0	0
Avg_Utilization_Ratio	0.134643	0.03852
Contacts_Count_12_mon	0.045616	0.046165
LOG_Total_Amt_Chng_Q4_Q1	0.210449	0.103309
Months_Inactive_12_mon	0.224374	0.131757
Total_Relationship_Count	0.324341	0.275342
LOG_Total_Ct_Chng_Q4_Q1	0.30413	0.279913
Total_Revolving_Bal	0.408928	0.429212
LOG_Total_Trans_Amt	0.562102	0.559685
Total_Trans_Ct	1	1

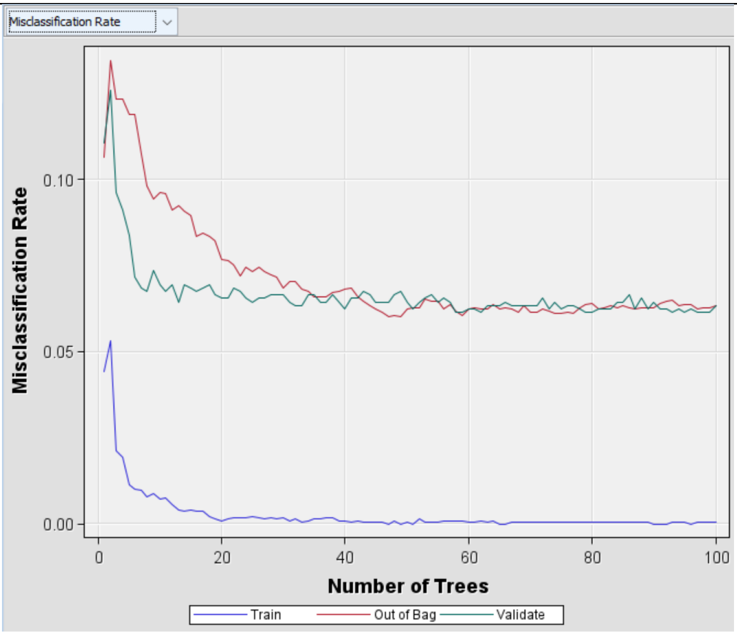
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	90.8907	91.8200	449	45.9570
1	0	9.1093	9.2213	45	4.6059
0	1	8.2816	8.1800	40	4.0942
1	1	91.7184	90.7787	443	45.3429

HP Forest

Predefined Settings:

- Maximum Number of Trees (100)
- Maximum Depth (50)
- Significance Level (0.05)
- Leaf Size (5)
- Variable Importance Method: Loss Reduction



Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.01021	0.051287
DIV	Divisor for ASE	4554	1954
MAX	Maximum Absolute Error	0.512698	0.923667
NOBS	Sum of Frequencies	2277	977
RASE	Root Average Squared Error	0.101045	0.226467
SSE	Sum of Squared Errors	46.49694	100.2151
DISF	Frequency of Classified Cases	2277	977
MISC	Misclassification Rate	.0004392	0.06346
WRONG	Number of Wrong Classifications	1	62

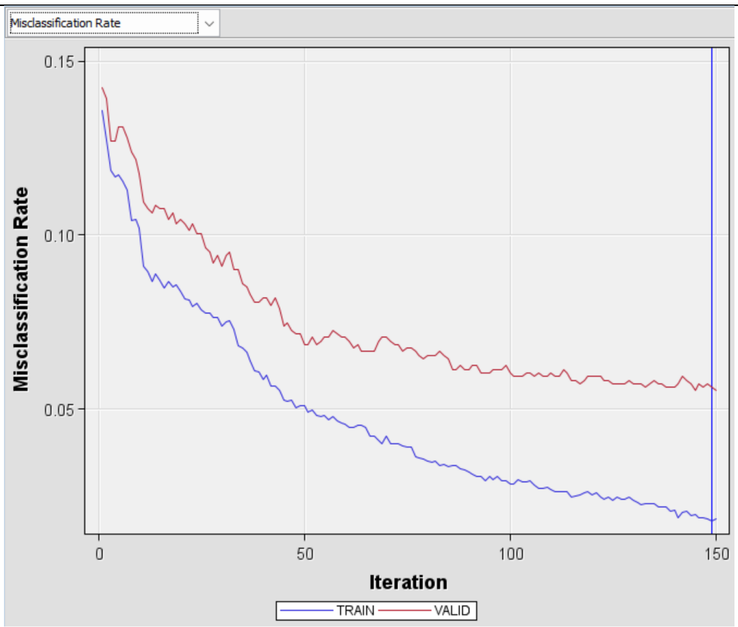
Loss Reduction Variable Importance

Variable	Number of Rules	Gini	OOB Gini	Valid Gini	Margin	OOB Margin	Valid Margin
Total_Trans_Ct	2914	0.147575	0.10076	0.10163	0.295151	0.250128	0.251284
LOG_Total_Trans_Amt	2238	0.094256	0.06133	0.07059	0.188511	0.154489	0.161218
Total_Revolving_Bal	1754	0.062450	0.03356	0.03215	0.124899	0.095271	0.094082
Avg_Utilization_Ratio	1017	0.037653	0.02048	0.01990	0.075305	0.058410	0.057702
Total_Relationship_Count	612	0.019967	0.01402	0.00920	0.039934	0.034288	0.028619
LOG_Total_Ct_Chng_Q4_Q1	2021	0.049367	0.01309	0.01408	0.098733	0.063191	0.065766
Months_Inactive_12_mon	382	0.007939	0.00490	0.00512	0.015877	0.012531	0.013107
Contacts_Count_12_mon	401	0.007773	0.00273	0.00152	0.015545	0.010073	0.009241
LOG_Total_Amt_Chng_Q4_Q1	1626	0.030996	0.00207	-0.00001	0.061993	0.032615	0.033055
LOG_Avg_Open_To_Buy	1446	0.016084	-0.00896	-0.00758	0.032168	0.006903	0.007737
LOG_Credit_Limit	1695	0.018085	-0.01098	-0.01270	0.036169	0.006745	0.004296

Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	94.3867	92.8425	454	46.4688
1	0	5.6133	5.5328	27	2.7636
0	1	7.0565	7.1575	35	3.5824
1	1	92.9435	94.4672	461	47.1853

<table>
<tr>
<td>Gradient Boosting (2)</td>
<td>Modified Settings:
▪ N Iterations (150)
▪ Maximum Branch (4)
▪ Maximum Depth (2)
▪ Leaf Fraction (0.001)</td>
<td>



<table>
<tr><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr>
<tr><td>_NOBS_</td><td>Sum of Frequencies</td><td>2277</td><td>977</td></tr>
<tr><td>_SUMW_</td><td>Sum of Case Weights Times Freq</td><td>4554</td><td>1954</td></tr>
<tr><td>_MISC_</td><td>Misclassification Rate</td><td>0.019763</td><td>0.055271</td></tr>
<tr><td>_MAX_</td><td>Maximum Absolute Error</td><td>0.939209</td><td>0.957116</td></tr>
<tr><td>_SSE_</td><td>Sum of Squared Errors</td><td>100.0706</td><td>82.31928</td></tr>
<tr><td>_ASE_</td><td>Average Squared Error</td><td>0.021974</td><td>0.042129</td></tr>
<tr><td>_RASE_</td><td>Root Average Squared Error</td><td>0.148237</td><td>0.205253</td></tr>
<tr><td>_DIV_</td><td>Divisor for ASE</td><td>4554</td><td>1954</td></tr>
<tr><td>_DFT_</td><td>Total Degrees of Freedom</td><td>2277</td><td>.</td></tr>
</table>

<table>
<tr><th>Variable Name</th><th>Importance</th><th>Validation Importance</th><th>Ratio of Validation to Training Importance</th></tr>
<tr><td>LOG_Total_Trans_Amt</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>Total_Trans_Ct</td><td>0.959439</td><td>0.961392</td><td>1.002036</td></tr>
<tr><td>Total_Revolving_Bal</td><td>0.58822</td><td>0.545959</td><td>0.928153</td></tr>
<tr><td>LOG_Total_Ct_Chng_Q4_Q1</td><td>0.384048</td><td>0.341897</td><td>0.890244</td></tr>
<tr><td>LOG_Total_Amt_Chng_Q4_Q1</td><td>0.351716</td><td>0.297993</td><td>0.847254</td></tr>
<tr><td>Total_Relationship_Count</td><td>0.256226</td><td>0.201467</td><td>0.786289</td></tr>
<tr><td>Months_Inactive_12_mon</td><td>0.191283</td><td>0.184028</td><td>0.96207</td></tr>
<tr><td>Contacts_Count_12_mon</td><td>0.14941</td><td>0.065204</td><td>0.436409</td></tr>
<tr><td>LOG_Avg_Open_To_Buy</td><td>0.123477</td><td>0.037655</td><td>0.304958</td></tr>
<tr><td>Avg_Utilization_Ratio</td><td>0.1032</td><td>0.020304</td><td>0.19674</td></tr>
<tr><td>LOG_Credit_Limit</td><td>0.086701</td><td>0</td><td>0</td></tr>
</table>

Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

<table>
<tr><th>Target</th><th>Outcome</th><th>Target Percentage</th><th>Outcome Percentage</th><th>Frequency Count</th><th>Total Percentage</th></tr>
<tr><td>0</td><td>0</td><td>94.4785</td><td>94.4785</td><td>462</td><td>47.2876</td></tr>
<tr><td>1</td><td>0</td><td>5.5215</td><td>5.5328</td><td>27</td><td>2.7636</td></tr>
<tr><td>0</td><td>1</td><td>5.5328</td><td>5.5215</td><td>27</td><td>2.7636</td></tr>
<tr><td>1</td><td>1</td><td>94.4672</td><td>94.4672</td><td>461</td><td>47.1853</td></tr>
</table>

</td>
</tr>
</table>

Summary Comparisons of Tree-Based Models

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Boost2	Gradient Boosting (2)	0.055271	0.021974	0.019763	0.042129
	HPDMForest	HP Forest	0.063460	0.010210	0.000439	0.051287
	Boost	Gradient Boosting	0.073695	0.063218	0.066315	0.068685
	Tree	Decision Tree	0.080860	0.068332	0.079930	0.071318
	HPTree	HP Tree	0.087001	0.038345	0.046113	0.075054

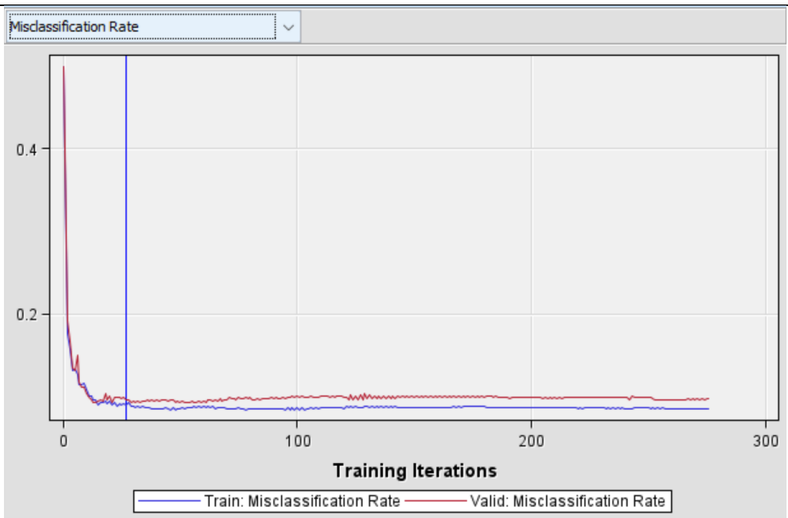
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Boost2	Gradient Boosting (2)	TRAIN	Attrition_Flag		8	1101	37	1131
Boost2	Gradient Boosting (2)	VALIDATE	Attrition_Flag		27	462	27	461
Tree	Decision Tree	TRAIN	Attrition_Flag		75	1031	107	1064
Tree	Decision Tree	VALIDATE	Attrition_Flag		38	448	41	450
Boost	Gradient Boosting	TRAIN	Attrition_Flag		53	1040	98	1086
Boost	Gradient Boosting	VALIDATE	Attrition_Flag		32	449	40	456
HPTree	HP Tree	TRAIN	Attrition_Flag		51	1084	54	1088
HPTree	HP Tree	VALIDATE	Attrition_Flag		45	449	40	443
HPDMForest	HP Forest	TRAIN	Attrition_Flag		.	1137	1	1139
HPDMForest	HP Forest	VALIDATE	Attrition_Flag		27	454	35	461

2.5.2 Neural Networks Models

The 5 neural networks models are as follows.

Table 6: Neural Network Modelling

<table>
<tr><th>Model</th><th>Optimization Properties</th><th>Validation Results</th></tr>
<tr><td>Neural Network</td><td>Predefined Settings:
▪ Method Selection Criterion (Profit/Loss)</td><td>


<table>
<tr><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr>
<tr><td>_AIC_</td><td>Akaike's Information Criterion</td><td>1212.214</td><td>.</td></tr>
<tr><td>_AVERR_</td><td>Average Error Function</td><td>0.230174</td><td>0.248434</td></tr>
<tr><td>_ASE_</td><td>Average Squared Error</td><td>0.068673</td><td>0.073587</td></tr>
<tr><td>_DFE_</td><td>Degrees of Freedom for Error</td><td>2195</td><td>.</td></tr>
<tr><td>_DIV_</td><td>Divisor for ASE</td><td>4554</td><td>1954</td></tr>
<tr><td>_ERR_</td><td>Error Function</td><td>1048.214</td><td>485.4394</td></tr>
<tr><td>_FPE_</td><td>Final Prediction Error</td><td>0.073804</td><td>.</td></tr>
<tr><td>_MAX_</td><td>Maximum Absolute Error</td><td>0.997498</td><td>0.996357</td></tr>
<tr><td>_MSE_</td><td>Mean Squared Error</td><td>0.071238</td><td>0.073587</td></tr>
<tr><td>_MISC_</td><td>Misclassification Rate</td><td>0.089592</td><td>0.095189</td></tr>
<tr><td>_DFM_</td><td>Model Degrees of Freedom</td><td>82</td><td>.</td></tr>
<tr><td>_NW_</td><td>Number of Estimated Weights</td><td>82</td><td>.</td></tr>
<tr><td>_WRONG_</td><td>Number of Wrong Classifications</td><td>204</td><td>93</td></tr>
<tr><td>_RASE_</td><td>Root Average Squared Error</td><td>0.262055</td><td>0.271268</td></tr>
<tr><td>_RFPE_</td><td>Root Final Prediction Error</td><td>0.271668</td><td>.</td></tr>
<tr><td>_RMSE_</td><td>Root Mean Squared Error</td><td>0.266905</td><td>0.271268</td></tr>
<tr><td>_SBC_</td><td>Schwarz's Bayesian Criterion</td><td>1682.124</td><td>.</td></tr>
<tr><td>_SUMW_</td><td>Sum of Case Weights Times Freq</td><td>4554</td><td>1954</td></tr>
<tr><td>_NOBS_</td><td>Sum of Frequencies</td><td>2277</td><td>977</td></tr>
<tr><td>_SSE_</td><td>Sum of Squared Errors</td><td>312.736</td><td>143.7881</td></tr>
<tr><td>_DFT_</td><td>Total Degrees of Freedom</td><td>2277</td><td>.</td></tr>
</table>

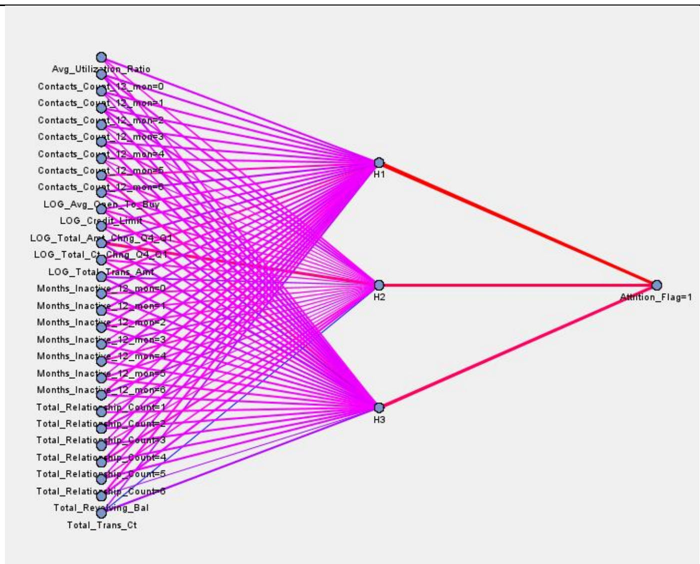
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

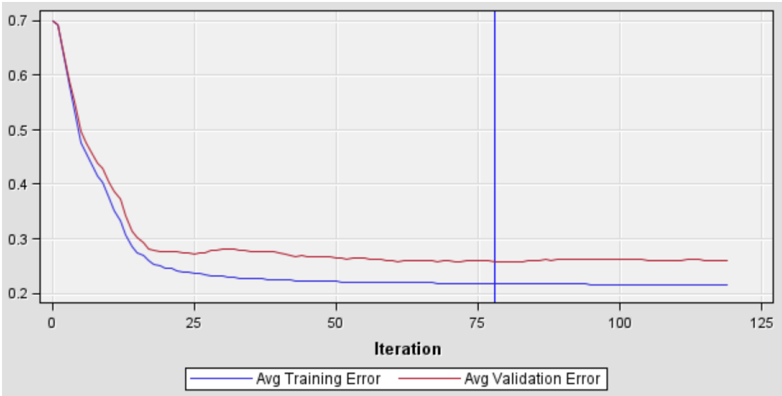
<table>
<tr><th>Target</th><th>Outcome</th><th>Target Percentage</th><th>Outcome Percentage</th><th>Frequency Count</th><th>Total Percentage</th></tr>
<tr><td>0</td><td>0</td><td>90.2439</td><td>90.7975</td><td>444</td><td>45.4452</td></tr>
<tr><td>1</td><td>0</td><td>9.7561</td><td>9.8361</td><td>48</td><td>4.9130</td></tr>
<tr><td>0</td><td>1</td><td>9.2784</td><td>9.2025</td><td>45</td><td>4.6059</td></tr>
<tr><td>1</td><td>1</td><td>90.7216</td><td>90.1639</td><td>440</td><td>45.0358</td></tr>
</table>
</td></tr>
</table>

HP Neural

Predefined Settings:

- Architecture (One Layer)
- Number of Hidden Neurons (3)
- Number of Hidden Layers (3)
- Target Activation Function (Identity)
- Number of Tries (2)
- Maximum Iterations (300)





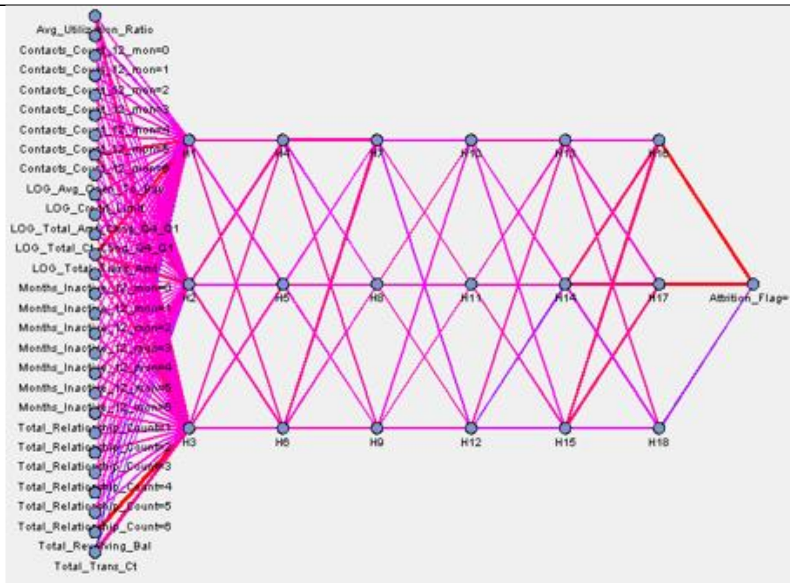
Fit Statistics	Statistics Label	Train	Validation
SSE	Sum of Squared Errors	298.2675	151.8091
NOBS	Sum of Frequencies	2277	977
RASE	Root Average Squared Error	0.255921	0.278732
WRONG	Number of Wrong Classifications	197	99
MISC	Misclassification Rate	0.086517	0.101331
MAX	Maximum Absolute Error	0.994613	0.993345
DISF	Frequency of Classified Cases	2277	977
DIV	Divisor for ASE	4554	1954
ASE	Average Squared Error	0.065496	0.077691

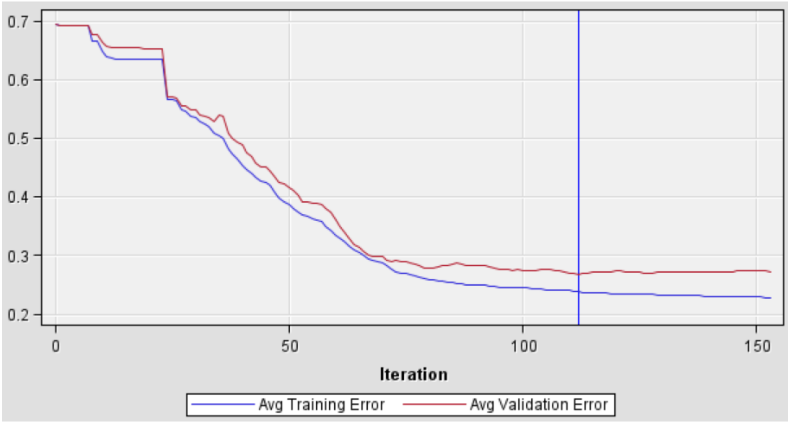
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	88.6905	91.4110	447	45.7523
1	0	11.3095	11.6803	57	5.8342
0	1	8.8795	8.5890	42	4.2989
1	1	91.1205	88.3197	431	44.1146

<table>
<tr>
<td>HP Neural (2)</td>
<td>Modified Settings:

Architecture (User-Defined)
Number of Hidden Neurons (3)
Number of Hidden Layers (6)
Target Activation Function (Identity)
Number of Tries (4)
Maximum Iterations (1000)

</td>
<td>




<table>
<tr><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr>
<tr><td>_ASE_</td><td>Average Squared Error</td><td>0.07341</td><td>0.080346</td></tr>
<tr><td>_DIV_</td><td>Divisor for ASE</td><td>4554</td><td>1954</td></tr>
<tr><td>_MAX_</td><td>Maximum Absolute Error</td><td>0.997182</td><td>0.992698</td></tr>
<tr><td>_NOBS_</td><td>Sum of Frequencies</td><td>2277</td><td>977</td></tr>
<tr><td>_RASE_</td><td>Root Average Squared Error</td><td>0.270943</td><td>0.283455</td></tr>
<tr><td>_SSE_</td><td>Sum of Squared Errors</td><td>334.3101</td><td>156.997</td></tr>
<tr><td>_DISF_</td><td>Frequency of Classified Cases</td><td>2277</td><td>977</td></tr>
<tr><td>_MISC_</td><td>Misclassification Rate</td><td>0.103206</td><td>0.111566</td></tr>
<tr><td>_WRONG_</td><td>Number of Wrong Classifications</td><td>235</td><td>109</td></tr>
</table>

Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

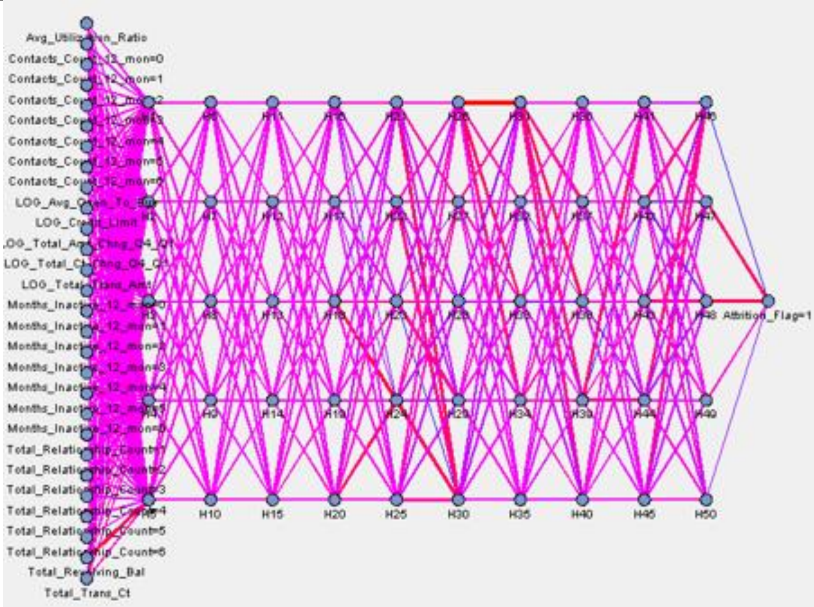
<table>
<tr><th>Target</th><th>Outcome</th><th>Target Percentage</th><th>Outcome Percentage</th><th>Frequency Count</th><th>Total Percentage</th></tr>
<tr><td>0</td><td>0</td><td>88.0000</td><td>89.9796</td><td>440</td><td>45.0358</td></tr>
<tr><td>1</td><td>0</td><td>12.0000</td><td>12.2951</td><td>60</td><td>6.1412</td></tr>
<tr><td>0</td><td>1</td><td>10.2725</td><td>10.0204</td><td>49</td><td>5.0154</td></tr>
<tr><td>1</td><td>1</td><td>89.7275</td><td>87.7049</td><td>428</td><td>43.8076</td></tr>
</table>
</td>
</tr>
</table>

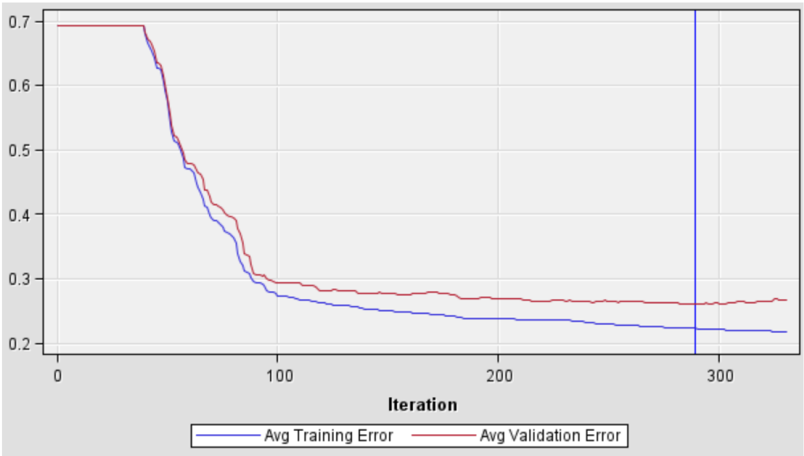
HP Neural (3)	Modified Settings:	

HP Neural (3)

Modified Settings:

- Architecture (User-Defined)
- Number of Hidden Neurons (3)
- Number of Hidden Layers (10)
- Target Activation Function (Identity)
- Target Activation Function (Identity)
- Number of Tries (15)
- Maximum Iterations (1000)



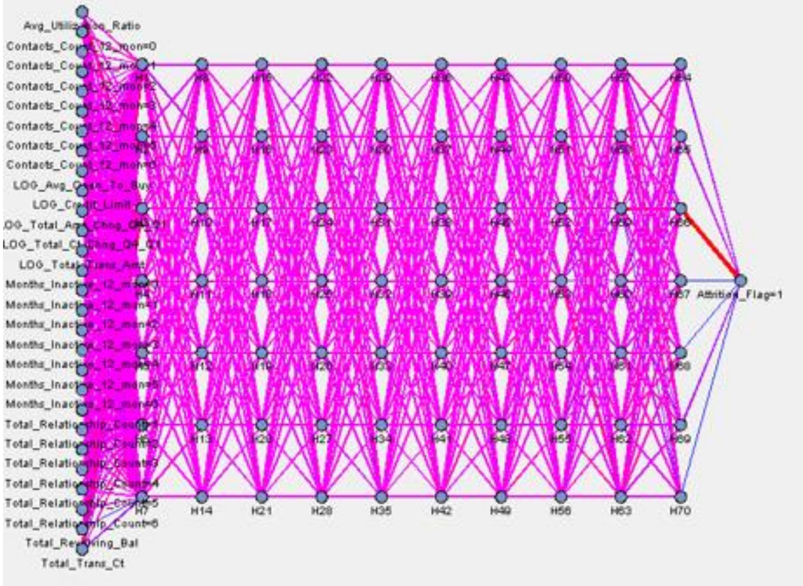


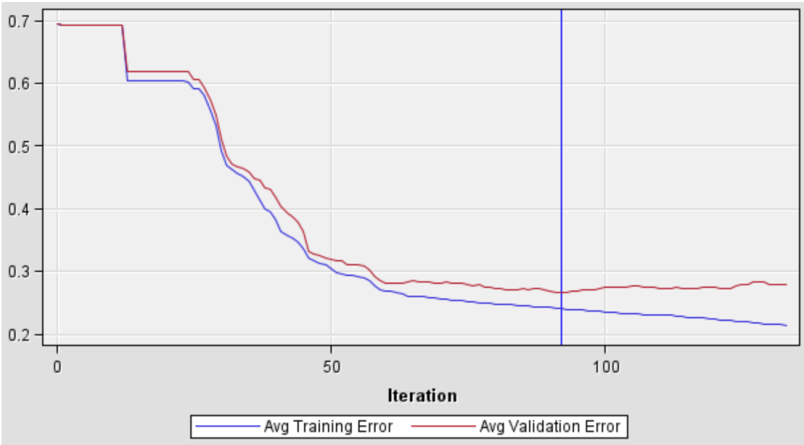
Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.068382	0.078503
DIV	Divisor for ASE	4554	1954
MAX	Maximum Absolute Error	0.994679	0.992857
NOBS	Sum of Frequencies	2277	977
RASE	Root Average Squared Error	0.2615	0.280184
SSE	Sum of Squared Errors	311.4126	153.3952
DISF	Frequency of Classified Cases	2277	977
MISC	Misclassification Rate	0.095301	0.107472
WRONG	Number of Wrong Classifications	217	105

```
Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

                          Target      Outcome     Frequency       Total
Target     Outcome      Percentage   Percentage     Count       Percentage

  0           0           87.9447     91.0020        445          45.5476
  1           0           12.0553     12.5000         61           6.2436
  0           1            9.3418      8.9980         44           4.5036
  1           1           90.6582     87.5000        427          43.7052
```

<table>
<tr>
<td>HP
Neural
(4)</td>
<td>Modified
Settings:

Architecture (User-Defined)
Number of Hidden Neurons (3)
Number of Hidden Layers (10)
Target Activation Function (Identity)
Target Activation Function (Identity)
Number of Tries (20)
Maximum Iterations (1000)

</td>
<td>




Data Role=VALIDATE Target Variable=Attrition_Flag Target Label=' '

<table>
<tr><th>Target</th><th>Outcome</th><th>Target Percentage</th><th>Outcome Percentage</th><th>Frequency Count</th><th>Total Percentage</th></tr>
<tr><td>0</td><td>0</td><td>88.9571</td><td>88.9571</td><td>435</td><td>44.5241</td></tr>
<tr><td>1</td><td>0</td><td>11.0429</td><td>11.0656</td><td>54</td><td>5.5271</td></tr>
<tr><td>0</td><td>1</td><td>11.0656</td><td>11.0429</td><td>54</td><td>5.5271</td></tr>
<tr><td>1</td><td>1</td><td>88.9344</td><td>88.9344</td><td>434</td><td>44.4217</td></tr>
</table>

<table>
<tr><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr>
<tr><td>_ASE_</td><td>Average Squared Error</td><td>0.073832</td><td>0.079541</td></tr>
<tr><td>_DIV_</td><td>Divisor for ASE</td><td>4554</td><td>1954</td></tr>
<tr><td>_MAX_</td><td>Maximum Absolute Error</td><td>0.991966</td><td>0.990863</td></tr>
<tr><td>_NOBS_</td><td>Sum of Frequencies</td><td>2277</td><td>977</td></tr>
<tr><td>_RASE_</td><td>Root Average Squared Error</td><td>0.271721</td><td>0.28203</td></tr>
<tr><td>_SSE_</td><td>Sum of Squared Errors</td><td>336.2329</td><td>155.4229</td></tr>
<tr><td>_DISF_</td><td>Frequency of Classified Cases</td><td>2277</td><td>977</td></tr>
<tr><td>_MISC_</td><td>Misclassification Rate</td><td>0.104523</td><td>0.110542</td></tr>
<tr><td>_WRONG_</td><td>Number of Wrong Classifications</td><td>238</td><td>108</td></tr>
</table>
</td>
</tr>
</table>

Summary Comparisons of Neural Network Models

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Neural	Neural Network	0.09519	0.068673	0.08959	0.073587
	HPNNA	HP Neural	0.10133	0.065496	0.08652	0.077691
	HPNNA3	HP Neural (3)	0.10747	0.068382	0.09530	0.078503
	HPNNA4	HP Neural (4)	0.11054	0.073832	0.10452	0.079541
	HPNNA2	HP Neural (2)	0.11157	0.073410	0.10321	0.080346

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
HPNNA2	HP Neural (2)	TRAIN	Attrition_Flag		109	1012	126	1030
HPNNA2	HP Neural (2)	VALIDATE	Attrition_Flag		60	440	49	428
HPNNA	HP Neural	TRAIN	Attrition_Flag		97	1038	100	1042
HPNNA	HP Neural	VALIDATE	Attrition_Flag		57	447	42	431
Neural	Neural Network	TRAIN	Attrition_Flag		91	1025	113	1048
Neural	Neural Network	VALIDATE	Attrition_Flag		48	444	45	440
HPNNA3	HP Neural (3)	TRAIN	Attrition_Flag		113	1034	104	1026
HPNNA3	HP Neural (3)	VALIDATE	Attrition_Flag		61	445	44	427
HPNNA4	HP Neural (4)	TRAIN	Attrition_Flag		122	1022	116	1017
HPNNA4	HP Neural (4)	VALIDATE	Attrition_Flag		54	435	54	434

2.6 Model Interpretation

Before delving into the evaluation and interpretation of the models created for each of the model group separately, it is important to understand the meaning of the performance metrics in the context of this study, which is the prediction of customer churn for the credit card services domain. In the said domain, the Precision metric measures the accuracy of correctly predicted churn cases against all predicted churns. Recall or also known as sensitivity on the other hand, examines the model's ability to capture all actual churns, underscoring the model's efficacy in identifying churn cases. The metric of specificity emphasizes on accurately capturing non-churned customers. Lastly, the F1 score acts to mediate the precision and recall metrics, thereby making it an important metric for when there is an imbalanced class, just like the case of this study.

Given the above, the more severe error in the case of this study would be the false negatives or Type II error where the model has failed to predict churn. This failure is critical because it represents a forgone opportunity for engagement to retain the customers, possibly resulting in the explicit loss of revenue and the continuing values that the customer could have contributed to the bank. That said, while false positive is undesirable as well, it is the lesser evil because it involves wrongfully predicting that customers will churn when in fact they would not, essentially leading to wasted resources on retention efforts. While such efforts might be costly to the bank, it is however less severe as it does not undermine customer relationships like how a false negative would. In essence, the goal is to select a model that minimizes false negatives to boost customer retention and to secure the bank's revenue. Also, referring back to the *Section 1.1*, it should be noted that research indicates that acquiring new customers in the case of the banking sector is significantly much costlier than retaining existing customers, once again justifying for the severity of false negatives in this study.

2.6.1 Interpretation and Recommendations (Tree-Based Model)

From the Fit Statistics Table under the 'Summary Comparisons of Tree-Based Models' part in *Section 2.5.1*, it is evident that misclassification rate was the lowest in the Boost2 model. While there is some disparity between the misclassification rates for the training (0.019763) and validation (0.055271) dataset, the difference (roughly 3.55%) is within the acceptable threshold of 5%. Hence, it can be concluded that the Boost2 model is likely generalizing well with unseen data, and that there is no issue of overfitting present. Despite being compared to HP Tree and HP Forest which are known to be more powerful than any conventional tree-based models, the

Gradient Boosting model (Boost2) with minimal modifications (N Iterations (150), and Maximum Branch (4)) was able to outperform its high-performance counterparts. The Subseries Plot for the Boost2 model demonstrates that misclassification rate achieved its lowest at 149th iteration. In the ROC curve diagram shown below (*Figure 3*), it is apparent that the Gradient Boosting (2) (Boost2) model, as depicted by the brown line, is closest to the top left corner of the square for the validation dataset. The AUC for the Boost2 model is also the largest among all the other tree-based models.

From the Event Classification Table under the 'Summary Comparisons of Tree-Based Models' part in *Section 2.5.1* and remembering that the objective is to minimize false negatives, the Boost2 model in the validation set is the model that has the lowest number of false negatives (27), implying that the model is the best at identifying true churn cases among the other competing models. In other words, the Boost2 model was found to have the highest Recall value among the rest. At the same time, the Boost2 model has also shown an equal number of false positive (27), suggesting that a good balance is maintained between sensitivity and precision. In terms of the variable importance, which is one of the objectives of this study, insights can be drawn form the Variable Importance statistics generated in *Section 2.5.1* of the Gradient Boosting (2) model. The top 5 factors with high variable worth are namely, LOG_Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, LOG_Total_Ct_Chng_Q4_Q1, and LOG_Total_Amt_Chng_Q4_Q1. As these factors are ranked highest in both Training Importance and Validation Importance, it is proposed that these factors are indeed indicative of the churn of credit card customers in the banking sector.

Given the identified factors, several recommendations are formulated to mitigate customer attrition and they are as follows:

1. The bank could enhance customer engagement by introducing personalized offerings and rewards that are based on the customer's transaction amounts and counts to incentivize them to further increase their transaction amount and the number of transactions that they will make, ultimately reducing their potential of churning.

2. The bank could equip customers with better tools such as frequent balance alerts in order to help them better manage their total revolving balance. At the same time, banks could incentivize their customers with lower interest rates on their revolving balances with the aim of fostering responsible credit usage, all of which could help to reduce customer attrition for the banks.

3. Banks could also keep a close eye on any drastic changes in their customers' transaction behaviours as denoted by LOG_Total_Ct_Chng_Q4_Q1 and LOG_Total_Amt_Chgn_Q4_Q1 so that proactive measures and support could be introduced promptly as and when drastic changes are detected, once again minimizing the risk of customer churns.

4. Besides that, banks could also strengthen customer relationships by providing benefits for longer tenure or more frequent interactions, as indicated by Total_Relationship_Count and Months_Inactive_12_mon, to encourage continued business with the bank.

5. Lastly, for accounts with lower card utilization ratio, banks could introduce programs that encourage the use of available credit without promoting unmanageable debt, yet at the same time, improving customer loyalty and perceived value of the banks' services.

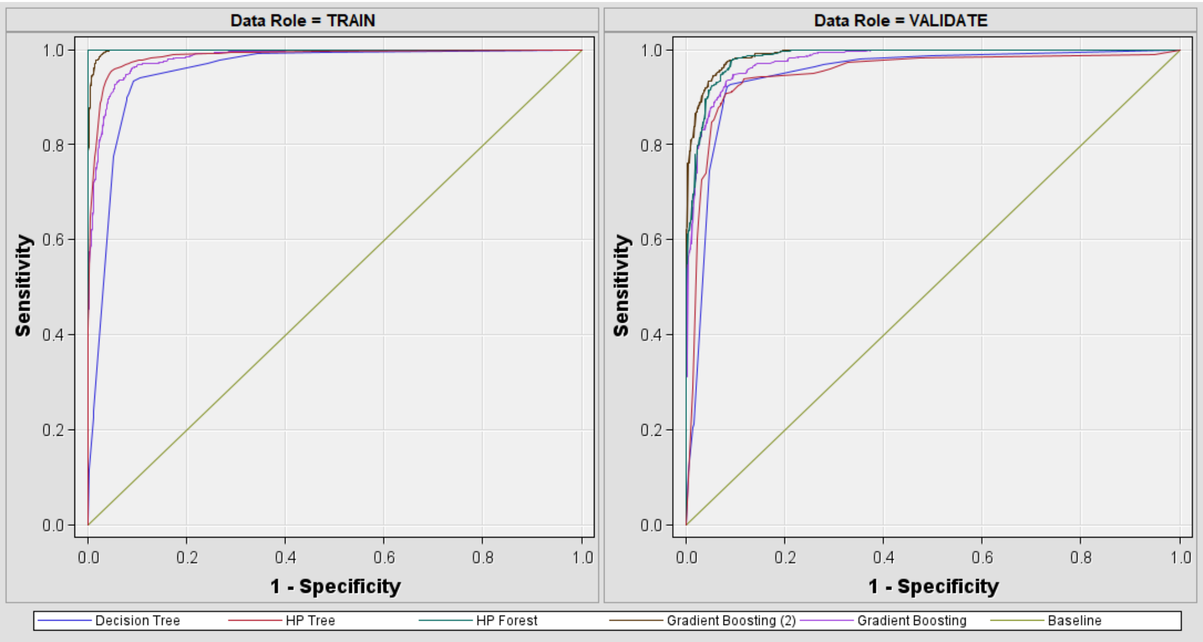


Figure 3: ROC Curve of Tree-Based Models

2.6.2 Interpretation and Recommendations (Neural Network Model)

When making comparisons between the neural network models created and based on the Fit Statistics Table under the under the ‘Summary Comparisons of Neural Network Models’ in *Section 2.5.2,* findings are that the Neural Network model has had the lowest misclassification rate (0.09519) among the rest of the models, for its validation dataset. Differences between the model’s misclassification rates for the validation (0.09519) and training dataset (0.08959) were also not significantly different from one another, hence implying that the model does generalize well with unseen data and that there is no issue of overfitting. This finding shed light on the adequacy of a conventional neural network in managing the given classification task effectively, demonstrating that, for this particular dataset, the introduction of more complex hyperparameter-optimized neural networks does not translate into substantial performance gains. Referring to the Event Classification Table under the similar sections previously mentioned, conclusions on the best model are found to be consistent with the above findings. While the above evaluation looks at the misclassification rates, the Event Classification Table calls for focus to be placed on the number of false negatives contained within the model. At the same time, consideration has to be given to the number of false positives as well to ensure that a balance is stroked as targeting too many customers senselessly with retention efforts would mean wasted resources, and hence is undesirable.

Following the expectations set on the false positives and negatives of the analysis, comparisons between the neural network models created indicates that the “Neural Network” model is indeed the best model as it has the lowest number of false negatives (48) and an almost equal number of false positive (45) in its validation dataset. The worst model however will be the HP Neural (3) model where false negative was the highest (61) and the false negative had the largest disparity between the false positive (44). Nonetheless, it is important to recognize that the HP Neural (3) model's performance, while not optimal, still holds up reasonably well when evaluated against the other contenders. In the ROC curve diagram shown below (*Figure 4*), it is apparent that consistent with the claims above, the Neural Network model, as depicted by the blue line, is closest to the top left corner of the square for the validation dataset. The AUC for that model is also the largest among all the other competing models.

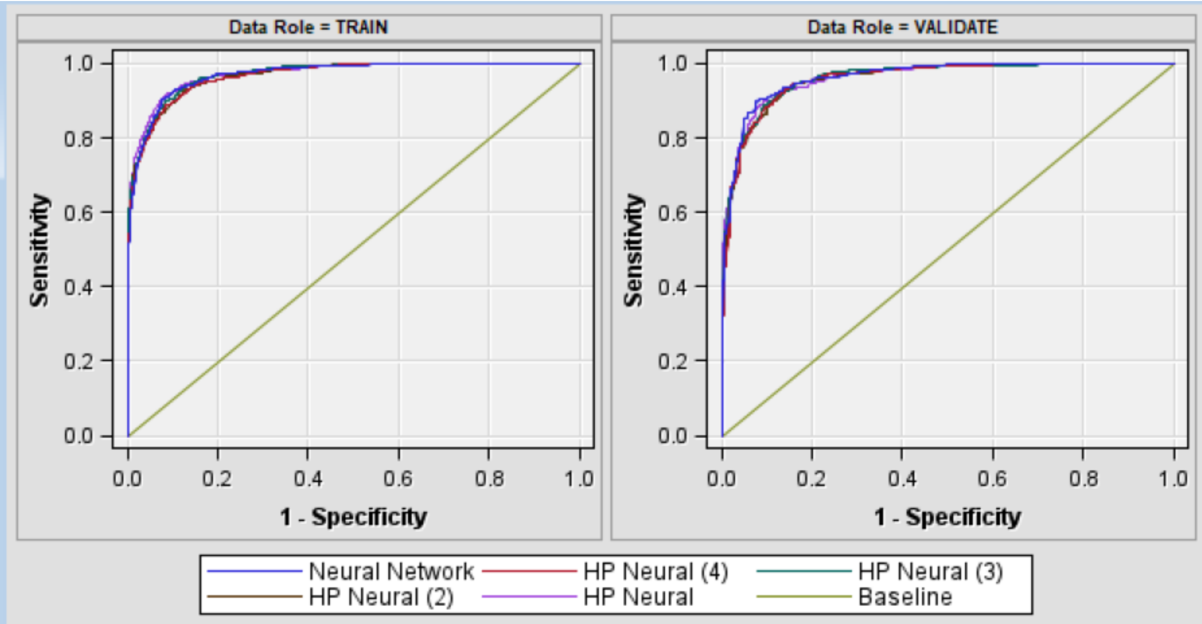


Figure 4: ROC Curve of Neural Network Models

Before proceeding with the interpretation of the best neural network identified (Neural Network), it is important to address the black box issue of neural networks. Neural networks are considered as black boxes because they exhibit the characteristics of having hidden layers which makes understanding and interpreting their decision-making process complicated and challenging. Hence, to aid in the interpretation of these complex models, a surrogate model ("Description Tree") was introduced into the data pipeline of this study. The surrogate model's insights into variable importance are intriguing, as there is consensus between the tree-based and neural network models on the significance of four variables (LOG_Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, and LOG_Total_Ct_Chng_Q4_Q1). However, a divergence arises in their respective fifth-ranked variable - the Gradient Boosting (2) model prioritizes LOG_Total_Amt_Chng_Q4_Q1 (*Section 2.5.1*), whereas the Neural Network model highlights Total_Relationship_Count as a top-five factor (*Figure 5*). Hence, this variation suggests that while this study can confidently focus on the four agreed-upon variables for immediate strategies to address customer churn, the fifth variable identified by each model—LOG_Total_Amt_Chng_Q4_Q1 by the Gradient Boosting (2) model and Total_Relationship_Count by the Neural Network model, warrants further detailed analysis in future investigations to fully understand their respective contributions to the prediction model.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Total_Trans_Ct		8	1.0000	1.0000	1.0000
Total_Revolving_Bal		2	0.5281	0.5742	1.0873
Total_Relationship_Count		3	0.3881	0.3971	1.0231
LOG_Total_Trans_Amt	Transformed Total_Trans_Amt	4	0.3873	0.4408	1.1381
LOG_Total_Ct_Chng_Q4_Q1	Transformed Total_Ct_Chng_Q4_Q1	1	0.2562	0.3066	1.1965
Months_Inactive_12_mon		1	0.1844	0.2167	1.1750

Figure 5: Variable Importance Table (Description Tree)

2.7 Summary

From the analysis conducted using both tree-based and neural network models, substantial findings have been obtained that respond directly to the aims of this study. The tree-based models, particularly the Gradient Boosting 2 (Boost2) model, have demonstrated superior performance with the lowest misclassification rates, effectively fulfilling the first objective of developing a predictive model for customer churn in the credit card domain. The Boost2 model, with its fine-tuned parameters, achieved the lowest misclassification rate and maintained a strong balance between sensitivity and precision, indicating it as the best model among its peers for generalizing well to unseen data without overfitting. In line with the second objective, the study identified critical factors influencing customer churn rates through variable importance analysis. Both model types concurred on the significance of four key variables: LOG_Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, and LOG_Total_Ct_Chng_Q4_Q1. The fifth variable differed between the models, highlighting an area for subsequent analysis to elucidate its impact fully.

The insights derived from these predictive models have been translated into actionable recommendations, addressing the third objective. Banks are advised to enhance customer engagement strategies focused on transaction behaviour and revolving balances. Alerting customers to significant transaction changes and fostering stronger relationships through targeted benefits can also aid in reducing attrition. For accounts with low card utilization, the introduction of programs to encourage responsible credit use without accruing unsustainable debt could improve customer retention. However, future studies are encouraged to delve deeper into the divergent fifth variable to consolidate the understanding of its influence on customer churn. Overall, this study has successfully met its three-fold objectives, offering a robust model for churn prediction, identifying pivotal churn determinants, and furnishing bank managers with evidence-based recommendations for strategic decision-making.

References

Fatema Akbar Mohamed, & Ali Khalifa Al-Khalifa. (2023). A review of machine learning methods for predicting churn in the telecom sector. *2023 International Conference On Cyber Management And Engineering (CyMaEn)* (pp. 164-170). IEEE. https://doi.org/10.1109/cymaen57228.2023.10051108

Goyal, S. (2021). *Credit Card customers* [Data file]. Retrieved from https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Kumar, S. (2024, February 20). Customer retention versus customer acquisition. Forbes. https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=79763d821c7d

Pahul Preet Singh, Fahim Islam Anik, Rahul Senapati, Sinha, A., Nazmus Sakib, & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, *7*(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Comput Sci*, *6*, e267. https://doi.org/10.7717/peerj-cs.267

Saradhi, V. V., & Palshikar, G. K. (2011) Employee churn prediction. *Expert Systems With Applications*, *38*, 19-30. https://doi.org/10.1016/j.eswa.2010.07.134

Wu, Z., & Li, Z. (2021). Customer churn prediction for commercial banks using customer-value-weighted machine learning models. *Journal of Credit Risk*, *17*(4), 15-42. DOI: 10.21314/JCR.2021.011