ASIA PACIFIC UNIVERSITY OF TECHNOLOGY & INNOVATION

GROUP ASSIGNMENT

GROUP MEMBERS: 1) LOKE WENG KHAY

2) LEE HAN SEN

3) LAU ZHI YI

INTAKE NUMBER : UC2F2102CS(DA)

MODULE NAME : Data Mining and Predictive Modelling

MODULE ID : CT119-3-2-DMPM

LECTURER : Dr. Preethi Subramanian

HAND OUT DATE : 10th August 2021

HAND IN DATE : 1st November 2021

Table of Contents

1 <u>Problem Statement</u>

Credit cards have played a crucial role in this ever-evolving world. A credit card is a payment method that allows customers to purchase goods and services without paying large sums of cash upfront. In addition, credit cards are able to provide the customer with security and convenience (Team, 2021).

Customers can apply for credit cards in any banks approved by the Central Bank of that country. Although the credit card business is lucrative, it also faces stiff competition between foreign and local banks. Based on the report by SHIFT Credit Card Processing, the average credit card per person in the United States is three credit cards (Credit Card Statistics, 2021). With this, banks will regularly give out rewards and bonuses to the customer, allowing them to stay longer and use the services that the banks provide. This is why it is crucial to give out bonuses and rewards to the right customer to avoid spending too much money and time on the least loyal customers.

2 Aim, Objectives and Scope

Aim

This research aims to apply data analysis techniques to understand the dataset's trend and generate a predictive or descriptive model to predict the outcome based on the historical dataset provided.

Objective

- To explore available datasets used to train and test generate the model
- To conduct pre-processing techniques and exploratory data analysis to get a general sense of the data.
- To use data mining techniques to analyse the dataset patterns and generate several predictive models to be used in this case study.

Scope

The dataset was published by Sakshi Goyal on a website called Kaggle in 2021. This dataset exists because a bank manager was worried that more and more customers are leaving their credit card service. Hence, the manager creates a credit card dataset to analyse the data to determine why customers are leaving and to predict customers who will drop their credit cards in the future. The dataset consists of 23 columns and 10128 rows.

3 Propose Methodology

For this project, SEMMA methodology will be applied. SEMMA methodology is a data-mining methodology developed by SAS Institute, a business analytic software company that access, analyse, and report the data to enhance decision-making for the company (Vishesh, 2020). SEEMA methodology tends to come out with previous undiscovered patterns, which might be an advantage in decision-making (SAS Institute Inc, 2017).

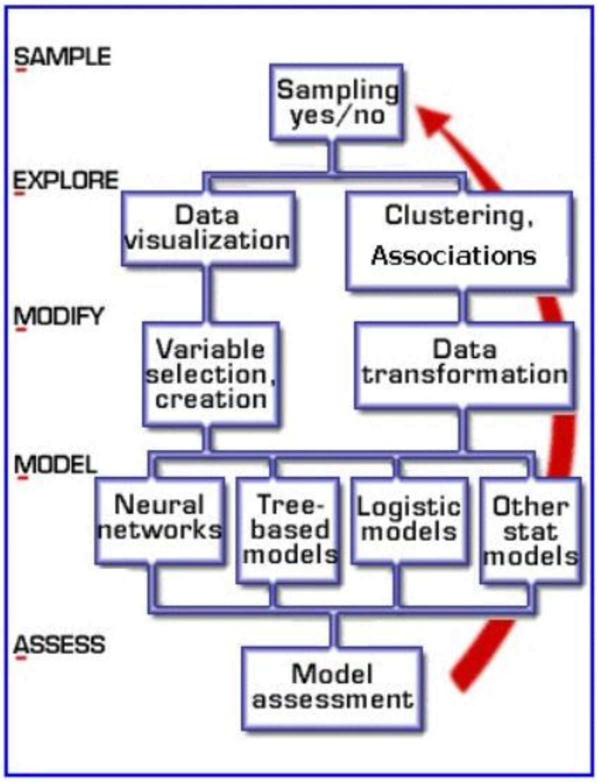


Figure 1: SEMMA Methodology *(SAS, 2017)*

Based on the figure above, there are five different levels in the SEMMA methodology, which are sample, explore, modify, model, and assess. A dataset named “BankChurner” was taken from Kaggle and uploaded to SAS Enterprise Miner to perform the study.

To start the study, sample data will be collected and uploaded to SAS Enterprise Miner. Then, the second and third level, which is “Explore” and “Modify” will be performing EDA (Explanatory Data Analysis) to ensure data cleanliness, searching for data relationship and trends to gain ideas, and transforming the useful variables into the models (SAS Institute Inc, 2017).

After explore and modify stage, it proceeds to model. During the model level, data will be taken into various modeling techniques such as neural networks, decision trees, and regression (Vishesh, 2020). Different modeling techniques will be chosen to apply according to their situation (Umair Shafique, 2014).

Lastly will be the assess level. The model that is done in the model stage is now evaluated for its usefulness and reliability. The data can now be tested and estimates its performance from the data mining process (SAS Institute Inc, 2017).

However, not all the levels need to be included in the SEMMA, depending on the situation. Also, some of the levels might repeat more than once to process to the next level.

4 Process Flow Diagram

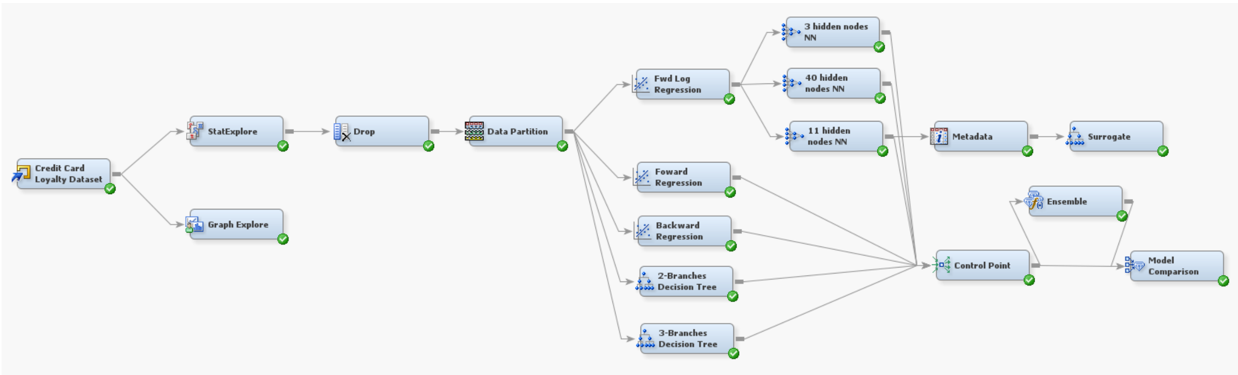


Figure 2: Process Flow Diagram

The figure above shows the process flow diagram to build various models for classification and prediction. These various models will be handed over to three persons to finish them. Loke Weng Khay will be building the 2-Branches and 3 Branches Decision Tree Models. Lee Han Sen will build the Forward and Backward Regression Models. Lau Zhi Yi will build the 3,11 and 40 Hiddens Neural Network Models.

5 Exploratory Data Analysis

Exploratory Data Analysis is the process of reviewing and performing initial studies on the data source to identify the data quality issues such as missing values, identifying outliers, noisy data, and many more. It also allows the data scientist to make assumptions with the help of the summary generated by the SAS Enterprise Miner (Exploratory Data Analysis, 2021).

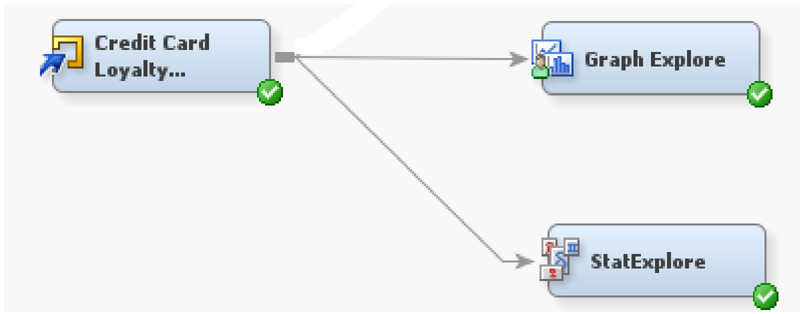


Figure 3: Nodes for Exploratory Data Analysis

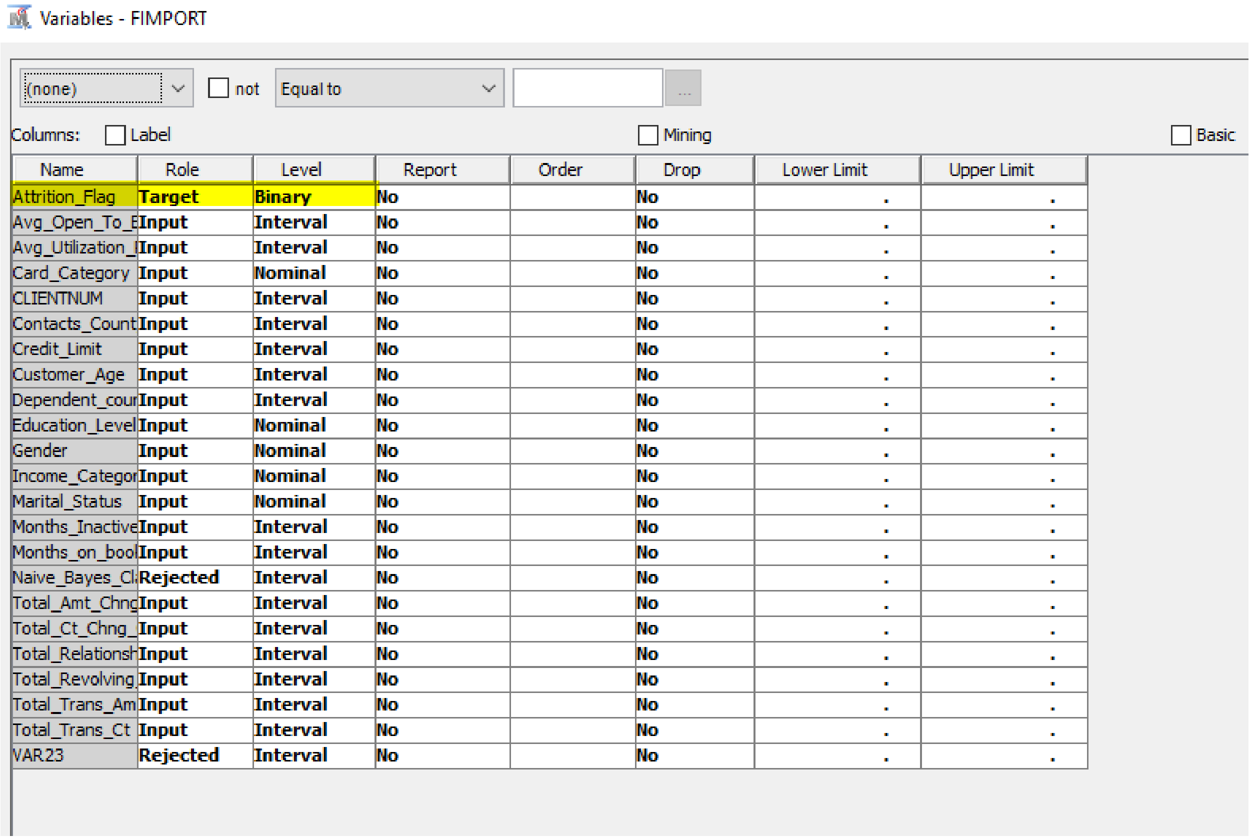
In figure 3, there are three nodes needed to perform Exploratory Data Analysis. The two nodes are:

- Graph Explore
- StatExplore

StatExplore is used in exploratory data analysis to get a general sense of the dataset. The StatExplore node can also be used to select variables for analysis, for profiling cluster and predictive model. Furthermore, it can be used to compute standard univariate and bivariate distribution statistics and correlation statistics for interval variables based on interval input and target (StatExplore Node, 2017).

The Graph Explore node is a graphical representation of the data that allows the data scientist to create multiple graphs that can be used to analyst the data source. Graph Explore node provides a simpler representation of data compared to StatExplore node.

Before any analysis can be done, a variable called “Attrition Flag” has to be set from Input to Target to allow it to train and validate the model based on the target variable. The below figure is the change for the variable role that has been done in the SAS Enterprise Miner.



Variables - FIMPORT

(none) | not | Equal to

Columns: Label | Mining | Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Attrition_Flag	Target	Binary	No		No	.	.
Avg_Open_To_E	Input	Interval	No		No	.	.
Avg_Utilization_	Input	Interval	No		No	.	.
Card_Category	Input	Nominal	No		No	.	.
CLIENTNUM	Input	Interval	No		No	.	.
Contacts_Count	Input	Interval	No		No	.	.
Credit_Limit	Input	Interval	No		No	.	.
Customer_Age	Input	Interval	No		No	.	.
Dependent_cour	Input	Interval	No		No	.	.
Education_Level	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Income_Categor	Input	Nominal	No		No	.	.
Marital_Status	Input	Nominal	No		No	.	.
Months_Inactive	Input	Interval	No		No	.	.
Months_on_bool	Input	Interval	No		No	.	.
Naive_Bayes_Cl	Rejected	Interval	No		No	.	.
Total_Amt_Chng	Input	Interval	No		No	.	.
Total_Ct_Chng_	Input	Interval	No		No	.	.
Total_Relationsh	Input	Interval	No		No	.	.
Total_Revolving	Input	Interval	No		No	.	.
Total_Trans_Am	Input	Interval	No		No	.	.
Total_Trans_Ct	Input	Interval	No		No	.	.
VAR23	Rejected	Interval	No		No	.	.

Figure 4: Change of variable “Attrition Flag” role in SAS Enterprise Miner

5.1 <u>Exploratory Data Analysis</u>

i. Missing variable

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	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	Attrition_Flag	TARGET	2	0	Existing Customer	83.93	Attrited Customer	16.07

Figure 5: Summary for all variables in StatExplore (Part 1)

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Avg_Open_To_Buy	INPUT	7469.14	9090.685	10127	0	3	3474	34516	1.661697	1.798617
Avg_Utilization_Ratio	INPUT	0.274894	0.275691	10127	0	0	0.176	0.999	0.718008	-0.79497
CLIENTNUM	INPUT	7.3918E8	36903783	10127	0	7.0808E8	7.1793E8	8.2834E8	0.995601	-0.61564
Contacts_Count_12_mon	INPUT	2.455317	1.106225	10127	0	0	2	6	0.011006	0.000863
Credit_Limit	INPUT	8631.954	9088.777	10127	0	1438.3	4549	34516	1.666726	1.808989
Customer_Age	INPUT	46.32596	8.016814	10127	0	26	46	73	-0.03361	-0.28862
Dependent_count	INPUT	2.346203	1.298908	10127	0	0	2	5	-0.02083	-0.68302
Months_Inactive_12_mon	INPUT	2.341167	1.010622	10127	0	0	2	6	0.633061	1.098523
Months_on_book	INPUT	35.92841	7.986416	10127	0	13	36	56	-0.10657	0.4001
Total_Amt_Chng_Q4_Q1	INPUT	0.759941	0.219207	10127	0	0	0.736	3.397	1.732063	9.993501
Total_Ct_Chng_Q4_Q1	INPUT	0.712222	0.238086	10127	0	0	0.702	3.714	2.064031	15.68929
Total_Relationship_Count	INPUT	3.81258	1.554408	10127	0	1	4	6	-0.16245	-1.00613
Total_Revolving_Bal	INPUT	1162.814	814.9873	10127	0	0	1276	2517	-0.14884	-1.14599
Total_Trans_Amt	INPUT	4404.086	3397.129	10127	0	510	3899	18484	2.041003	3.894023
Total_Trans_Ct	INPUT	64.85869	23.47257	10127	0	10	67	139	0.153673	-0.36716

Figure 6: Summary for all variables in StatExplore (Part 2)

Based on figure 5 and figure 6, there are no missing data for all variables in this dataset. Hence, no imputation has to be applied for the missing data.

ii. Handling Outlier

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Avg_Open_To_Buy	INPUT	7469.14	9090.685	10127	0	3	3474	34516	1.661697	1.798617
Avg_Utilization_Ratio	INPUT	0.274894	0.275691	10127	0	0	0.176	0.999	0.718008	-0.79497
CLIENTNUM	INPUT	7.3918E8	36903783	10127	0	7.0808E8	7.1793E8	8.2834E8	0.995601	-0.61564
Contacts_Count_12_mon	INPUT	2.455317	1.106225	10127	0	0	2	6	0.011006	0.000863
Credit_Limit	INPUT	8631.954	9088.777	10127	0	1438.3	4549	34516	1.666726	1.808989
Customer_Age	INPUT	46.32596	8.016814	10127	0	26	46	73	-0.03361	-0.28862
Dependent_count	INPUT	2.346203	1.298908	10127	0	0	2	5	-0.02083	-0.68302
Months_Inactive_12_mon	INPUT	2.341167	1.010622	10127	0	0	2	6	0.633061	1.098523
Months_on_book	INPUT	35.92841	7.986416	10127	0	13	36	56	-0.10657	0.4001
Total_Amt_Chng_Q4_Q1	INPUT	0.759941	0.219207	10127	0	0	0.736	3.397	1.732063	9.993501
Total_Ct_Chng_Q4_Q1	INPUT	0.712222	0.238086	10127	0	0	0.702	3.714	2.064031	15.68929
Total_Relationship_Count	INPUT	3.81258	1.554408	10127	0	1	4	6	-0.16245	-1.00613
Total_Revolving_Bal	INPUT	1162.814	814.9873	10127	0	0	1276	2517	-0.14884	-1.14599
Total_Trans_Amt	INPUT	4404.086	3397.129	10127	0	510	3899	18484	2.041003	3.894023
Total_Trans_Ct	INPUT	64.85869	23.47257	10127	0	10	67	139	0.153673	-0.36716

Figure 7: Skewness level for all variables from this dataset

From this figure, we can see that there is no outlier issue for all the variables in this dataset. Therefore, no outlier is needed to be fixed or handled for this dataset.

iii. Variable Worth

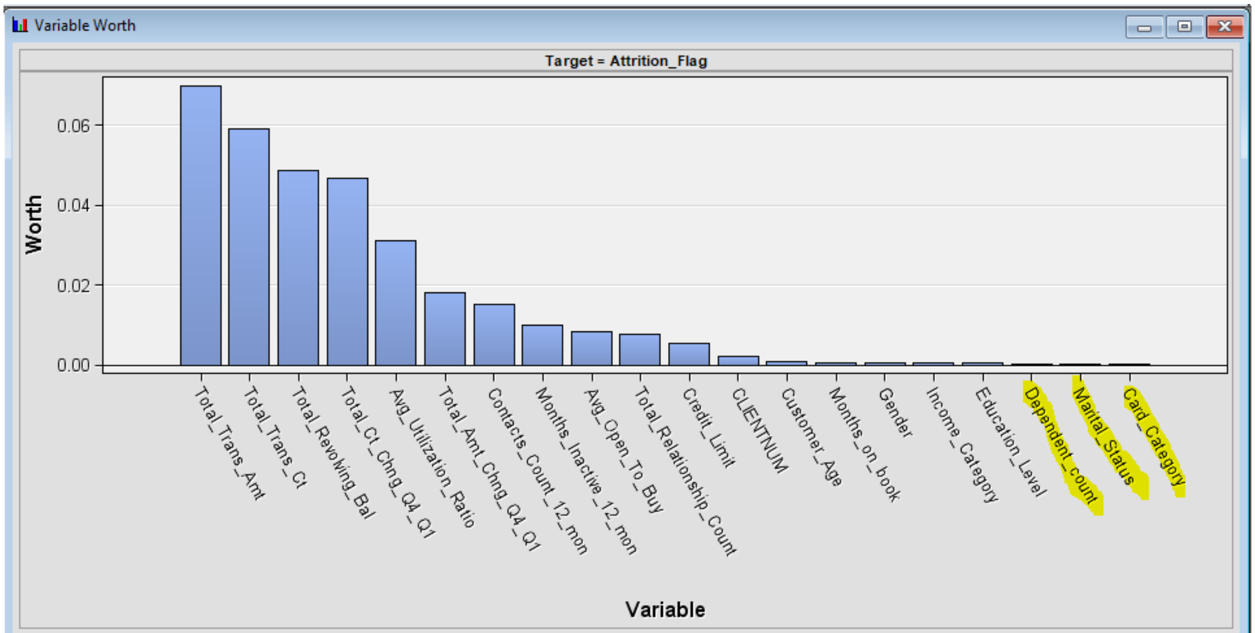


Figure 8: Variable Worth for all variables in the dataset

From figure 8, we can see that the bottom three variable in the bar chart has the lowest variable worth, which is "Dependent_count", "Marital_Status" and "Card_Category". For the three variables, it will be dropped when performing data pre-processing.

iv. Low number of cluster group/number of level

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	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	Attrition_Flag	TARGET	2	0	Existing Customer	83.93	Attrited Customer	16.07

Figure 9: Group cluster

From figure 9, there is only two group cluster available for the variable “Attrition_Flag” which is suitable to be the Target Role as it helps prevent complication when creating a model from this dataset.

v. Chi Variable

Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=Attrition_Flag

Input	Chi-Square	Df	Prob
Gender	14.0682	1	0.0002
Income_Category	12.8323	5	0.0250
Education_Level	12.5112	6	0.0515
Marital_Status	6.0561	3	0.1089
Card_Category	2.2342	3	0.5252

Figure 10: Chi-Square Statistics

figure 10 represents the chi-square statistics for class variables. Class variables such as “Gender” and “Income_Category” represent the best variable to be used to predict the target variables. This is because the class variable probability is below 0.05, which is the best to be used to predict the target variables.

5.2 Graphical Analysis

Attrition Flag (Bar Chart)

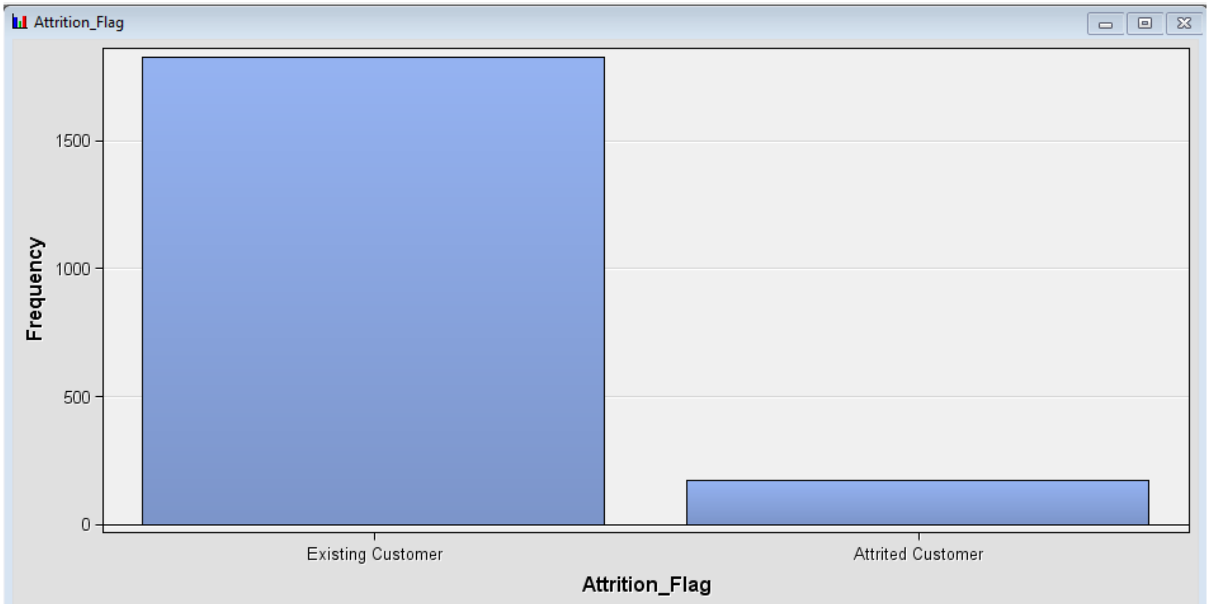


Figure 11: Bar chart for attrition flag

Figure 11 shows the bar chart for the attrition flag of the customer in the dataset. There are a total of 2000 records in this dataset. The frequency of existing customers is 1829, whereas the frequency of attrited customers is 171.

Customer age (histogram)

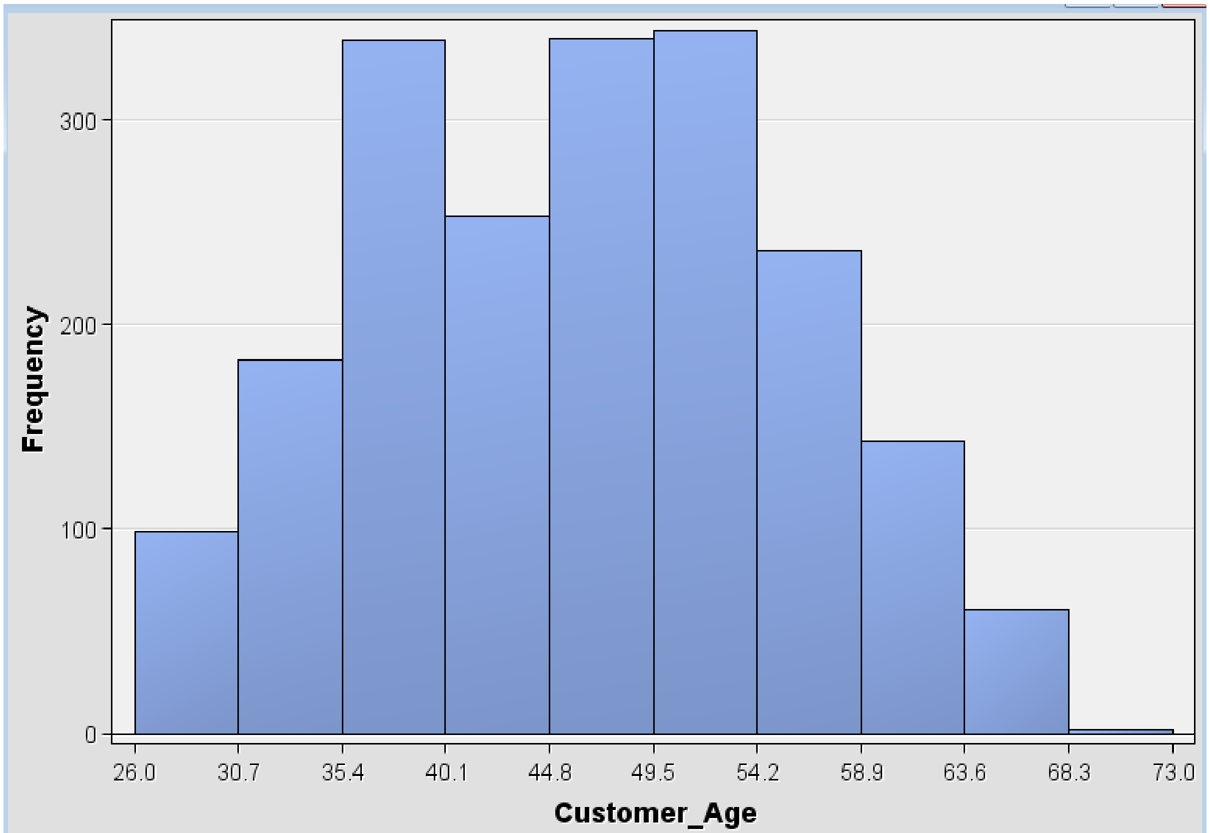


Figure 12: Histogram for customer age

Figure 12 shows the histogram of the age of the customers. By observing the histogram, it clearly shows that the youngest customer aged 26 whereas the oldest customer aged 73. There are outliers in this dataset which is customers from the age 68.3 to 73. This histogram is a left-skewed histogram as the skewness value of this histogram is -0.03361. The mean value for the customer age is 42.32596, and the standard deviation value for the customer is 8.016814.

Gender (bar chart)

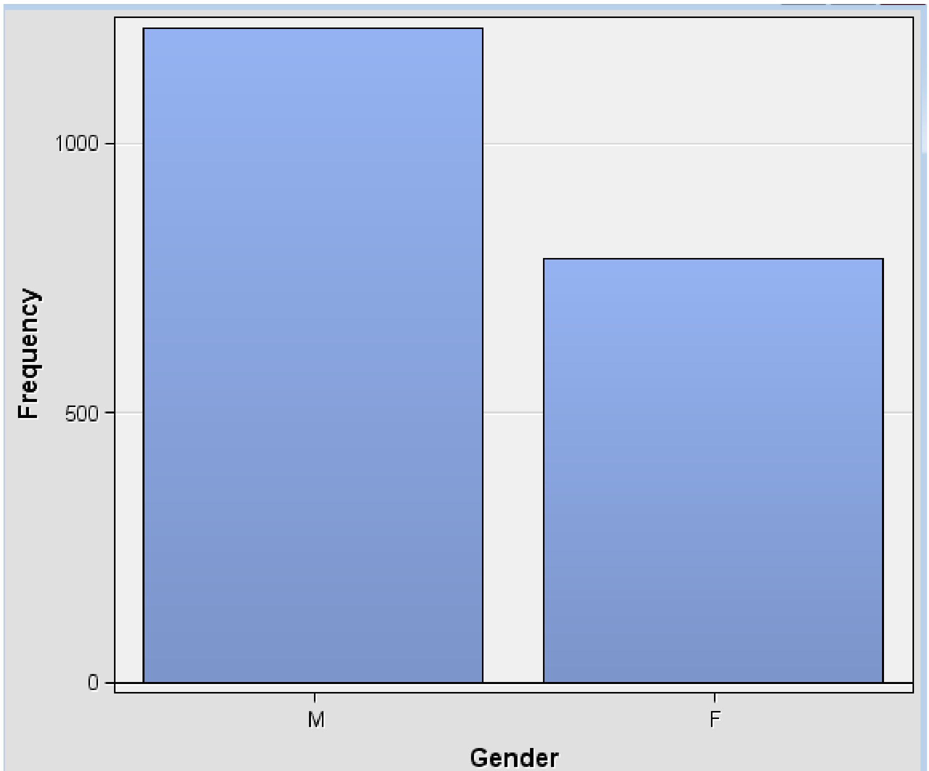


Figure 13: Bar chart for gender

Figure 13 shows the bar chart for customer gender. There are a total of 2000 customer records in this dataset. Most of the customers are male, having a total of 1215 frequencies. At the same time, the total number of female customers is 785.

Income Category (Pie Chart)

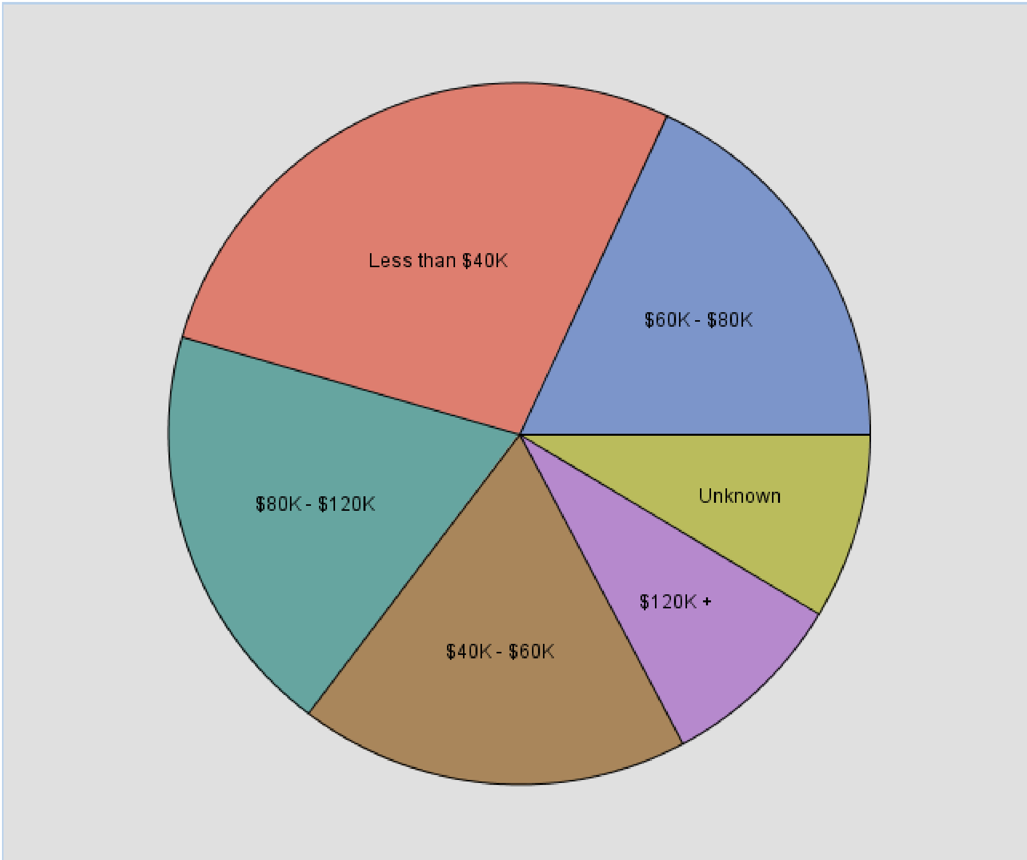


Figure 14: Pie chart for income category

Figure 14 shows the income category of the customer. By observing the pie chart, the majority of the customer falls in the category of less than $40k, having 547 customers, which cover 27.35% of the total chart. The least percentage category of the pie chart is the customer having their income more than $120k, having a total number of 172 customers and covering 8.6% of the total percentage. The second-largest customer income category falls at $80k to $120k. Having a total number of 385 customers and covers 19.25%. The third-largest customer income category falls at $60k to $80k. Having a total number of 362 customers and covers 18.1%. The fourth-largest customer income category falls at $40k to $60k. Having a total number of 359 customers and covers 17.95%. However, this dataset contains 175 unknown records of customer income.

Education Level (Pie Chart)

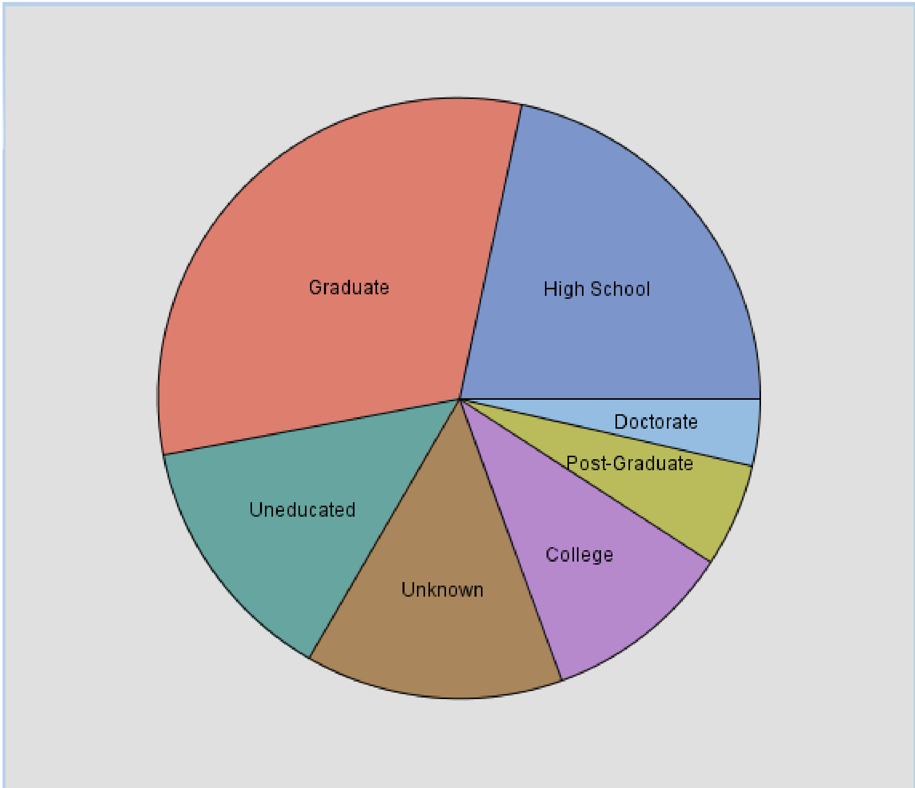


Figure 15: Pie chart for education level

Figure 15 shows the education level of the customer. By observing the pie chart, most customers fall in the category of graduate, having 630 customers and covering 31.5% of the total chart. The least percentage category of the pie chart is customers achieving doctorate education level, having 73 numbers of customers and covering 3.65% of the total percentage.

Marital Status (Bar Chart)

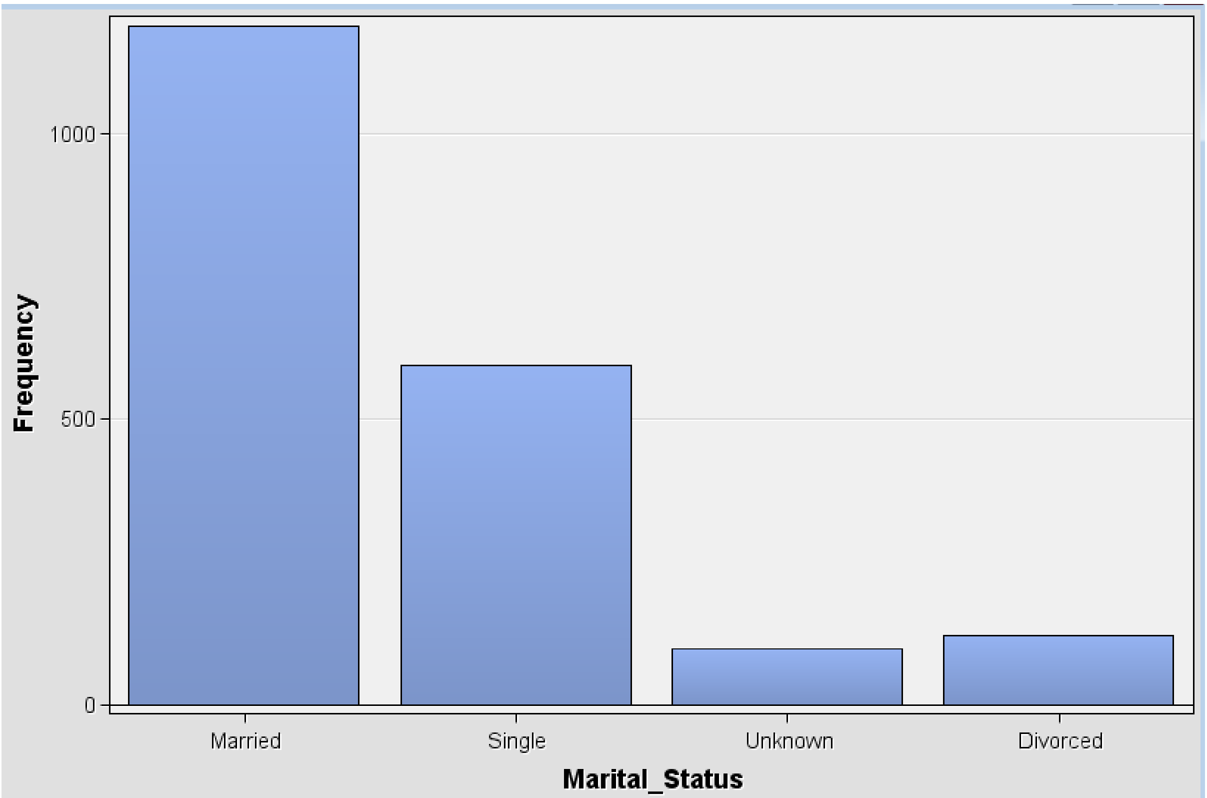


Figure 16: Bar chart for marital status

Figure 16 shows the marital status of the customer. This bar chart consists of a total of 2000 records of the customer. By observing the bar chart, most of the customers are married, having a number of 1189 records. Excluding the unknown category, the lowest record of martial status falls at the divorced category, having only 121 records.

Months Inactive 12 months (Bar chart)

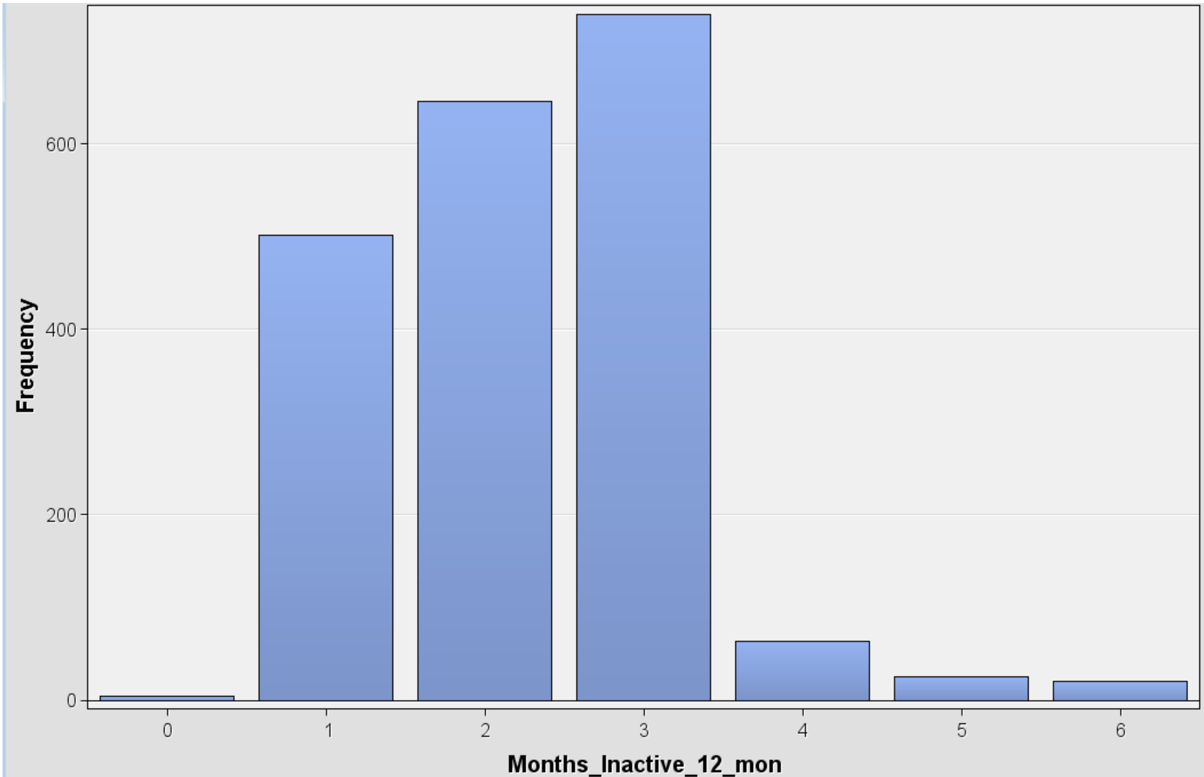


Figure 17: Bar chart for months inactive in the last 12 months

Figure 17 shows the bar chart for the existing customer that has their card inactive for the past 12 months. By observing the bar chart, the highest frequency number of months inactive is 3 months which has 740 customers. The lowest frequency number of months inactive is 0 months, which only has 4 records.

Relationship between card category and Attrition Flag (Bar chart)

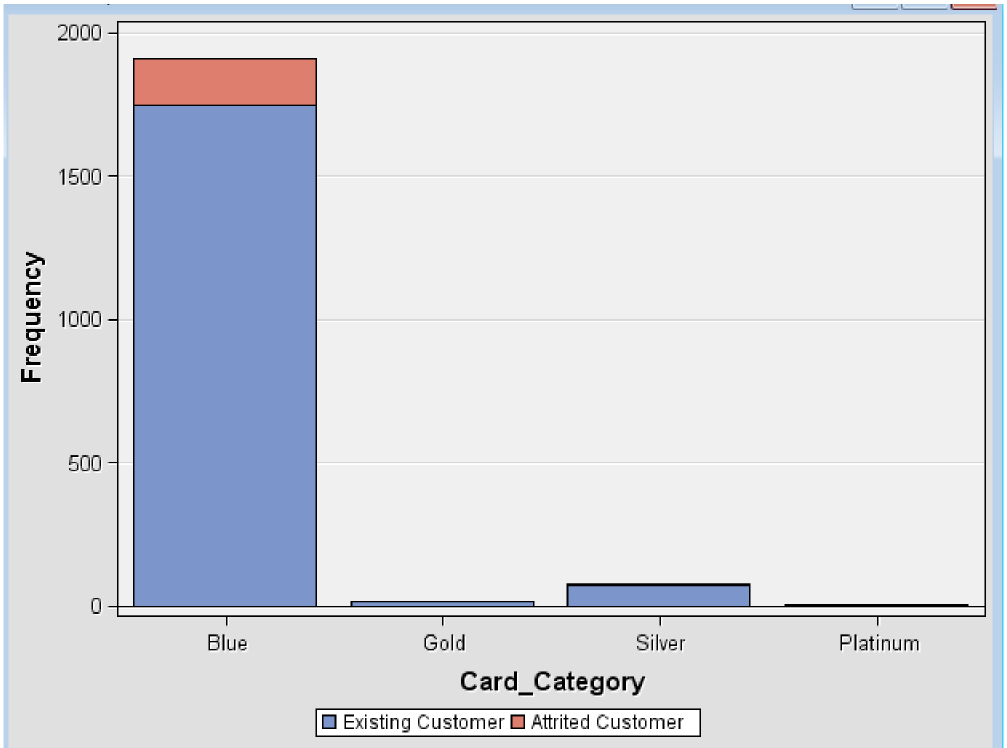


Figure 18: Bar chart for the relationship between card category and attrition flag

Figure 18 shows the relationship between card category and attrition flag. By observation, the blue card category has the highest number of customers, which includes 1744 existing customers and 166 attrited customers. The platinum card category has the least number of customers, having only 2 existing customers.

Relationship between customer age and months on book (Scatter plot)

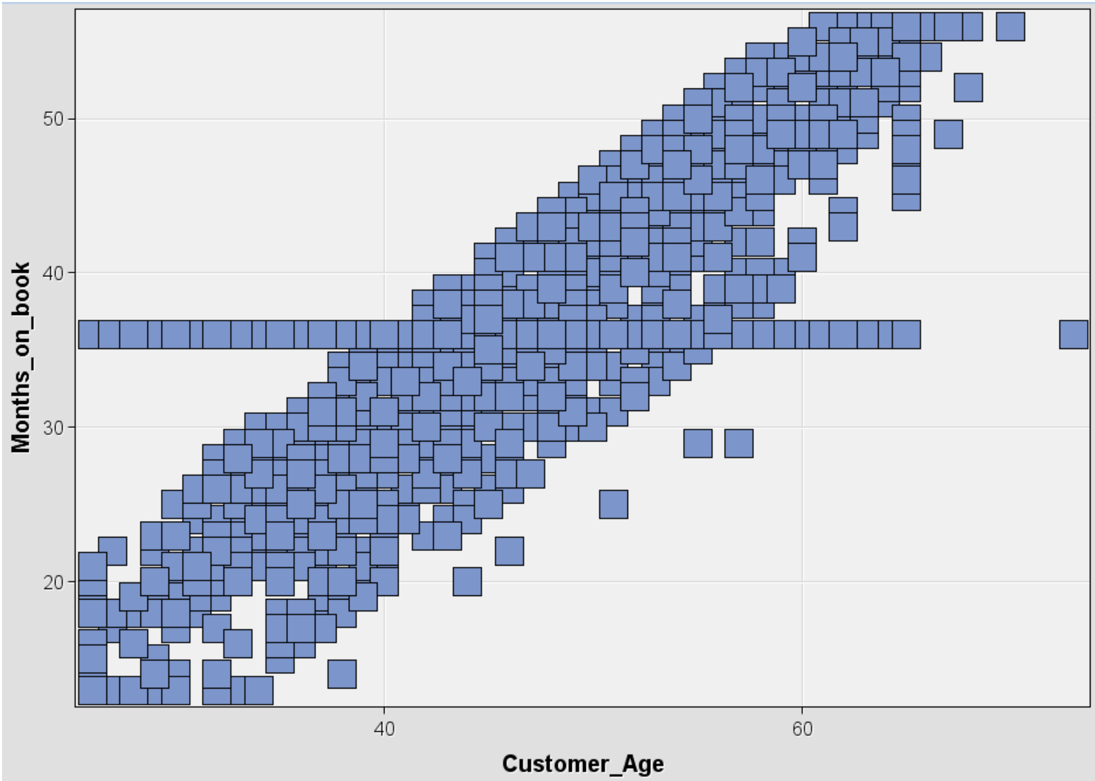


Figure 19: Scatter plot for customer age based on months on the book

Figure 19 shows the scatter plot for the relationship between customer age and period of relationship with the bank. By observation, the notable pattern for this figure will be the higher the customer age, the more the months that they had relationship with the bank. In general, the older the customer, the more loyal to the bank they are.

Relationship between total trans ct, customer age and Attrition Flag (Scatter plot)

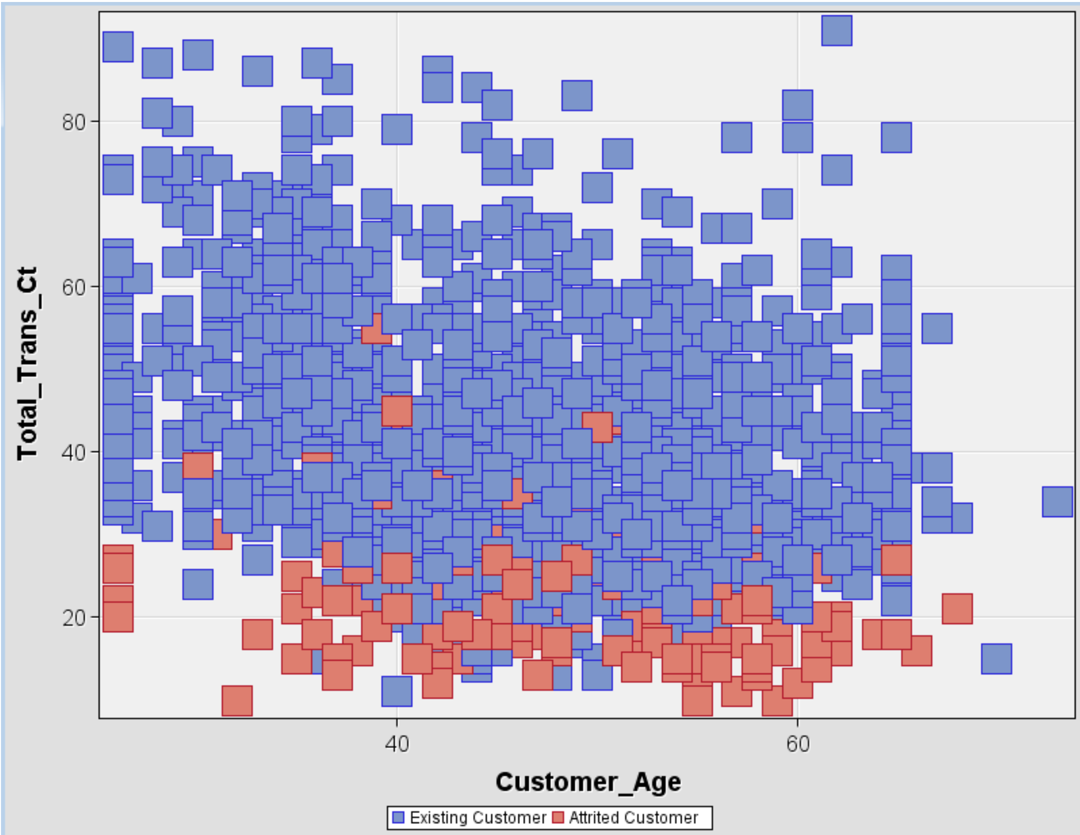


Figure 20: Scatter plot for the relationship between total transaction count based on customer age and attrition flag.

Figure 20 shows the scatter plot for the relationship between total transaction count based on customer age and attrition flag. A notable pattern that can be discovered in this graph is no matter how old the customer is, most of the transaction amounts are above 20. This applies to the existing customer only.

Relationship between total relationship count and Attrition Flag (Bar chart)

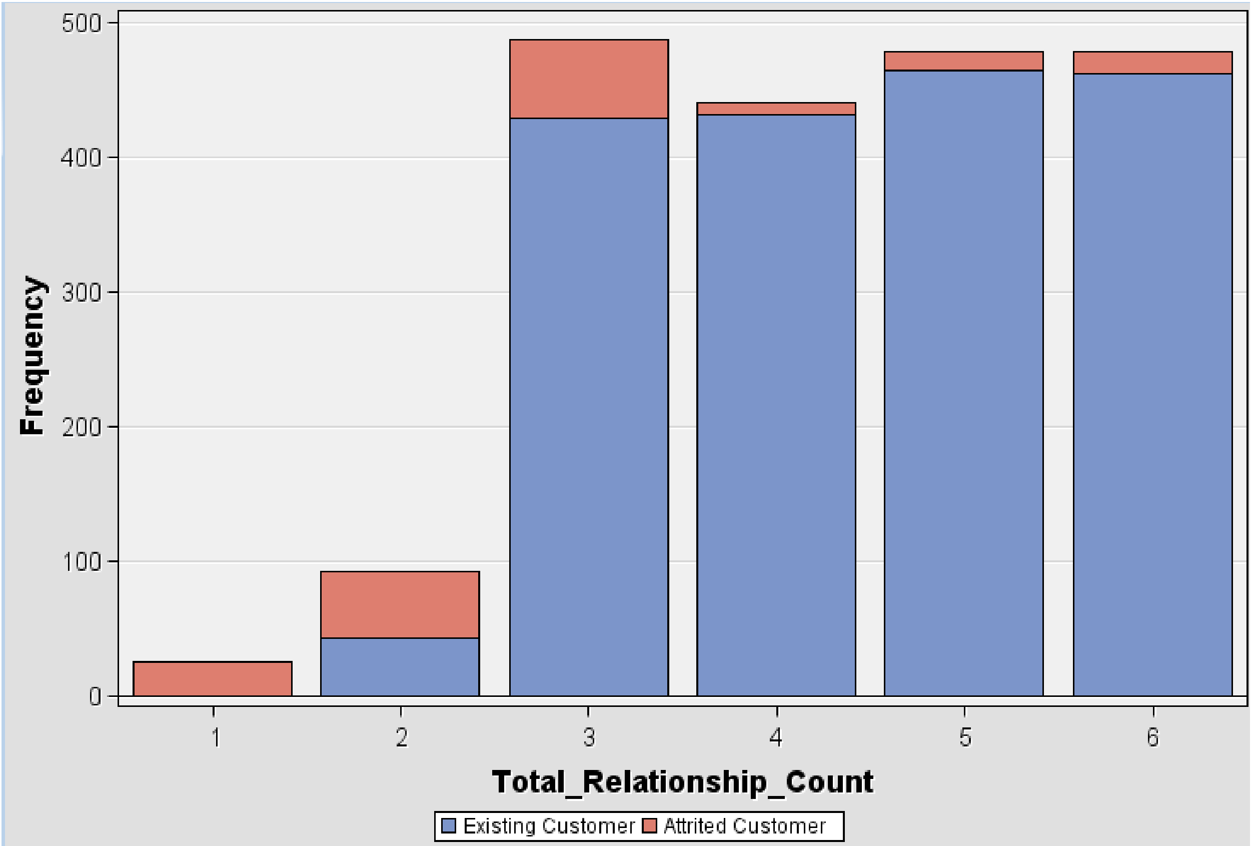


Figure 21: Bar chart for total relationship count group by attrition flag

The figure shows the bar chart for the total relationship count group by attrition flag. By observing the graph above, most of the customers having at least 3 relationships count with the bank. The relationship counts here means total products by the bank held by the customer.

6 Data Pre-Processing

Pre-processing is the process of refining and fixing any problems that exist in the dataset. Pre-Processing techniques are the process of resolving issues, errors, incomplete and inconsistent data. Without pre-processing, it can cause the machine learning model not to have the best accuracy and precision rate overall.

In this section, we will explain a few preprocessing techniques that need to be applied to the dataset before any development of machine learning starts to happen. For this dataset, there is no missing data, outlier issues, and inconsistent data. Operations needed to be done to drop low-worth variables and perform data partitioning.

6.1. Drop Low Worth Variable

From figure 8, we can see that the lowest variable worth which is “Dependent_count”, “Marital_Status”, and “Card_Category”. For the three variables, it will be dropped in this section. The figure below is the process of dropping the three variables

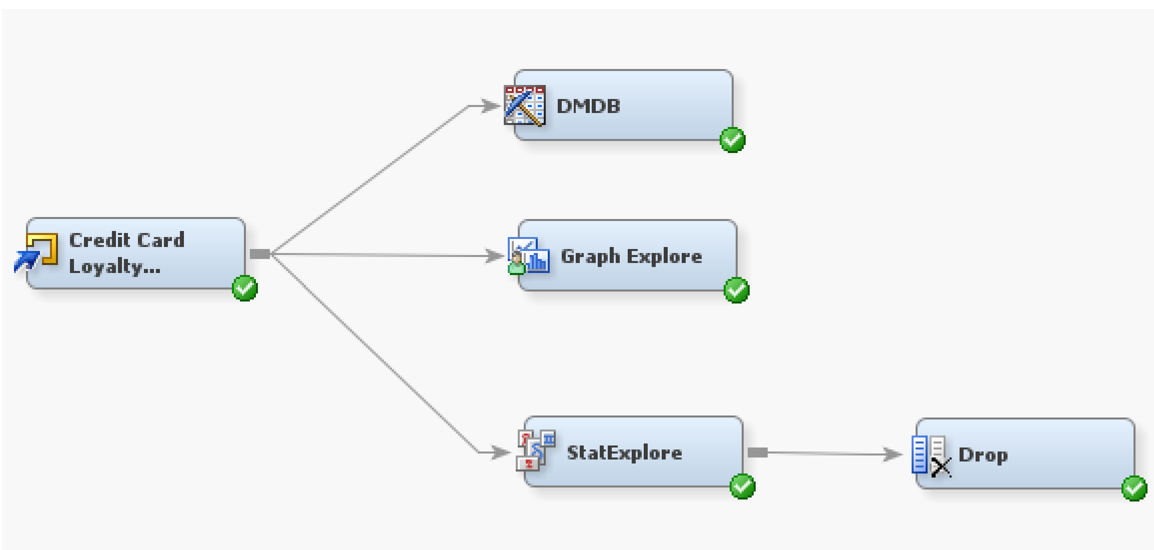
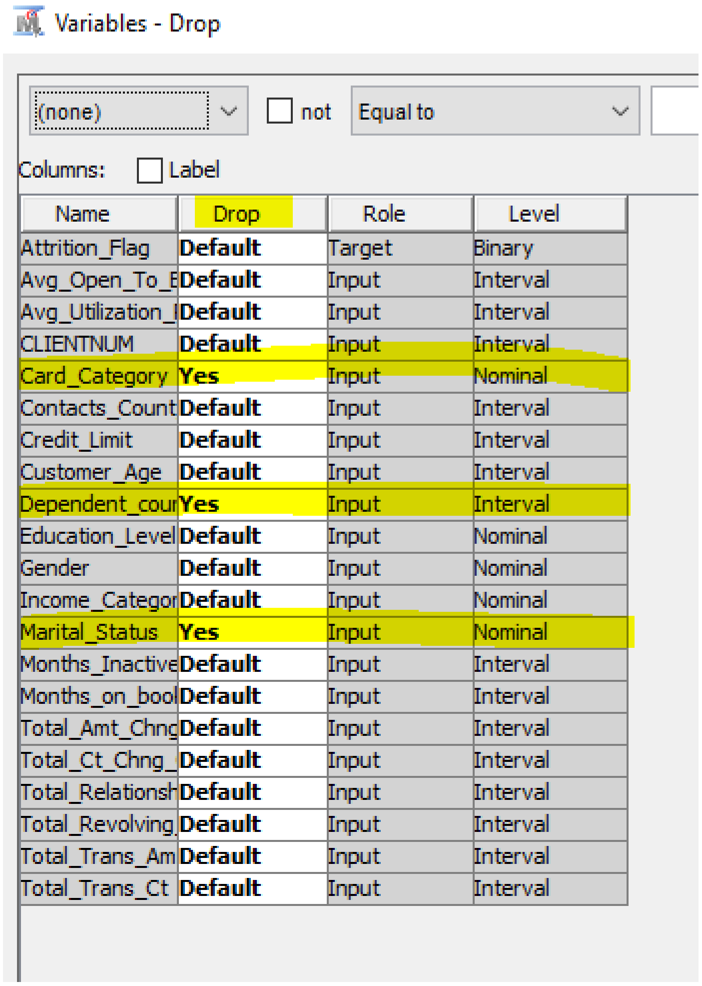


Figure 22: The “DROP” node used to drop low worth variable


Variables - Drop

(none) | not | Equal to

Columns: Label

Name	Drop	Role	Level
Attrition_Flag	Default	Target	Binary
Avg_Open_To_E	Default	Input	Interval
Avg_Utilization_	Default	Input	Interval
CLIENTNUM	Default	Input	Interval
Card_Category	Yes	Input	Nominal
Contacts_Count	Default	Input	Interval
Credit_Limit	Default	Input	Interval
Customer_Age	Default	Input	Interval
Dependent_cour	Yes	Input	Interval
Education_Level	Default	Input	Nominal
Gender	Default	Input	Nominal
Income_Categor	Default	Input	Nominal
Marital_Status	Yes	Input	Nominal
Months_Inactive	Default	Input	Interval
Months_on_bool	Default	Input	Interval
Total_Amt_Chng	Default	Input	Interval
Total_Ct_Chng_	Default	Input	Interval
Total_Relationsh	Default	Input	Interval
Total_Revolving	Default	Input	Interval
Total_Trans_Am	Default	Input	Interval
Total_Trans_Ct	Default	Input	Interval

Figure 23: Setting to drop the variable in the “DROP” node

6.2. Data Partition

Once the low variable data have been dropped, the following steps are to perform data partitioning. To make this possible is uses “Data Partition” Node. The cleaned dataset is to partition the data into two sets which are the training set and validation set, before the development of any of the machine learning models. This is important as it allows the models to validate themselves once the models are done training themselves. The figure below is the process of applying the “Data Partition” Node.

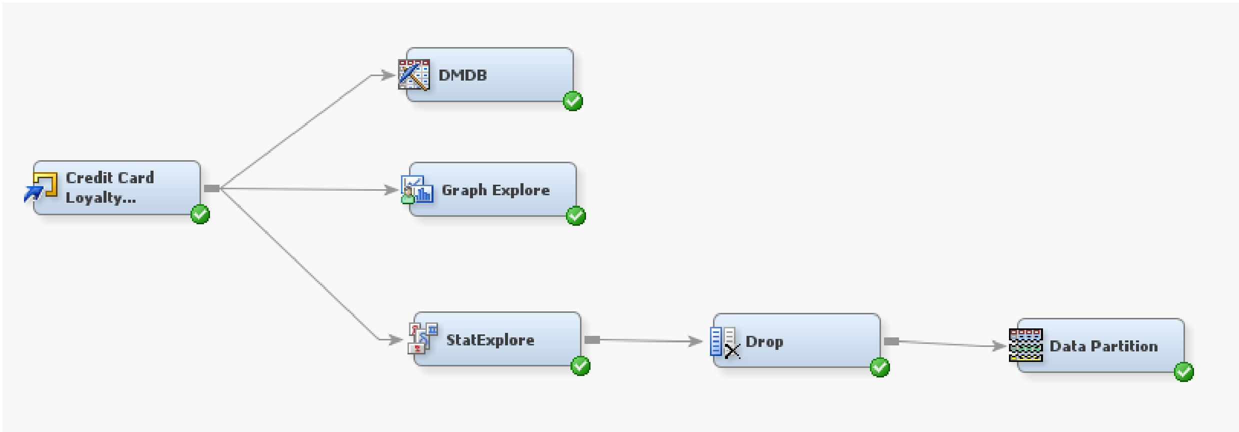
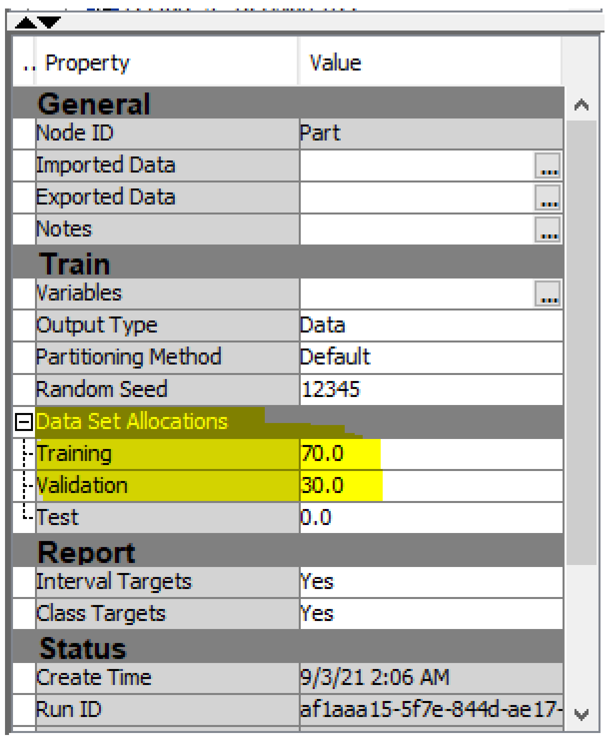


Figure 24: The “Data Partition” node used to partition the cleaned dataset into two dataset



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	9/3/21 2:06 AM
Run ID	af1aaa15-5f7e-844d-ae17-

Figure 25: Setting for the “Data Partition” node

7 Decision Tree (Loke Weng Khay – TP062166)

7.1. Diagram Flow for decision tree

A Decision Tree is a supervised learning algorithm where it can solve regression and classification problems. The goal of a decision tree is to create a training model using a dataset or training data which will allow the decision tree to predict the correct class or a specific target variable (Chauhan, n.d.).

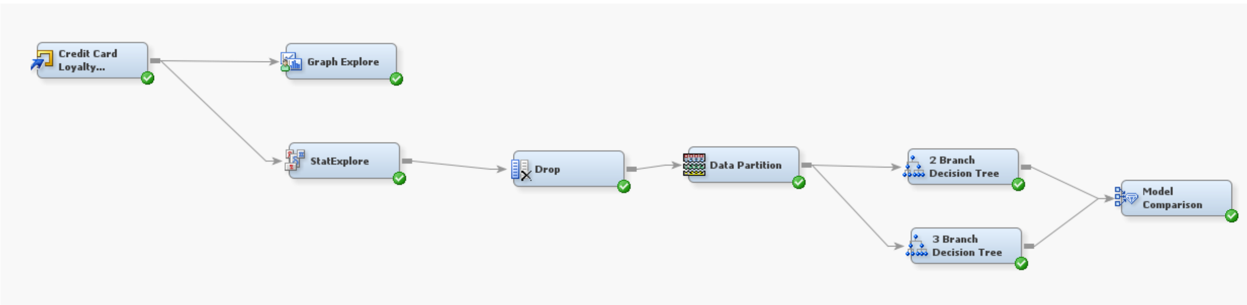


Figure 26: Diagram Flow for decision tree

The figure above represents the Diagram Flow for the decision tree. Two types of decision trees will be developed, which are 2 Branch Decision Tree and 3 Branch Decision Tree. The process flow diagram is developed using SAS Enterprise Miner.

7.2. Model Construction and Optimization

7.2.1. Decision Tree After Pruning (2 Branches)

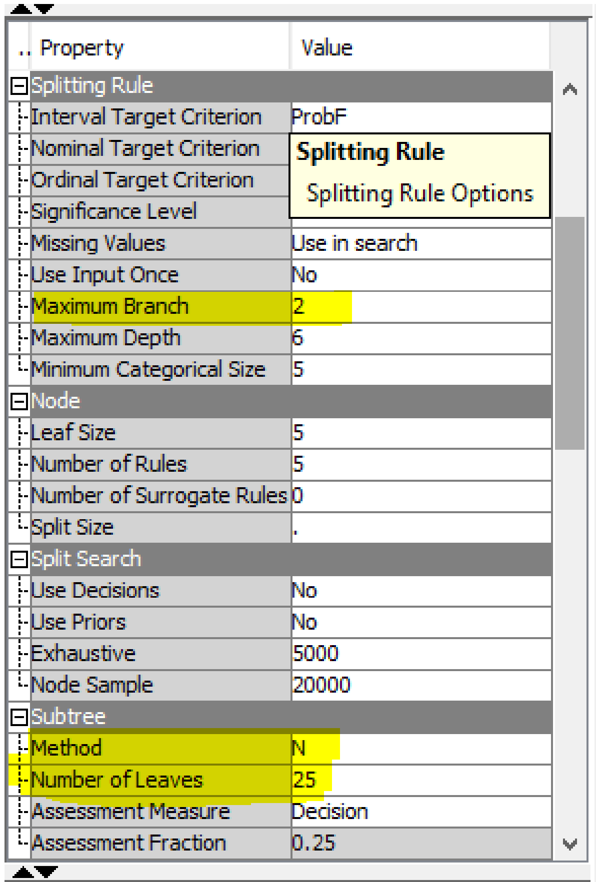


Figure 27: Settings for 2 Branch Decision Tree

Figure 27 represents the settings needed to be applied to generate a 2 Branch Decision Tree with 25 leaves. The number of leaves is based on the misclassification rate graph from figure 29. The graph displays the optimal number of leaves where the misclassification rate is the lowest with no overfitting.

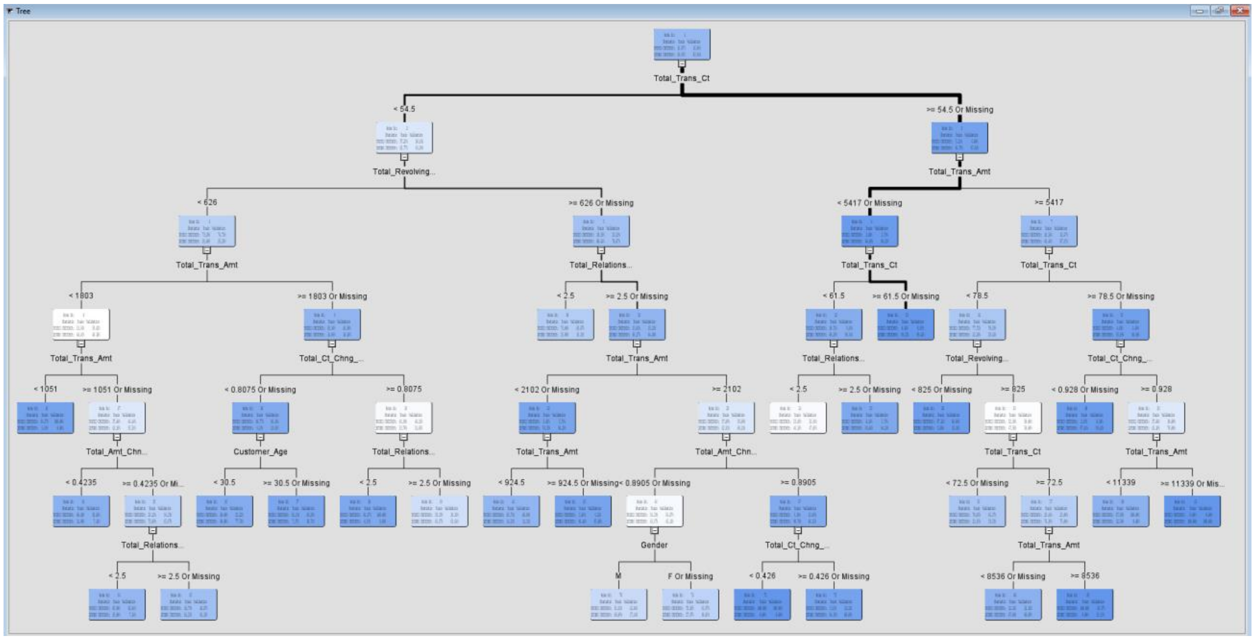


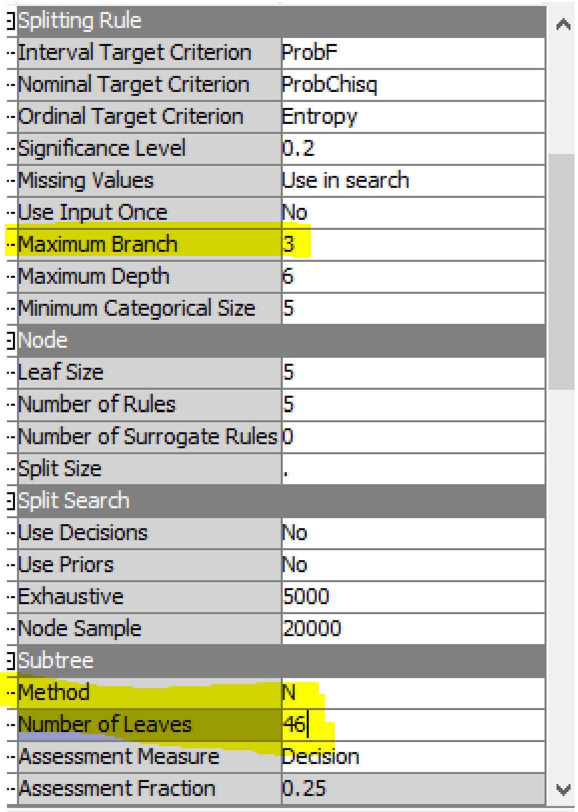
Figure 28: 2 Branch Decision Tree After Pruning

Once the settings have been applied, the figure above is the 2 Branch Decision Tree after pruning. All the variables that are used to develop the decision tree are “Total_Trans_Ct”, “Total_Revolving_Bal”, “Total_Trans_Amt”, “Total_Relationship_Count”, “Total_Amt_Chng_Q4_Q1”, “Total_Ct_Chng_Q4_Q1”, “Gender”, and “Customer_Age”. With the eight variables, it is able to develop a decision tree that is able to predict and distinguish between the loyal customer and non-loyal customer in this credit card industry.

Figure 29: Misclassification Rate for 2 Branch Decision Tree

Based on figure 29 shows the validation misclassification rate is at 0.0550 or (5.5%) at 25 leaves. The training misclassification rate is at 0.0516 or (5.1%) at 25 leaves. This shows that there is no overfitting or underfitting. The Subtree Assessment Plot graph also displays the optimal number of leaves where the misclassification rate is the lowest for validation and training. This is important as it shows that the misclassification rate will not improve anymore if the number of leaves increases.

7.2.2. Decision Tree After Pruning (3 Branches)



Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	N
Number of Leaves	46
Assessment Measure	Decision
Assessment Fraction	0.25

Figure 30: Settings for 3 Branch Decision Tree

Figure 30 represents the settings needed to be applied to generate a 3 Branch Decision Tree with 46 leaves. The number of leaves is based on the misclassification rate graph from figure 32. The graph displays the optimal number of leaves where the misclassification rate is the lowest with no overfitting.

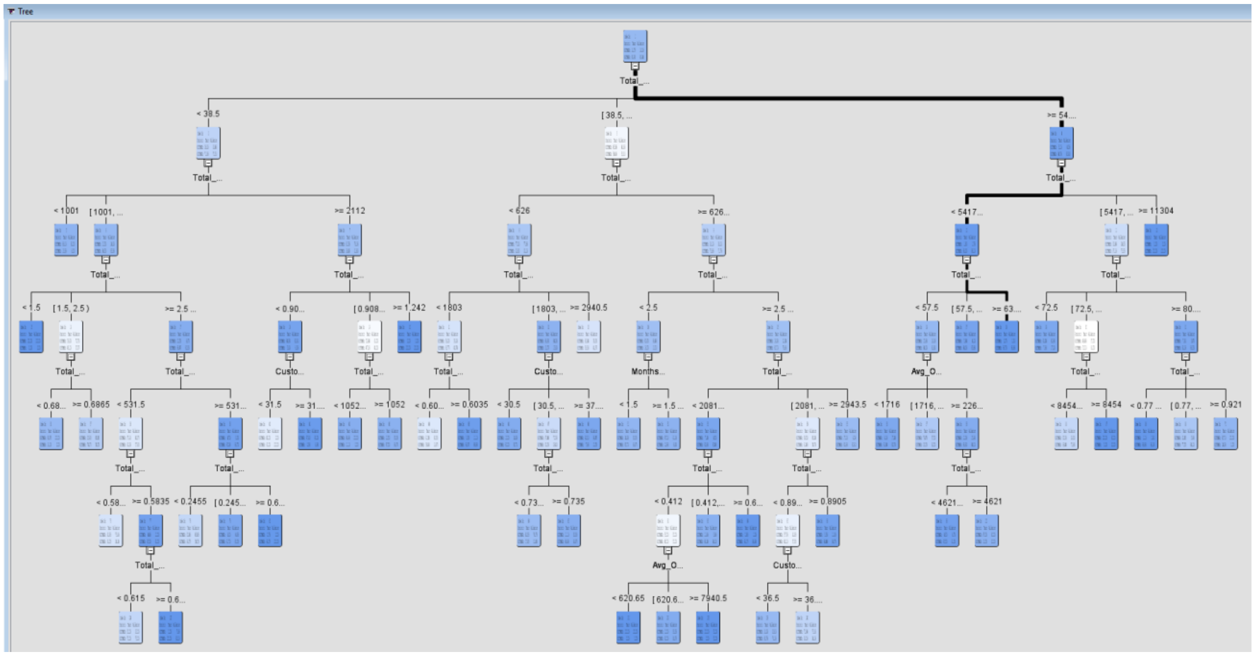


Figure 31: 3 Branch Decision Tree after Pruning

Once the settings have been applied, the figure above is the 3 Branch Decision Tree after pruning. All the variables that are used to develop the decision tree are “Total_Trans_Ct”, “Total_Trans_Amt”, “Total_Revolving_Bal”, “Total_Relationship_Count”, “Total_Amt_Chng_Q4_Q1”, “Total_Ct_Chng_Q4_Q1”, “Customer_Age”, “Avg_Open_To_Buy” and “Months_Inactive_12_mon”. With the nine variables, it is able to develop a decision tree that is able to predict and distinguish between the loyal customer and non-loyal customer in this credit card industry.

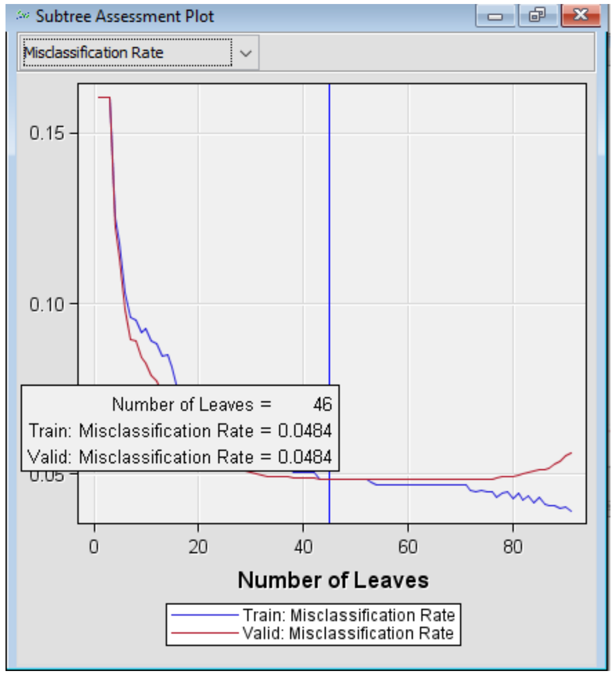


Figure 32: Misclassification Rate for 3 Branch Decision Tree

Based on figure 32 shows the validation misclassification rate is at 0.0484 or (4.84%) at 46 leaves. For the training, the misclassification rate is at 0.0484 or (4.84%) at 46 leaves. This shows that there is no overfitting or underfitting. The Subtree Assessment Plot graph displays the optimal number of leaves with the lowest misclassification rate for validation and training. This is important as it shows that the misclassification rate will not improve anymore if the number of leaves increases.

7.3. Model Validation

7.3.1. Summary of All Decision Tree Model

Decision Tree Method	No. of Leaves	Validation Misclassification Rate
Decision Tree After Pruning – 2 Branches	25	0.051
Decision Tree After Pruning – 3 Branches	46	0.048

Table 1: Summary of all decision tree model

Based on the summary above, we can conclude that the best decision tree model is the 3 Branch Decision tree with 46 leaves with the lowest Misclassification Rate of 0.048 or (4.8%). The second best decision tree model is the 2 Branch Decision tree with 25 leaves which has the lowest Misclassification Rate of 0.051 or (5.1%).

7.3.2. Event Classification Table

```
Event Classification Table

Data Role=TRAIN Target=Attrition_Flag Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

   174         970         169        5775


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

  False       True        False       True
Negative    Negative    Positive    Positive

   78          419         69         2473
```

Figure 33: Event Classification Table for 3 Branch Decision Tree with 46 leaves

Based on the figure above, represent the event classification table for the 3 Branch Decision Tree with 46 leaves. The Credit Card Loyalty Program dataset contains 10127 rows of data where 70% or 7088 rows of data is used to train the model, and 30% or 3039 rows of data is used to validate the model generated.

Based on the validation set, there were 2473 true positives where the customer will continue to use the bank's service, while there were 419 true negatives where the customer will defer to other banks. There were also 69 false positives, where the model has predicted the customer will stay with the banks but actually they will defer to other banks. This misclassification causes the banks a lot of money as banks try to give benefits and promotions to loyal customers to stay with the bank, but the customer ends up leaving for other banks. There were also 78 false negatives, where the model has predicted the customer will defer to other banks, but actually, the customer will stay with the banks. This misclassification does not cost the company a lot of money as the company only lost the opportunity for a loyal customer.

7.3.3. Precision and Accuracy

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Figure 34: Format to calculate the accuracy (Erika D, 2019)

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Figure 35: Format to calculate the accuracy (Erika D, 2019)

Data Role	Accuracy	Precision
Training Set	95.02%	97.16%
Validation Set	95.16%	97.29%

Table 2: Accuracy and Precision for the 3 Branch Decision Tree with 46 leaves

The table above represents the accuracy and the precision for the 3 Branch Decision Tree with 46 leaves. As shown in the table it can be said that the accuracy is above 95% for both validation and training sets. While for the precision, it is above 97% precision for both validation and training set. It can be seen that this 3 Branch Decision Tree is the best in making predictions on who will be the most loyal customer to the bank and the least loyal customer to the bank based on both the accuracy and its precision percentages.

7.4. Critical Interpretation of Outcomes

The table below is the node rule generated from the best decision tree model, which is the 3 Branch Decision Tree with 46 leaves.

No.	Node Rules & Explanation
1.	*--* Node = 12 *--* if Total_Trans_Ct >= 54.5 or MISSING AND Total_Trans_Amt >= 11304 then Tree Node Identifier = 12 Number of Observations = 537 Predicted: Attrition_Flag=Existing Customer = 1.00 Predicted: Attrition_Flag=Attrited Customer = 0.00 If the total transaction count is more and equal than 54.5 and the total transaction amount is more and equal than 11302, then there is 100% that the customer will stay as a loyal customer with the bank and 0% possibility of customer deferring to other banks
2.	*--* Node = 27 *--* if Total_Trans_Ct < 63.5 AND Total_Trans_Ct >= 57.5 AND Total_Trans_Amt < 5417 or MISSING then Tree Node Identifier = 27 Number of Observations = 471 Predicted: Attrition_Flag=Existing Customer = 0.94 Predicted: Attrition_Flag=Attrited Customer = 0.06 If the Total Transaction Count is less than 63.5 and Total Transaction Count is more and equal than 57.5 and Total Transaction Amount less than 5417 or missing, then there is a 6% possibility that the customer will defer other banks and 94% possibility of staying as a loyal customer to the bank

<table>
<tr>
<td>3.</td>
<td>
<pre>
--
 Node = 28
--
if Total_Trans_Ct >= 63.5 or MISSING
AND Total_Trans_Amt < 5417 or MISSING
then
 Tree Node Identifier   = 28
 Number of Observations = 2946
 Predicted: Attrition_Flag=Existing Customer = 1.00
 Predicted: Attrition_Flag=Attrited Customer = 0.00
</pre>
If Total Transaction Count is more and equal than 63.5 or Missing and Total Transaction Amount is less than 5417 or Missing, then there is a 100% possibility that the customer will stay as a loyal customer with the bank and 0% chance that the customer will defer to other banks
</td>
</tr>
</table>

Table 3: Node Rules for 3 Branches Decision Tree with 46 leaves

8 Logistic Regression (Lee Han Sen TP059717)

8.1 Diagram flow for logistic regression

In order to perform logistic regression in SAS Enterprise Miner, regression node should be selected and drag to the workspace.

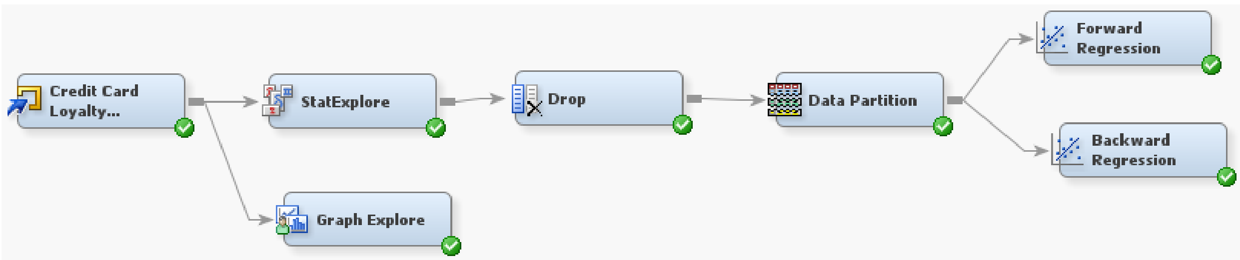


Figure 36: Diagram flow for logistic regression model

Figure above shows the diagram flow for the logistic regression. In this assignment, two logistic regressions will be done which is forward regression and backward regression.

8.2 Model construction and Optimization

8.2.1 Forward Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes
Selection Options	...

Figure 37: Settings for *forward logistic regression*

Figure above shows the settings that are required to do forward logistic regression. In order to create a forward logistic regression, the regression type needs to select logistic regression, selection model needs to be forward and selection criteria is set to validation misclassification as this assignment will be focusing on the misclassification rate of each logistic regression model.

Results for forward regression

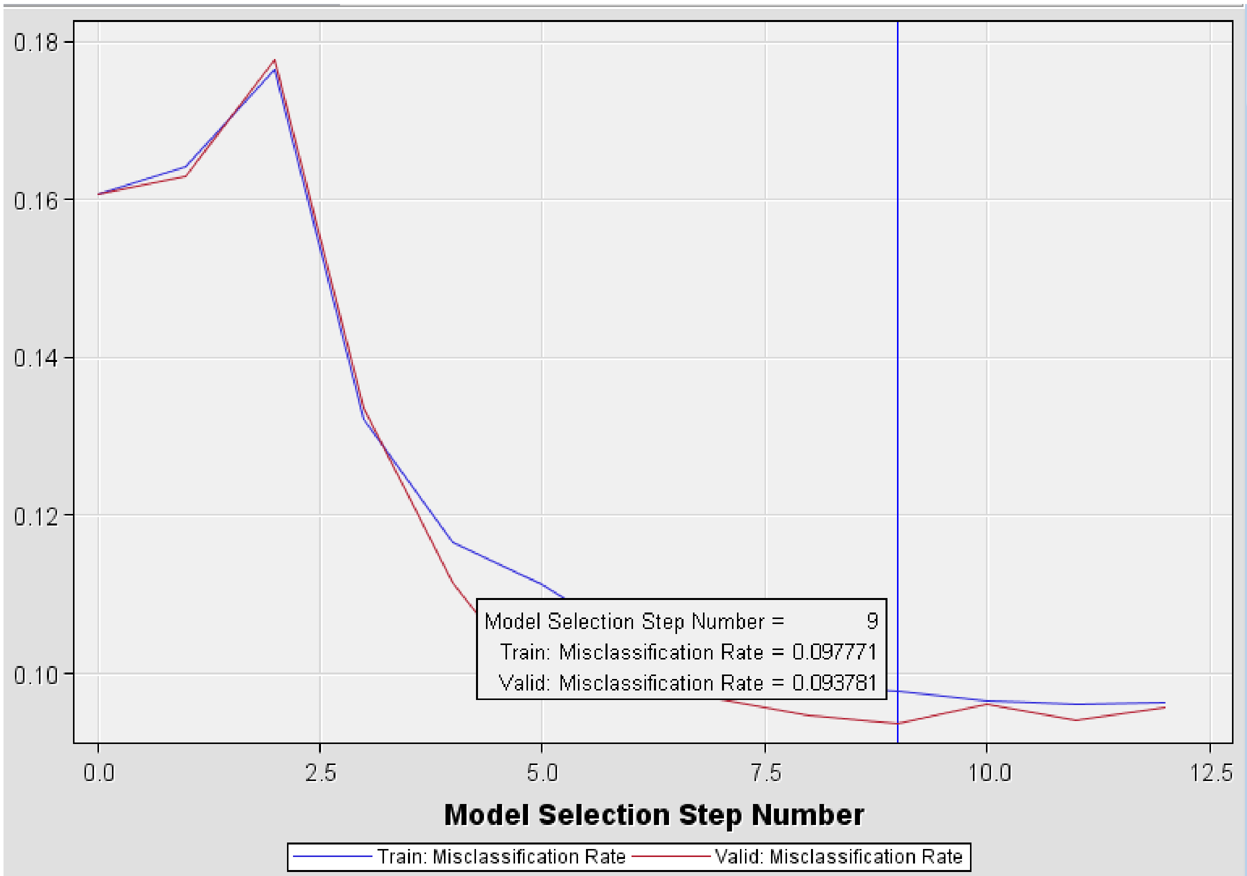
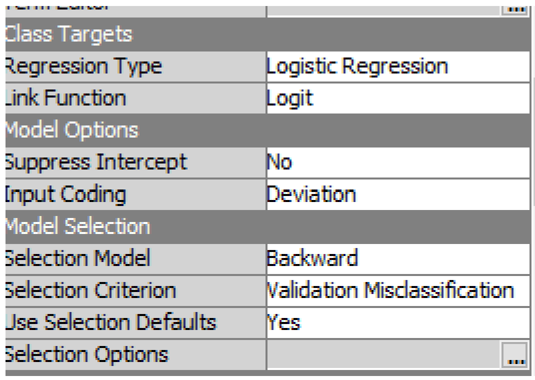


Figure 38: Results of misclassification rate for forward regression

Figure above shows the location of the lowest misclassification rate. At step 9, it indicates the lowest misclassification rate which means at this step, the error rate for both training and validation data set will be lowest. The misclassification rate will not be able to improve after the blue line (step 9). At step 9, the misclassification rate for training data set is 0.097771 or (9.7%) and for validation data set is 0.093781 (9.3%). Hence, the accuracy of model will be 90.63%.

8.2.2 Backward Regression



Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes
Selection Options	...

Figure 39 : Settings for backward logistic regression

Figure above shows the settings that are required to do backward logistic regression. In order to create a backward logistic regression, the regression type needs to select logistic regression, selection model needs to be backward and selection criteria is set to validation misclassification as this assignment will be focusing on the misclassification rate of each logistic regression model.

Results for backward regression

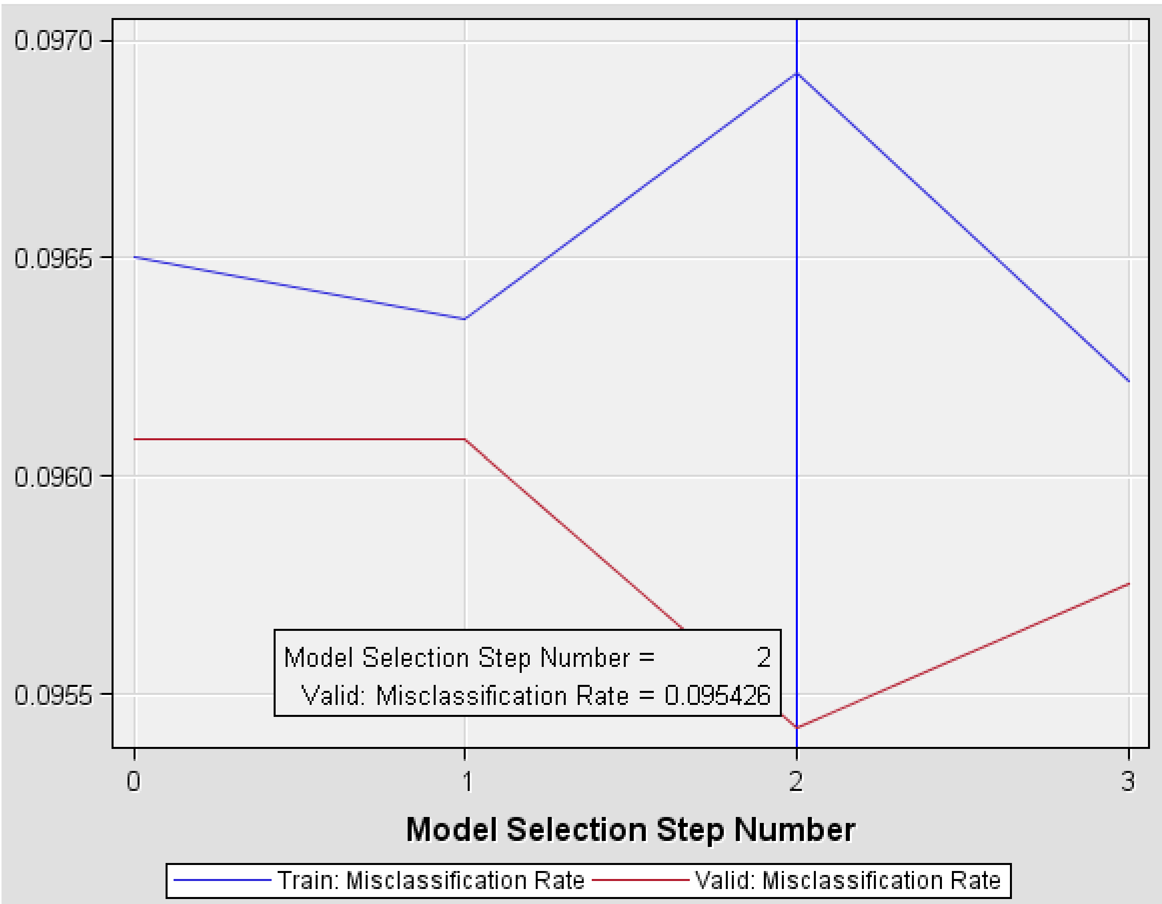


Figure 40: Results of misclassification rate for backward regression

Figure above shows the location of the lowest misclassification rate. At step 2, it indicates the lowest misclassification rate which means at this step, the error rate for both training and validation data set will be lowest. The misclassification rate will not be able to improve after the blue line (step 2). At step 2, the misclassification rate for training data set is 0.096924 or (9.6%) and for validation data set is 0.095426 (9.54%). Hence, the accuracy of the model will be 90.46%.

8.3 Model Validation

8.3.1 Summary of both forward and backward regression

Regression Method	Lowest misclassification rate step location	Validation Misclassification Rate	Model Accuracy
Forward Regression	9	0.093781 (9.37%)	90.63%
Backward Regression	2	0.095426 (9.54%)	90.46%

Table 4: Summary of both forward and backward regression

According to the table above, forward regression is better than backward regression as the misclassification rate is lower than backward regression.

8.3.2 Summary of Forward Regression

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	Total_Trans_Ct	1	1	938.7495	<.0001	0.1629
2	Total_Trans_Amt	1	2	534.9922	<.0001	0.1777
3	Total_Revolving_Bal	1	3	502.6028	<.0001	0.1336
4	Total_Ct_Chng_Q4_Q1	1	4	268.6361	<.0001	0.1115
5	Total_Relationship_Count	1	5	193.1674	<.0001	0.0971
6	Contacts_Count_12_mon	1	6	146.5858	<.0001	0.0981
7	Months_Inactive_12_mon	1	7	133.0666	<.0001	0.0967
8	Gender	1	8	60.3215	<.0001	0.0948
9	CLIENTNUM	1	9	17.4188	<.0001	0.0938
10	Customer_Age	1	10	9.7915	0.0018	0.0961
11	Income_Category	5	11	15.1915	0.0096	0.0941
12	Total_Amt_Chng_Q4_Q1	1	12	4.4723	0.0344	0.0958

Figure 41: Summary of *Forward Regression*

Figure above shows the summary of forward regression. As the variables enter the model, the validation misclassification rate gets reduced. After 12 important variables enters the model, the misclassification rate will be reduce to the lowest which is 0.0958 (9.5%). The variables that are needed to reduced the misclassification rate are “CLIENTNUM”, “Contacts_Count_12_mon”, “Gender   Months_Inactive_12_mon”, “Total_Ct_Chng_Q4_Q1”, “Total_Relationship_Count”, “Total_Revolving_Bal”, “Total_Trans_Amt” and “Total_Trans_Ct”.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
CLIENTNUM	1	17.3046	<.0001
Contacts_Count_12_mon	1	143.2061	<.0001
Gender	1	57.6076	<.0001
Months_Inactive_12_mon	1	130.7479	<.0001
Total_Ct_Chng_Q4_Q1	1	183.5258	<.0001
Total_Relationship_Count	1	178.0050	<.0001
Total_Revolving_Bal	1	314.0800	<.0001
Total_Trans_Amt	1	327.3201	<.0001
Total_Trans_Ct	1	691.7827	<.0001

Figure 42: Analysis of Effects

Figure above shows the analysis of effects. The higher the wald chi-square, the more important the variable. Hence, the two most important variable that forward regression should consider are “Total_Trans_Ct”, having 691.7828 Wald Chi-Square value and “Total_Trans_Amt”, having 327.3201 Wald Chi-Square value.

Odds Ratio Estimates

Effect		Attrition_Flag	Point Estimate
CLIENTNUM		Existing Customer	1.000
Contacts_Count_12_mon		Existing Customer	0.602
Gender	F vs M	Existing Customer	0.505
Months_Inactive_12_mon		Existing Customer	0.598
Total_Ct_Chng_Q4_Q1		Existing Customer	17.548
Total_Relationship_Count		Existing Customer	1.535
Total_Revolving_Bal		Existing Customer	1.001
Total_Trans_Amt		Existing Customer	1.000
Total_Trans_Ct		Existing Customer	1.115

Figure 43: *Odds Ratio Estimate*

Figure above shows the odds ratio estimate. Odds ratio estimate reveals the strength of association between two events. According to this odds ratio estimate, between gender male and female, it shows that female is 50.5% more than male to having the bank credit cards.

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

False Negative	True Negative	False Positive	True Positive
79	282	206	2472

Figure 44: Event Classification Table

Figure above shows the event classification table of forward regression. This dataset contains 10127 rows of data where 70% (7088) of the data are used for training, and the rest of 30% (3039) are for validation of model.

Based on the validation partition, 2472 true positive were detected correctly where the customer will still continue using the bank's credit card service. 282 true negatives were also detected correctly where the customer stops the terminates their bank credit card service. On the other hand, there are 206 false positive which means that the model predicted that the customer will still continue using the bank's credit card service while the actual truth is the customer terminates their bank's credit card service. False negative do also appear, having 79 records of it. False negative means by the model predicts the customer will terminates their bank's credit card service but the truth is the customer will still continue using the bank's credit card service. This forward regression predicts more false positive than more false negative, hence it might causes the bank giving unnecessary promotion to the customer who terminates the bank's credit card service. By summing all statistics up, this forward regression still achieve a 82% of accuracy.

8.4 **Sensitivity, Specificity, False Positive Rate, Precision and Accuracy**

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

Figure 45: *Format to calculate sensitivity (Kakanadan, 2020)*

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

Figure 46: F*ormat to calculate precision (Kakanadan, 2020)*

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

Figure 47: Format to calculate accuracy *(Kakanadan, 2020)*

Data Role	**Sensitivity**	**Precision**	**Accuracy**
Training Set	96.60%	92.12%	90.22%
Validation Set	96.90%	92.30%	90.62%

Table 5: Sensitivi*ty, Specificity, False Positive Rate, Precision and Accuracy for Forward Regression*

The table above shows the sensitivity, specificity, precision, and accuracy of forward regression model. Sensitivity rate shows the correctly predicted positive event among all positive events. This model achieves a 96.90% in validation data set, meaning that there are 96.90% of correctly predicted positive event among all positive events. Precision rate shows the ratio of true positive event among all predicted positive events. This model achieves a 92.30% in validation data set, meaning that there are 92.30% of correctly predicted positive event among all predicted positive events. Accuracy rate shows how often the model predicted correctly among all cases. This model achieves a 90.62% in validation data set, meaning that there are 90.62% of correctly predicted event among all events. Hence, this forward regression model considered as a good model as most of the prediction is up to 90% correct.

9 Neural Network (Lau Zhi Yi TP059579)

9.1 Diagram flow for Neural Network

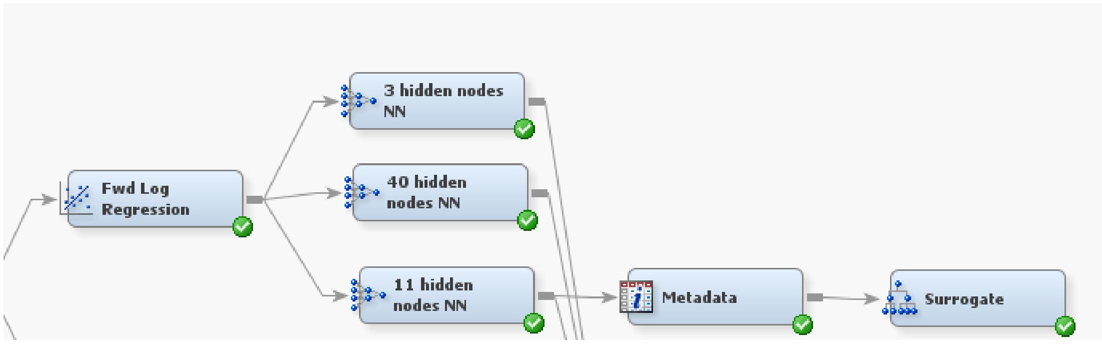


Figure 48: Diagram flow for Neural Network

Based on the figure above shows that Forward Log Regression is used to make Neural Network models. There are 3 types of Neural Network models, which are 3 Hidden Nodes Neural Network model, 11 Hidden Nodes Neural Network model and 40 Hidden Nodes Neural Network model. In addition, 11 Hidden Nodes Neural Network model will do the Surrogate Tree.

9.2 Model construction and Optimization

9.2.1 3 Hidden Nodes Neural Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

Figure 49: Network Settings for 3 Hidden Nodes Neural Network

Property	Value
Training Technique	Default
Maximum Iterations	1000
Maximum Time	5 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1

Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Figure 50: Optimization Settings for 3 Hidden Nodes Neural Network

The figures above show the settings required to be applied to generate a 3 Hidden Nodes Neural Network. Since this neural network has 3 hidden nodes, the number of hidden units will be set to 3. The number of maximum iterations is set to 1000 due to the maximum number of iterations is 1000. The minimum time in the value of maximum time is 5 minutes. Because we need to develop the neural network model as fast as possible, we put the shortest time in the maximum time section. Last, the preliminary training is set to unable.

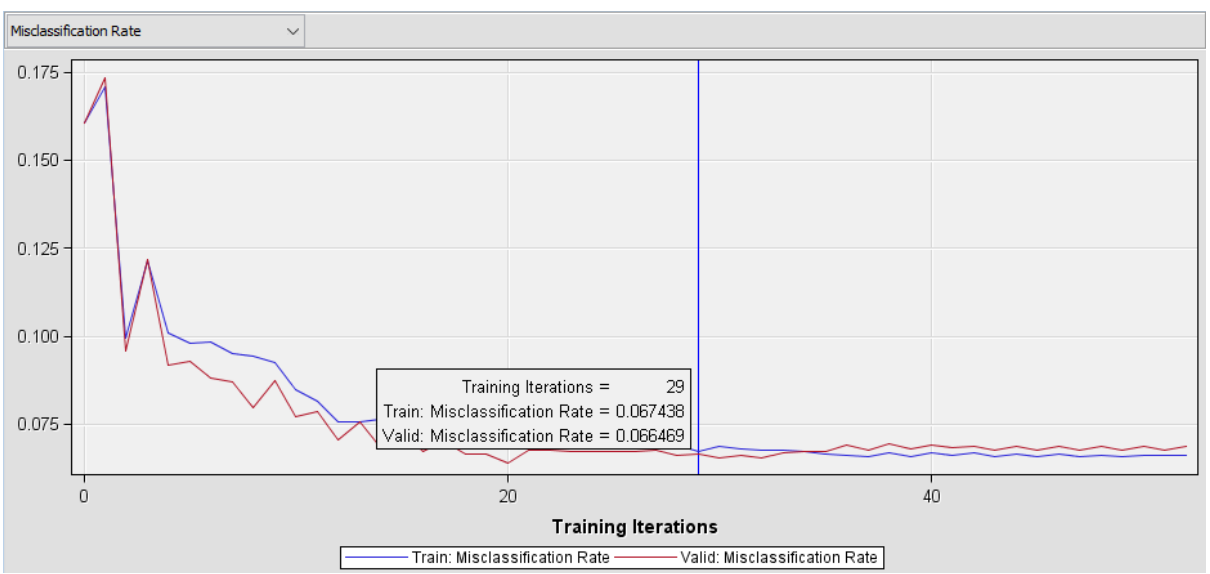


Figure 51: Misclassification rate of 3 Hidden Nodes Neural Network

Based on the figure above shows the straight blue line that indicates the location of the lowest misclassification rate. This is important because it shows that if the number of training iterations increases after the straight blue line, the misclassification rate will not improve. The lowest misclassification rate of the 3 Hidden Nodes Neural Network is 0.066469 or (6.66%). The number of training iterations with this lowest misclassification rate is 29. In addition, the lines are neither overfitted nor underfitted.

9.2.2 11 Hidden Nodes Neural Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	11
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Figure 52: Network Settings for 11 Hidden Nodes Neural Network

Property	Value
Training Technique	Default
Maximum Iterations	1000
Maximum Time	5 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Figure 53: Optimization Settings for 11 Hidden Nodes Neural Network

The figures above show the settings required to be applied to generate an 11 Hidden Nodes Neural Network. Since this neural network has 11 hidden nodes, the number of hidden units will be set to 11. The number of maximum iterations is set to 1000 due to the maximum number of iterations is 1000. The minimum time in the value of maximum time is 5 minutes. Because we need to develop the neural network model as fast as possible, we put the shortest time in the maximum time section. Last, the preliminary training is set to unable.

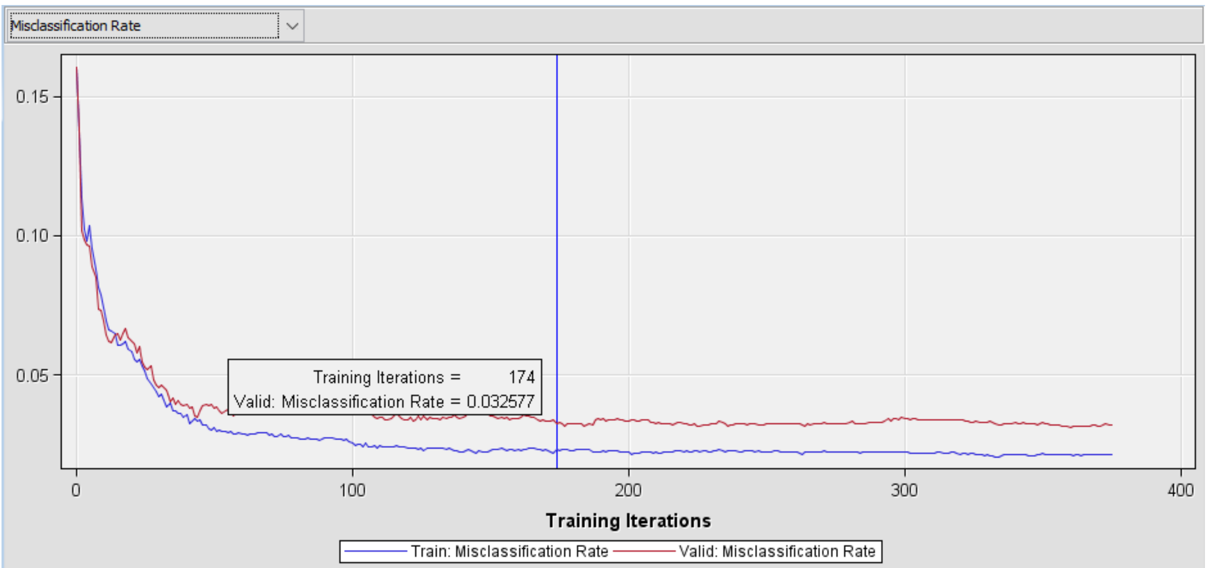


Figure 54: Misclassification rate of 11 Hidden Nodes Neural Network

Based on the figure above shows the straight blue line that indicates the location of the lowest misclassification rate. This is important because it shows that if the number of training iterations increases after the straight blue line, the misclassification rate will not improve. The lowest misclassification rate of the 11 Hidden Nodes Neural Network is 0.032577 or (3.26%). The number of training iterations with this lowest misclassification rate is 174. In addition, the lines are neither overfitted nor underfitted.

9.2.3 40 Hidden Nodes Neural Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	40
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default

Figure 55: Network Settings for 40 Hidden Nodes Neural Network

Property	Value
Training Technique	Default
Maximum Iterations	1000
Maximum Time	5 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Figure 56: Optimization Settings for 40 Hidden Nodes Neural Network

The figures above show the settings required to be applied to generate a 40 Hidden Nodes Neural Network. Since this neural network has 40 hidden nodes, the number of hidden units will be set to 40. The number of maximum iterations is set to 1000 due to the maximum number of iterations is 1000. The minimum time in the value of maximum time is 5 minutes. Because we need to develop the neural network model as fast as possible, we put the shortest time in the maximum time section. Last, the preliminary training is set to unable.

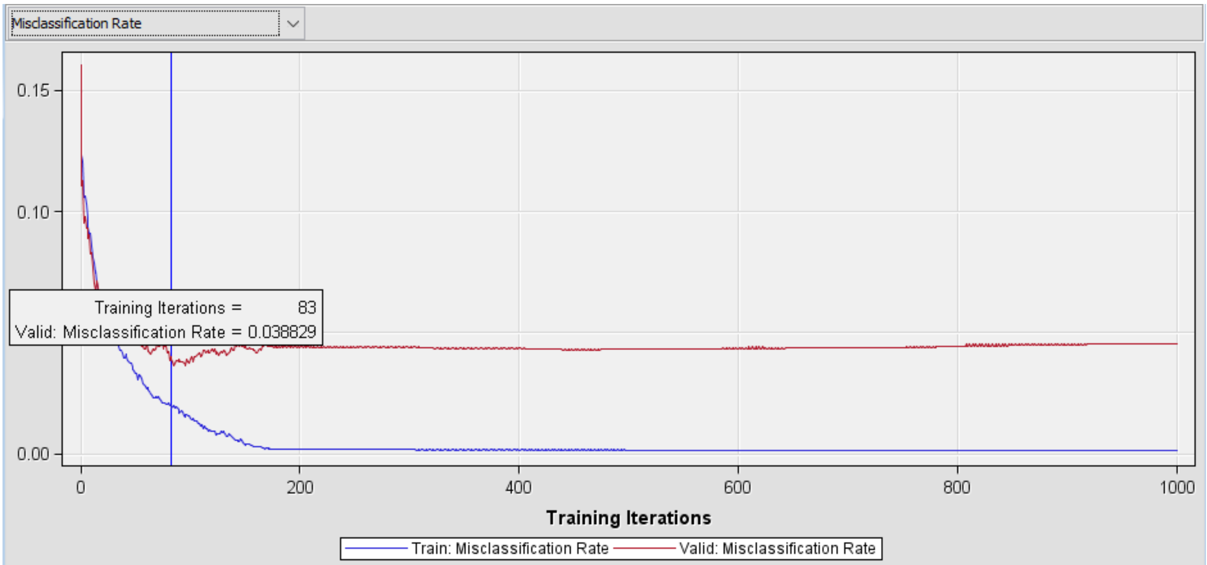


Figure 57: Misclassification rate of 40 Hidden Nodes Neural Network

Based on the figure above shows the straight blue line that indicates the location of the lowest misclassification rate. This is important because it shows that if the number of training iterations increases after the straight blue line, the misclassification rate will not improve. The lowest misclassification rate of the 40 Hidden Layer Neural Network is 0.038829 or (3.88%). The number of training iterations with this lowest misclassification rate is 83. In addition, the lines are neither overfitted nor underfitted.

9.3 Model Validation

9.3.1 Summary of all Neural Network Model

Neural Network Model	Hidden Nodes	Validation Misclassification Rate
3 Hidden Nodes Neural Network	3	0.066469
11 Hidden Nodes Neural Network	11	0.032577
40 Hidden Nodes Neural Network	40	0.038829

Table 6: Summary of all Neural Network Model

Based on the table above, the best Neural Network Model is the 11 Hidden Nodes Neural Networks, which have the lowest classification rate of 0.032577.

9.3.2 Event Classification Table

```
Event Classification Table

Data Role=TRAIN Target=Attrition_Flag Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

   73         1049         90         5876


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

  False       True        False       True
Negative    Negative    Positive    Positive

   54          443         45         2497
```

Figure 58: Event Classification Table of 11 Hidden Nodes Neural Network

Based on the figure above shows the event classification table of 11 Hidden Nodes Neural Network. The Credit Card Loyalty dataset contains 10127 rows of data, which have been partitioned into 70% (7088 rows) is used to train the model, and 30% (3039 rows) is used to validate the model. According to the verification set, the number of true positives is 2497, indicating that the customer will continue to use the bank credit card service. The number of true negatives is 443, which means that the customer will not utilize the bank credit card service. The number of false positives is 45, suggesting that the model anticipated that the customer would continue to use the bank credit card service, but they will not. This misclassification causes the banks to lose a lot of money in providing the benefits and promotions to the customer, but they would switch to using other banks' credit card services. The number of false negatives is 54, indicating that the model predicted that the customer would not use the bank credit card service, but they did. This misclassification causes the bank to have a high possibility of losing those loyal customers.

9.3.3 Accuracy, Precision, Recall, and F1 score

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP}
\end{aligned}
$$

Figure 59: Formula of Accuracy, Precision, Recall, and F1 score *(Gad, 2020)*

Based on the figure above, the formulas are used to calculate the accuracy, accuracy, recall, and F1 score of the 11 hidden node neural network models.

Data Role	**Accuracy**	**Precision**	**Recall**	**F1 score**
Training Set	97.70%	84.85%	98.77%	91.28%
Validation Set	96.74%	84.93%	97.88%	90.95%

Table 7: Accuracy, Precision, Recall and F1 score for Training and Validation Set

Based on the table above, the accuracy rate of the validation set is 96.74%, indicating that the 11 Hidden Nodes Neural Network model has high accuracy in predicting customer decisions. The recall rate of the verification set is 97.88%, which shows that the model has high accuracy in predicting customers who are loyal to bank credit card services. However, the precision rate (84.93%) is low compared to the accuracy and recall rates. This means that the model is less accurate in predicting customers who are not loyal to bank credit card services. Last, the F1 score rate is 90.95%, which is a weighted average of precision and recall (Solutions, 2016).

9.4 Critical Interpretation of Outcomes

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report	Model
Attrition_Flag	N	**Default**	Target	**Rejected**	Binary	**Default**	**Default**	**Default**	Neural5
Avg_Open_To_B	N	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
Avg_Utilization_I	N	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
CLIENTNUM	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Card_Category	Y	**Default**	Rejected	**Default**	Nominal	**Default**	**Default**	**Default**	
Contacts_Count	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Credit_Limit	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Customer_Age	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Dependent_cour	Y	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
Education_Level	N	**Default**	Rejected	**Default**	Nominal	**Default**	**Default**	**Default**	
F_Attrition_Flag	N	**Default**	Classification	**Default**	Nominal	**Default**	**Default**	**Default**	
Gender	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**	
I_Attrition_Flag	N	**Default**	Classification	**Default**	Nominal	**Default**	**Default**	**Default**	
Income_Categor	N	**Default**	Rejected	**Default**	Nominal	**Default**	**Default**	**Default**	
Marital_Status	Y	**Default**	Rejected	**Default**	Nominal	**Default**	**Default**	**Default**	
Months_Inactive	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Months_on_bool	N	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
Naive_Bayes_Cl	Y	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
P_Attrition_Flag	N	**Default**	Prediction	**Default**	Interval	**Default**	**Default**	**Default**	
P_Attrition_Flag	N	**Default**	Prediction	**Default**	Interval	**Default**	**Default**	**Default**	
R_Attrition_Flag	N	**Default**	Residual	**Default**	Interval	**Default**	**Default**	**Default**	
R_Attrition_Flag	N	**Default**	Residual	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Amt_Chng	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Ct_Chng_	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Relationsh	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Revolving	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Trans_Am	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
Total_Trans_Ct	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**	
U_Attrition_Flag	N	**Default**	Classification	**Target**	Nominal	**Default**	**Default**	**Default**	
VAR23	Y	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**	
WARN	N	**Default**	Assessment	**Default**	Nominal	**Default**	**Default**	**Default**	
dataobs	N	**Default**	ID	**Default**	Interval	**Default**	**Default**	**Default**	

Figure 60: Settings for Metadata

Based on the figure above shows the new role of Attrition_Flag has changed to Rejected, and the new role of U_Attrition_Flag has changed to Target.

The table below is node rules generated from 11 Hidden Nodes Neural Network model:

<table>
<tr><th>No.</th><th>Node Rules & Explanation</th></tr>
<tr><td>1.</td><td><pre>
--
 Node = 13
--
if Total_Trans_Ct >= 57.5 or MISSING
AND Total_Trans_Amt < 5417 or MISSING
then
 Tree Node Identifier   = 13
 Number of Observations = 3417
 Predicted: U_Attrition_Flag=Existing Customer = 0.99
 Predicted: U_Attrition_Flag=Attrited Customer = 0.01
</pre>
In this observation, the number of customers was 3417. If the total number of transactions on the credit card is greater than or equal to 57.5 and the total transaction amount on the credit card is less than 5417, 99% of customers are loyal customers.</td></tr>
<tr><td>2.</td><td><pre>
--
 Node = 43
--
if Total_Trans_Ct < 55.5
AND Total_Trans_Amt < 2058.5 AND Total_Trans_Amt >= 924.5 or MISSING
AND Total_Revolving_Bal >= 626 or MISSING
AND Total_Relationship_Count >= 2.5 or MISSING
then
 Tree Node Identifier   = 43
 Number of Observations = 1092
 Predicted: U_Attrition_Flag=Existing Customer = 0.96
 Predicted: U_Attrition_Flag=Attrited Customer = 0.04
</pre>
In this observation, the number of customers was 1092. If the total number of transactions on the credit card is less than 55.5, the total transaction amount is greater than 924.5 and less than 2058.5, the total revolving balance on credit card is greater and equal than 626, and the total relationship count is greater and equal than 2.5, 96% of customers are loyal customers.</td></tr>
</table>

Table 8: Node Rules for 11 Hidden Nodes Neural Network model

10 Conclusion

10.1 Diagram flow for Model Comparison

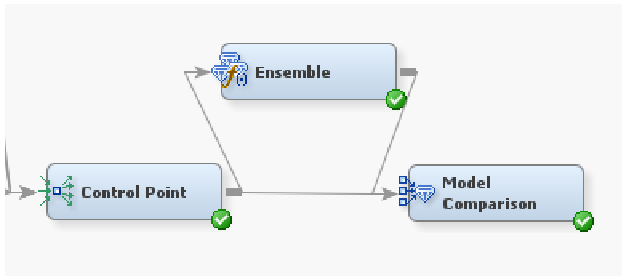


Figure 61: Model Comparison

The figure above shows the process of comparing models to get the best model. To make it easier for later nodes to connect to the models, all the models are connected to the Control Point node at the start of the process. Control Point node is connected to Ensemble node and Model Comparison node. Ensemble node will be used for creating the Ensemble model. In the property of the Ensemble node, the posterior probabilities in-class target section is set to ‘voting’, which means that the Ensemble node will take the majority target variable as a prediction result. Model Comparison node will be used for comparing the performance of all models including the Ensemble model.

10.2 Result of Model Comparison

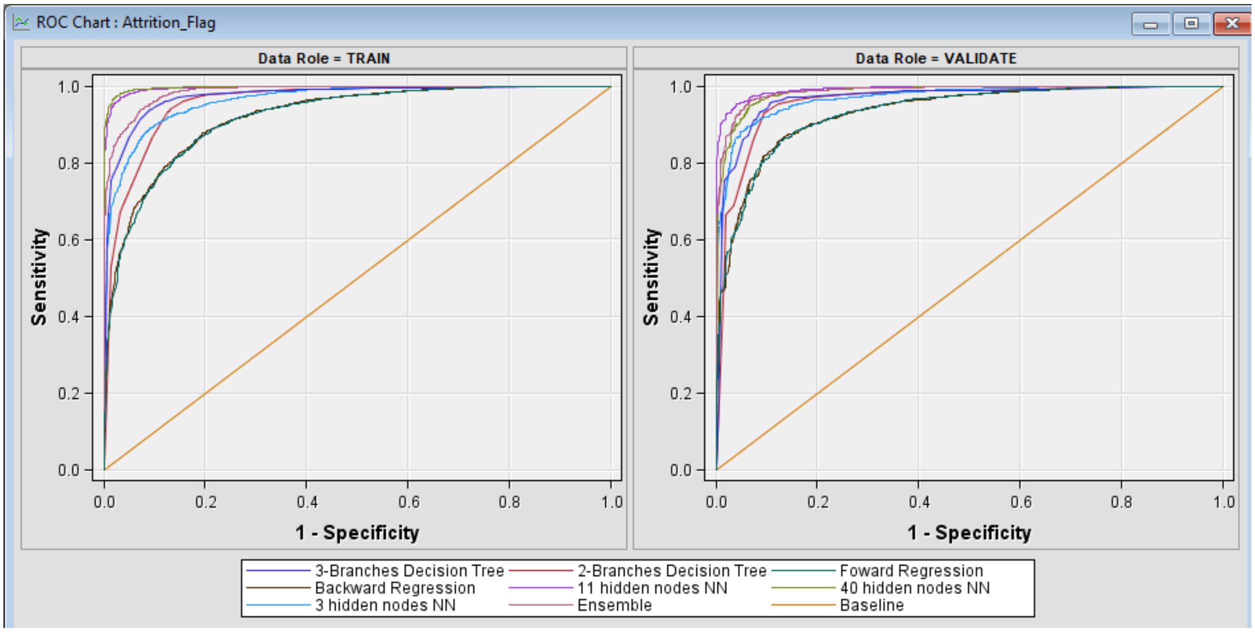


Figure 62: ROC charts

The ROC charts generated by Model Comparison are shown in the figure above. The ROC chart is used to compare the performance of each model. The curve closest to the upper left corner is the best model. According to the above figure, the purple curve is most relative to the upper left corner, indicating that the model representing the purple curve is the best model, which is 11 Hidden Nodes Neural Network model.

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Neural5	Neural5	11 hidden nodes NN	Attrition_Flag		0.032577	1193.787	0.017636
Neural2	Neural2	40 hidden nodes NN	Attrition_Flag		0.038829	1867.478	0.014871
Ensmbl	Ensmbl	Ensemble	Attrition_Flag		0.039816	.	0.030071
Tree3	Tree3	3-Branches Decision Tree	Attrition_Flag		0.048371	.	0.039075
Tree	Tree	2-Branches Decision Tree	Attrition_Flag		0.054952	.	0.043462
Neural	Neural	3 hidden nodes NN	Attrition_Flag		0.066469	2470.968	0.049906
Reg3	Reg3	Foward Regression	Attrition_Flag		0.093781	3468.371	0.072842
Reg2	Reg2	Backward Regression	Attrition_Flag		0.095426	3457.211	0.072149

Figure 63: Fit Statistics

The Fit Statistics generated by Model Comparison are shown in the figure above. The Fit Statistics can be used to compare the misclassification rate of each model. According to the above figure, the lowest misclassification rate among these models is 0.032577 or (3.26%), where the model is 11 Hidden Nodes Neural Network. Hence, based on the result of the ROC chart and the Fit Statistics, the best classification model for predicting loyal credit card customers is the 11 Hidden Node Neural Network Model.

References

Chauhan, N. S. (n.d.). *Decision Tree Algorithm, Explained*. Retrieved from Kdnuggets: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Credit Card Statistics. (2021, August). Retrieved from shiftprocessing: https://shiftprocessing.com/credit-card/

Deepa, M., & D., A. (2019, December). *Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques*. Retrieved from Research Gate: https://www.researchgate.net/publication/339974302_Survey_Paper_for_Credit_Card_Fraud_Detection_Using_Data_Mining_Techniques

Exploratory Data Analysis. (2021, September 28). Retrieved from IBM: https://www.ibm.com/cloud/learn/exploratory-data-analysis

Gad, I. m. (2020, July 14). *What is the best metric (precision, recall, f1, and accuracy) to evaluate the machine learning model for imbalanced data?* Retrieved from ResearchGate: https://www.researchgate.net/post/What_is_the_best_metric_precision_recall_f1_and_accuracy_to_evaluate_the_machine_learning_model_for_imbalanced_data

Kakanadan, U. (2020, March 14). *Analytics Vidhya*. Retrieved from Regression and performance metrics — Accuracy, precision, RMSE and what not!: https://medium.com/analytics-vidhya/regression-and-performance-metrics-accuracy-precision-rmse-and-what-not-223348cfcafe

SAS. (2017, August 30). *SAS Help Center*. Retrieved from Introduction to SEMMA: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

Solutions, E. (2016, September 9). *Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures*. Retrieved from EXSILIO: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

StatExplore Node. (2017, August 30). Retrieved from SAS® Help Center: https://documentation.sas.com/doc/en/emref/14.3/p11763vwqrp4f3n1cslg56nfbgv3.htm

Team, T. R. (2021, November 14). *Introduction to credit cards*. Retrieved from Ringgit Plus: https://ringgitplus.com/en/blog/personal-finance-news/introduction-to-credit-cards.html

Erika D. (2019, December 9). *Accuracy, Recall & Precision. When it comes to evaluating how well a… | by Erika D | Medium*. Medium. https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d