Table of Contents

Bank Customers Churn Prediction

1. Introduction

Financial institutions are finding hard to retain their customers due to the competitive environment in the banking industry. Due to the fierce competition, banks are constantly coming with innovative and attractive financial products with the intention of attracting new customers and at the same time retaining their existing customer base. Meanwhile, bank customers have plenty of options and would choose the bank that can offer the best product and services. For this reason, banks are constantly finding ways to identify and prevent customer churn before it happens. By implementing the use of analytics, banks can uncover hidden trends and patterns that could lead to customer churn. As a result, banks can reap many long-term benefits which include higher earnings, larger customer base, stronger branding and many more.

2. Problem Statement

Expensive to capture new market share

Customer churn presents a huge problem for financial institutions as it is much more expensive to capture the market share of new customers compared to retaining existing ones. According to Wertz (2018), acquiring new customers are 5 times more costly than retaining existing customers. Hence, it is advantageous for banks to foresee which client is inclined towards churning so that preventive measures can be taken early to minimize the probability of the client churning.

3. Aim

To develop a prediction model which can identify banking customers who are likely to churn

4. Objectives

a) To identify the characteristics of churned customers

b) To explore the relationship between credit score, estimated salary and bank balance

c) To build a prediction model in SAS Enterprise Miner based on the prominent features to predict bank customers churn

d) To recommend solutions for banks to retain their existing customers

5. Methodology

The methodology used in this assignment would be the SEMMA model. The SEMMA model consists of six sequential steps, namely sample, explore, modify, model and assess. As shown in figure 1 below is the flow of the SEMMA data mining method.

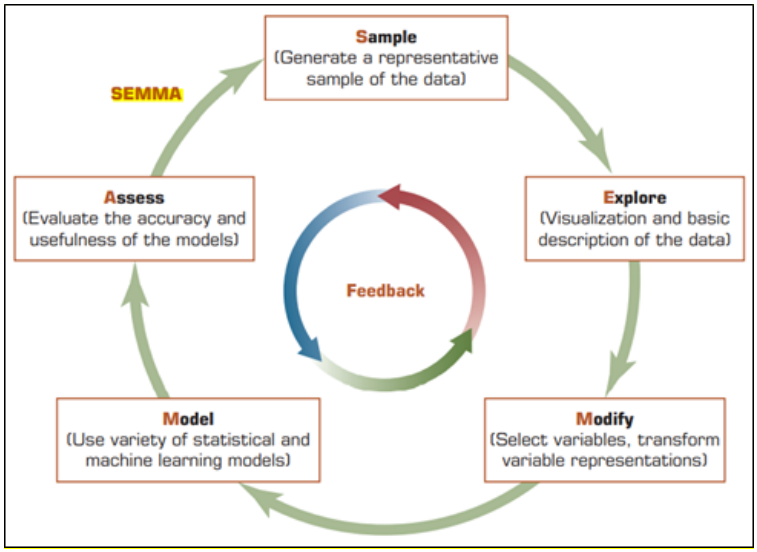


Figure 1 : SEMMA Model Diagram (Sharda et al., 2018)

a) Sample

The data is split into training and testing set with a ratio of 70% for the training set and 30% for the validation set. Due to the imbalance target variable in the dataset, oversampling is used to handle the class imbalance problem in later stages.

b) Explore

A comprehensive data exploratory analysis is conducted with the help of Tableau to uncover the trends in the dataset. Also, the summary statistics and correlation statistics of each variables will be analysed at this stage using SAS Enterprise Miner.

c) Modify

The data is modified by transforming the variables before going into model selection. Some variables with outliers, for example the “Age” variable would need to be transformed in order to eliminate or minimize the outliers.

d) Model

Decision tree and stepwise logistic regression would be used to predict the target variable as it has a binary outcome.

e) Assess

The models created are evaluated using metrics like ROC/AUC curve, mean square error and many more. If the model is valid, the accuracy will be high when it is being fed with the training and validation data. In the end, the model should be able to predict whether the customer has churned or is still with the bank.

6. Dataset Preparation & EDA

6.1 Dataset Metadata

The dataset contains 10,000 individual bank customers across three different countries and their respective financial information. The dataset is obtained from Kaggle which can be accessed through this link https://www.kaggle.com/datasets/santoshd3/bank-customers.

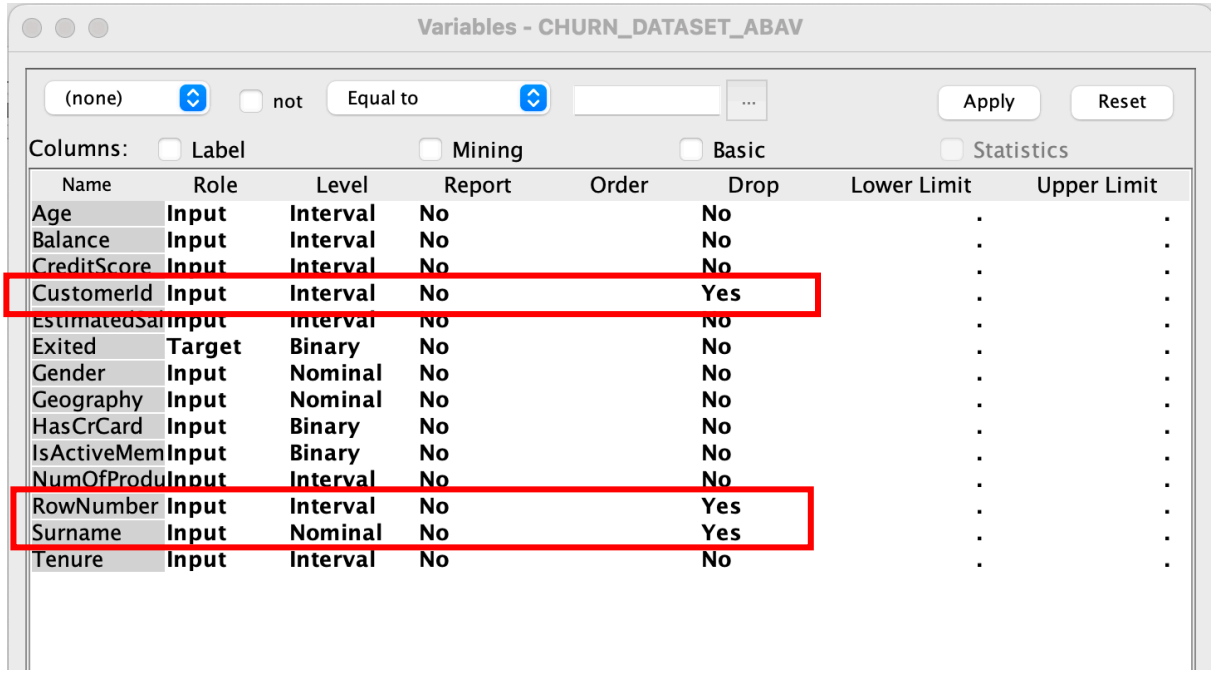
There are a total of 14 attributes and 10,000 records in the dataset. However, there are three attributes namely, "RowNumber", "Surname" and "CustomerID" which will be excluded from the dataset as these attributes do not contribute any meaningful interpretation to the model. As shown in table 1 below is the list of attributes in the dataset.

No.	Variables	Description	Data Type
1	~~RowNumber~~ (Removed during model building)	The record number of each customer	Nominal
2	~~CustomerId~~ (Removed during model building)	Random values assigned to each customer	ID
3	~~Surname~~ (Removed during model building)	Surname of a customer	Nominal
4	CreditScore	The customer's location	Interval
5	Gender	Gender of the customer	Categorical
6	Age	Age of the customer	Interval

7	Tenure	Number of years that the customer has been a client of the bank.	Nominal
8	Balance	The bank balance of the customer	Interval
9	NumOfProducts	Number of products that the customer has purchased from the bank	Nominal
10	HasCrCard	Customer has a credit card. (0 = No, 1 = Yes)	Binary
11	IsActiveMember	Active customers (0 = No, 1 = Yes)	Binary
12	EstimatedSalary	Customer's estimated salary	Interval
13	Exited (Target Variable)	Whether or not the customer left the bank. (0=No,1=Yes)	Binary

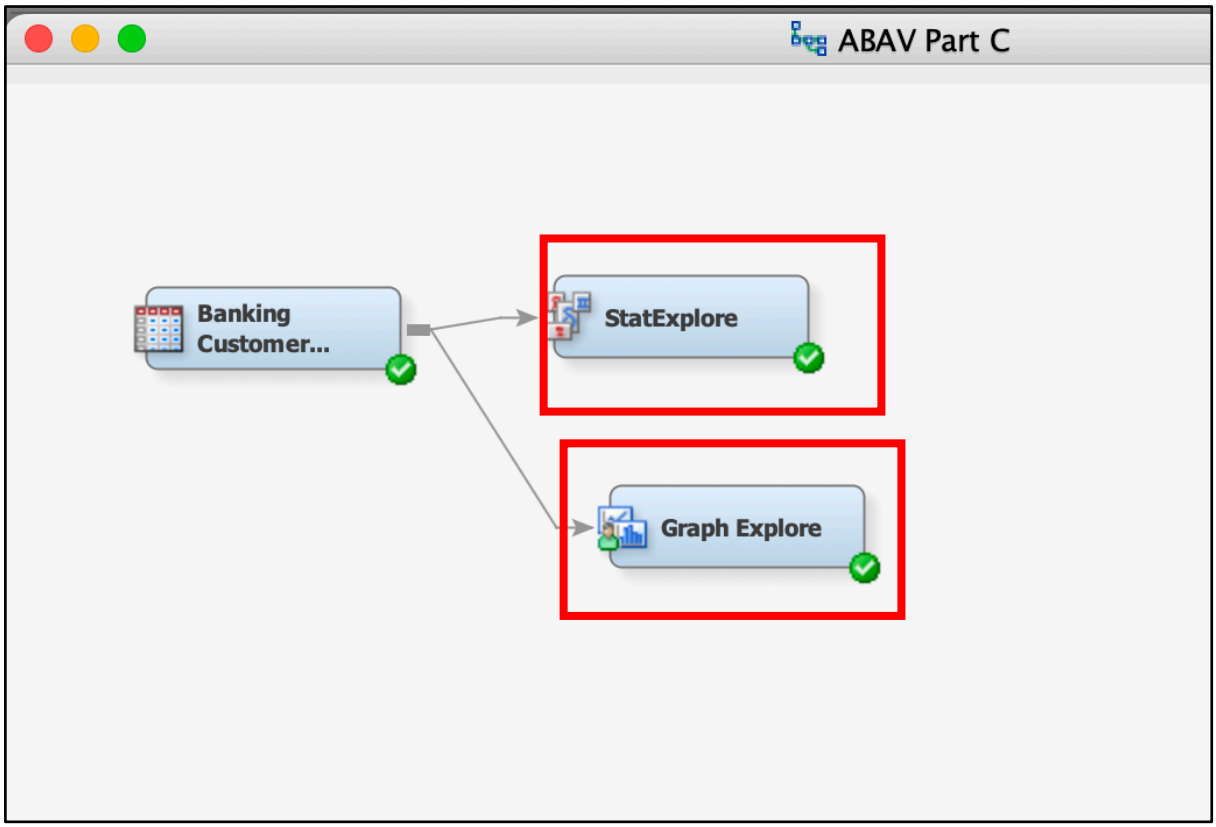
Table 1: Description of variables in the dataset

6.2 Exploratory Data Analysis



Before performing data exploration on the dataset, the variables "RowNumber", "Surname" and "CustomerID" are dropped.

6.2.1 Nodes Used During Initial Data Exploration



During the initial data exploration, both StatExplore and Graph Explore nodes are used.

6.2.2 Summary Statistics from StatExplore

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	Gender	INPUT	3	8	Male	54.53	Female	45.39
TRAIN	Geography	INPUT	3	0	France	50.14	Germany	25.09
TRAIN	HasCrCard	INPUT	2	0	1	70.55	0	29.45
TRAIN	IsActiveMember	INPUT	2	0	1	51.51	0	48.49
TRAIN	NumOfProducts	INPUT	4	0	1	50.84	2	45.90
TRAIN	Tenure	INPUT	12	6	2	10.47	1	10.35
TRAIN	Exited	TARGET	2	0	0	79.63	1	20.37

Interval Variable Summary Statistics
(maximum 500 observations printed)

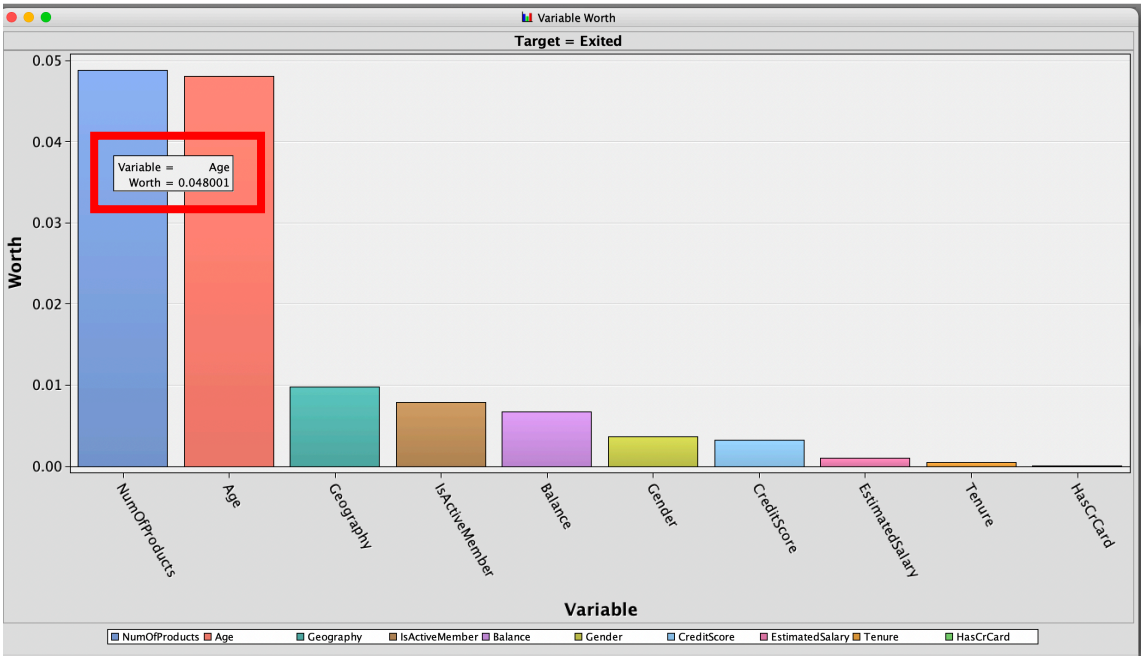
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	38.92155	10.49046	9994	6	18	37	92	1.011241	1.393612
Balance	INPUT	76485.89	62397.41	10000	0	0	97188.62	250898.1	-0.14111	-1.48941
CreditScore	INPUT	650.5288	96.6533	10000	0	350	652	850	-0.07161	-0.42573
EstimatedSalary	INPUT	100090.2	57510.49	10000	0	11.58	100187.4	199992.5	0.002085	-1.18152

Based on the class variable summary statistics and interval variable summary statistics generated in StatExplore, it is observed that there are eight missing values in “Gender”, six

missing values in “Tenure” and six missing values in “Age”. These missing values will be imputed accordingly during the data processing phase.

6.2.3 Variable Worth Chart



Besides that, the StatExplore node shows the top ten variable worth of the independent variables in predicting the target variable. In the figure above, it is known that the variable “NumofProducts” and “Age” are having very similar variable worth of about 0.048. Since the variable worth of these two variables are significantly higher compared to the rest of the variables, it can be said that “NumOfProducts” and “Age” have significant predictive power on the dependent variable (Exited) in this dataset. Also, it can be observed that the variable worth has reduced significantly starting “EstimatedSalary”.

6.2.4 Chi-Square Statistics

Output

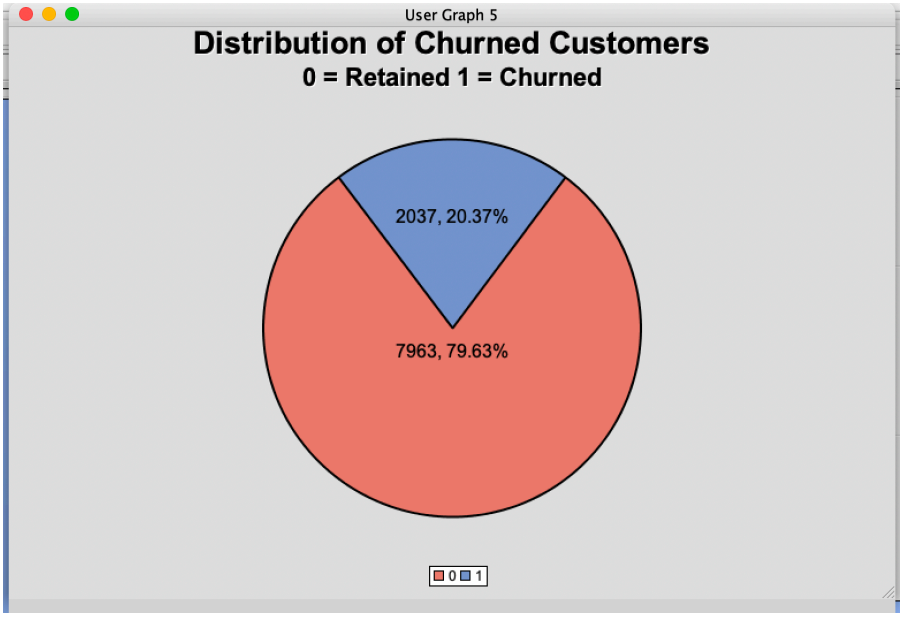
Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=Exited

Input	Chi-Square	Df	Prob
NumOfProducts	1503.6294	3	<.0001
Age	1177.7974	5	<.0001
Geography	301.2553	2	<.0001
IsActiveMember	243.7604	1	<.0001
Balance	185.7437	4	<.0001
Gender	113.3468	2	<.0001
CreditScore	21.1481	4	0.0003
Tenure	13.9475	11	0.2359
EstimatedSalary	2.4884	4	0.6467
HasCrCard	0.5095	1	0.4754

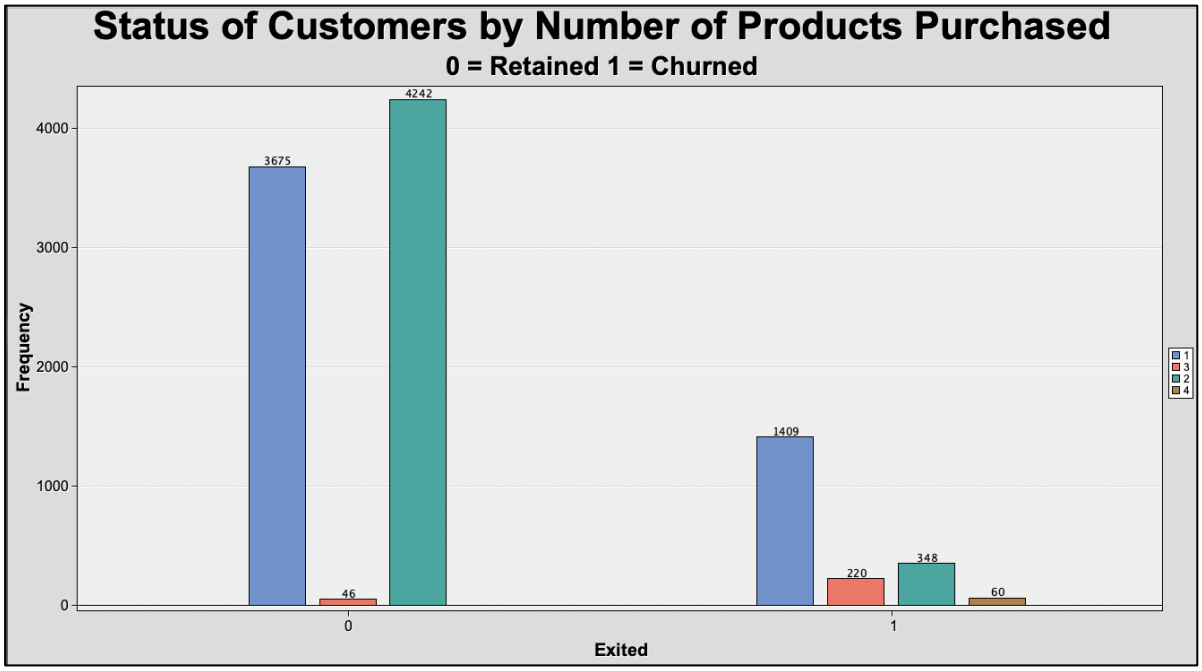
The chi-square statistics table is being referred to determine which variables are to be used during our model creation. It is found that the first 7 variables are having p-values of less than 0.05 each, which indicates that they are significant. Hence, "NumofProducts", "Age", "Geography", "IsActiveMember", "Balance", "Gender" and "CreditScore" are selected when building the model because they are significant predictors.

6.2.5 Distribution of the target variable (Exited)



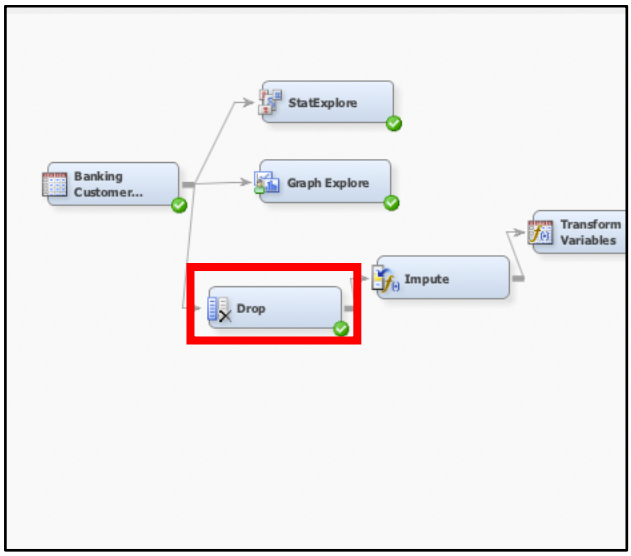
Based on the pie chart above, it is discovered that 2,037 customers (20.37%) have churned whereas 7,963 bank customers (79.63%) are still with the bank. The number of customers who have chosen to stayed with the bank is skewed to one side, which indicates that there is a class imbalance problem. This will cause the prediction model to be biased towards predicting customers who did not churn. Hence, this class imbalance issue should be dealt with during the pre-processing phase.

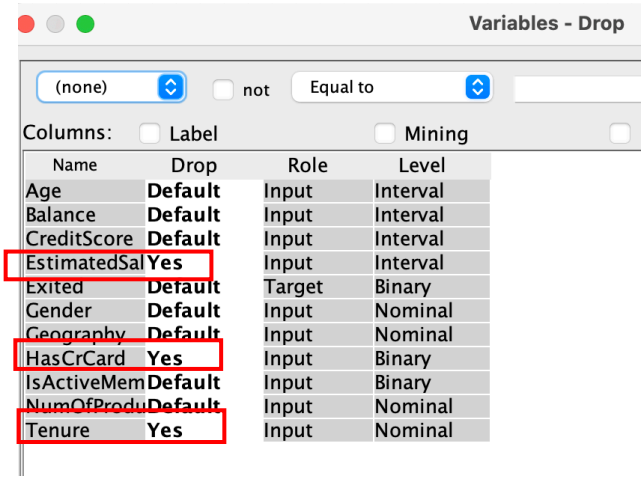
6.2.6 Bar chart of "NumofProducts" vs "Exited"



Looking at the customer churn status by number of products purchased, it is noticed that all customers who have purchased or subscribed to four bank products have churned. This might indicate that the products offered by the banks are causing dissatisfaction among the customers.

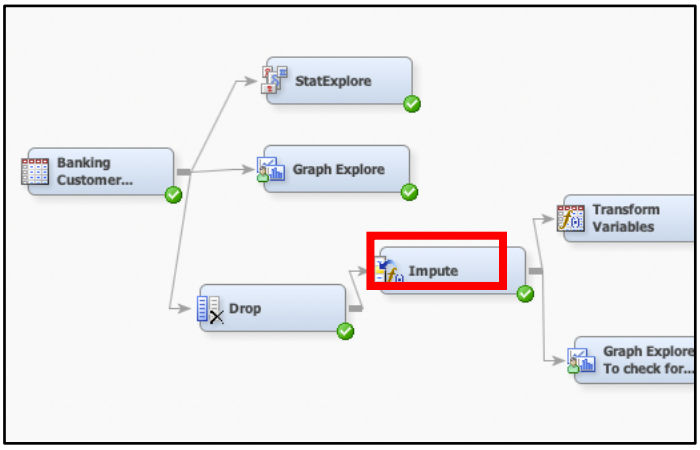
6.3 Removal of Variables

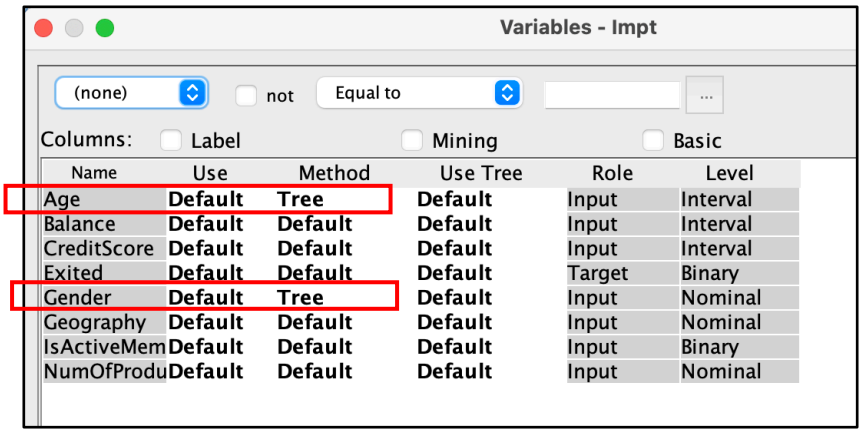




The "Drop" node is used to remove four independent variables which are found to be insignificant because of their extremely low variable worth and their p-values are more than 0.05 during the exploratory data analysis phase.

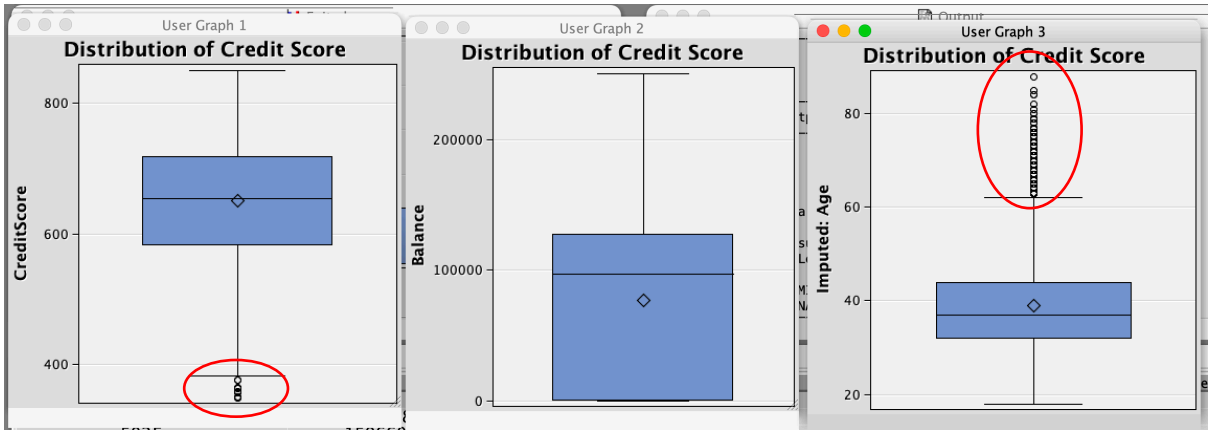
6.4 Imputation Of Missing Values



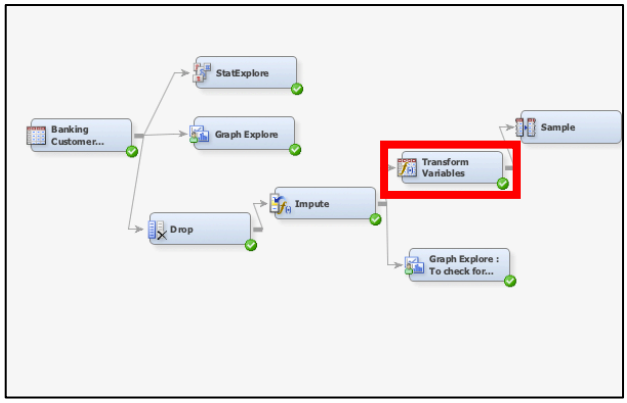


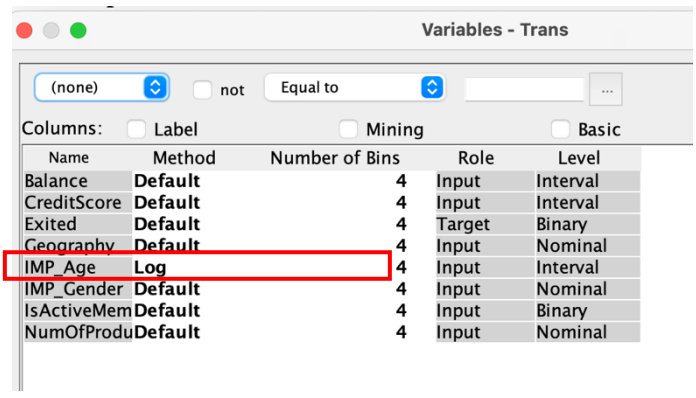
After dropping the insignificant variables, the "Impute" node is used to impute the missing values found in the variables. Since the variable "Tenure" was dropped previously, there are now only two variables with missing values (Gender and Age). The missing values will be imputed using the "Tree" method instead of using mean/mode. This is to prevent us from introducing biasness into the dataset.

6.5 Data Transformation

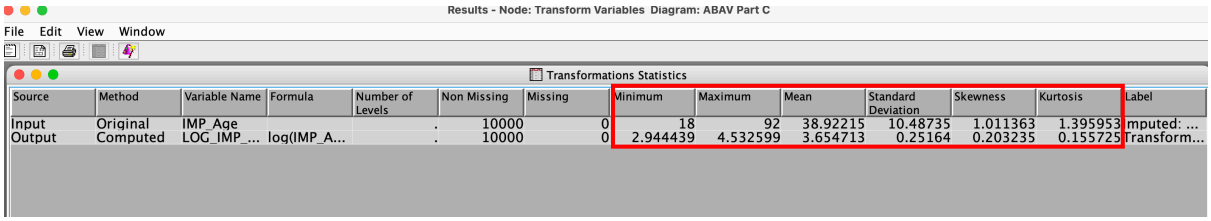


Based on the boxplot graph of all the continuous variables, it is found that there are extreme values in the variable "Age" and "CreditScore". These outliers or extreme values may spoil or mislead the prediction model resulting in less accurate models and ultimately poorer results. However, it is irrational to remove these outliers all together as they are genuine data. Therefore, transformation is required to reduce the outliers. As for "CreditScore", no transformation is needed as the outliers are mild outliers.



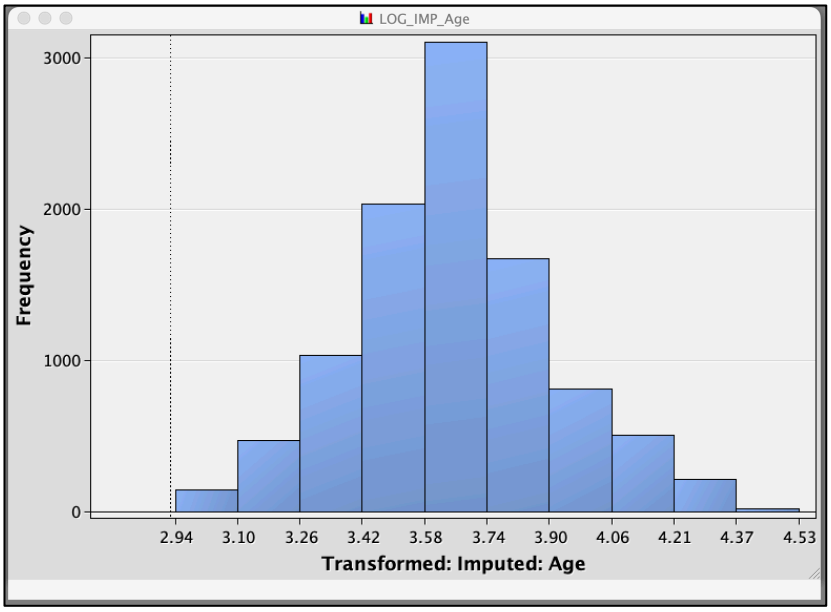


The "Transform Variables" node is used to perform a log transformation on the "Age" variable.



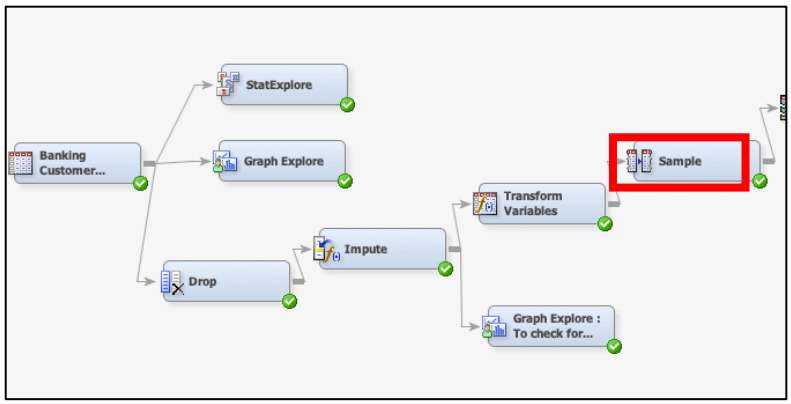
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	IMP_Age		.	10000	0	18	92	38.92215	10.48735	1.011363	1.395953	Imputed: ...
Output	Computed	LOG_IMP_...	log(IMP_A...	.	10000	0	2.944439	4.532599	3.654713	0.25164	0.203235	0.155725	Transform...

After transformation, it is noticed that the standard deviation, skewness and kurtosis of the variables have reduced significantly.



Besides that , it is also observed that the histogram of the transformed variable is now looking like a normal distribution. Hence, the transformation process is successful.

6.6 Class Imbalance

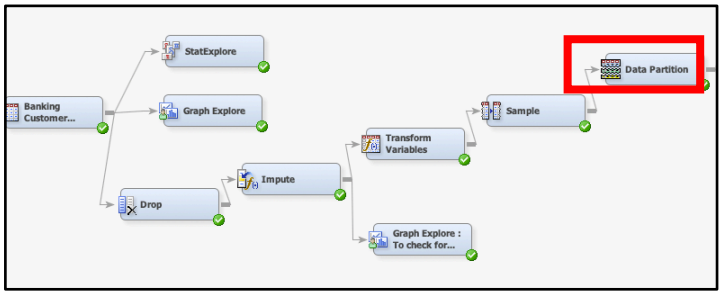


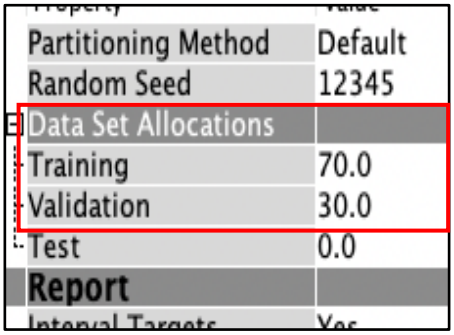
Property	Value
Type	Percentage
Observations	
Percentage	50.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event

```
Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

                 Numeric     Formatted     Frequency
Variable          Value         Value         Count      Percent     Label

 Exited               0             0          7963        79.63     Exited
 Exited               1             1          2037        20.37     Exited


Data=SAMPLE

                 Numeric     Formatted     Frequency
Variable          Value         Value         Count      Percent     Label

 Exited               0             0          2037           50     Exited
 Exited               1             1          2037           50     Exited
```

To solve the class imbalance problem, "Sample" node is used. Under the properties, the sample size of 50% and equal sampling method is being selected. After the sampling process, the total sample size is now 4,074 and half of the sample consists of customers who churned and the other half consists of customers who are still with the bank. The target variable is now perfectly balanced.

6.7 Data Partition



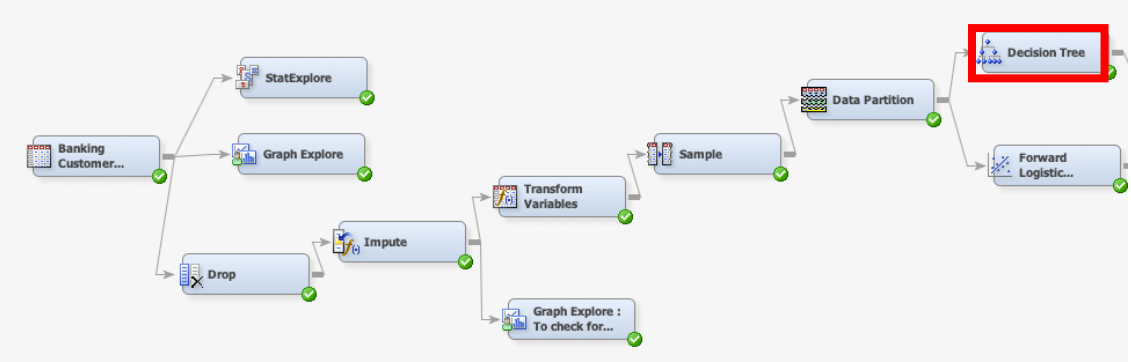



```
Partition Summary

                                          Number of
Type              Data Set                Observations

DATA         EMWS1.Smpl_DATA                  4074
TRAIN        EMWS1.Part_TRAIN                 2851
VALIDATE     EMWS1.Part_VALIDATE              1223
```

The dataset is split into 70% training and 30% validation. This is done to ensure that the prediction model can be tested on a different sample of unseen data which can help to assess the prediction model much better. It is observed that the training dataset now consists of 2,851 observations while the validation dataset has 1,223 observations.

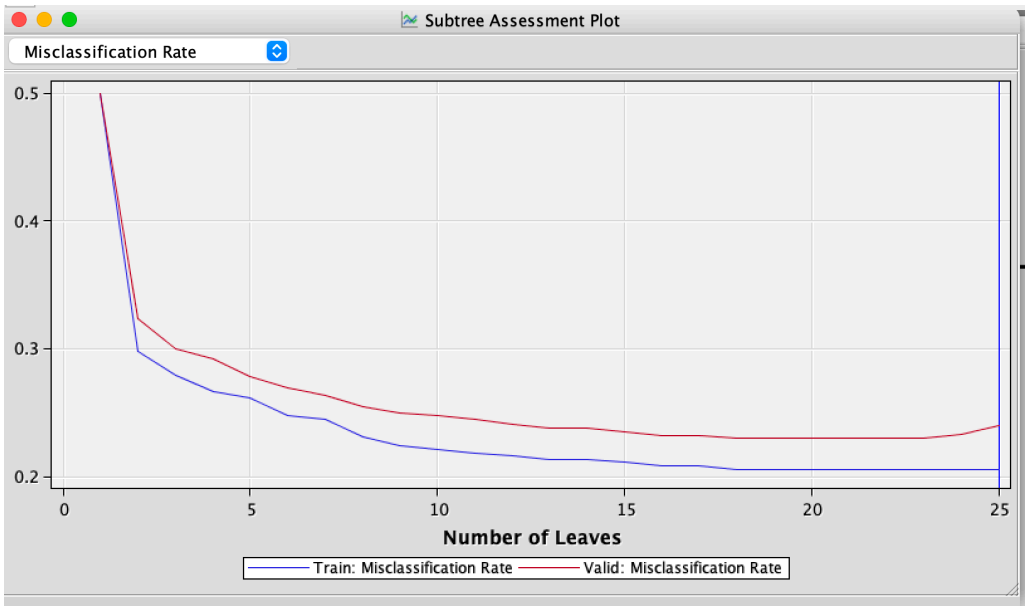
7. Model Construction, Optimization and Validation

7.1 Decision Tree Model



The first model is built using the "Decision Tree" node. In the case of predicting whether a customer has left the bank, decision tree is a great way to map all the possible outcomes and is able to show the bank what are the factors that are highly likely to cause customer churn.

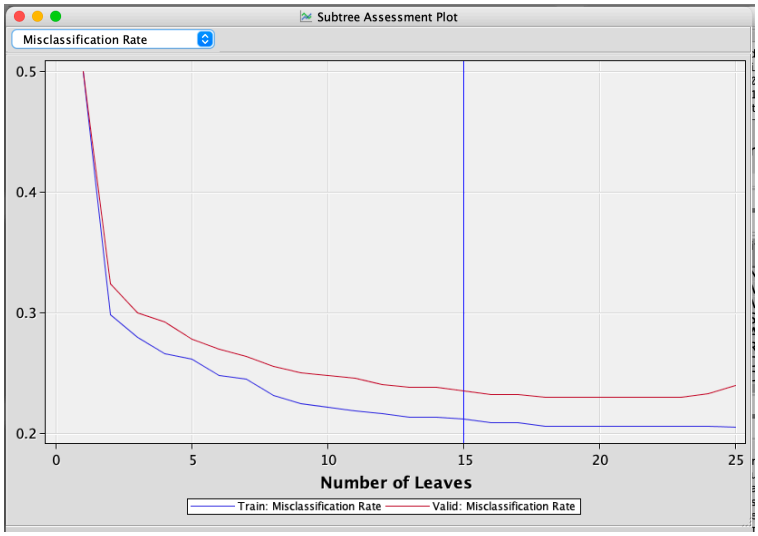
7.1.1 Decision Tree Model (Before Optimization)

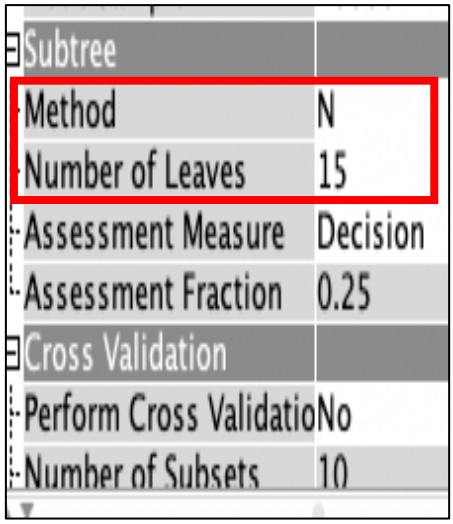


Based on the subtree assessment plot, it is noticed that the change in misclassification rate has stabilized around 15 leaves. Although there is still a slight decrease of 0.002 in terms of the

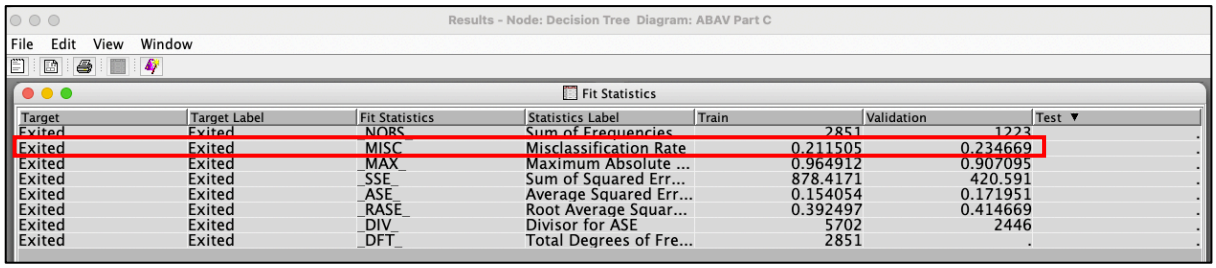
misclassification rate for both training and validation data, the amount is insignificant. Hence, pruning is required to remove the subtree that results in the least information gain.

7.1.2 Decision Tree Model (Optimized)





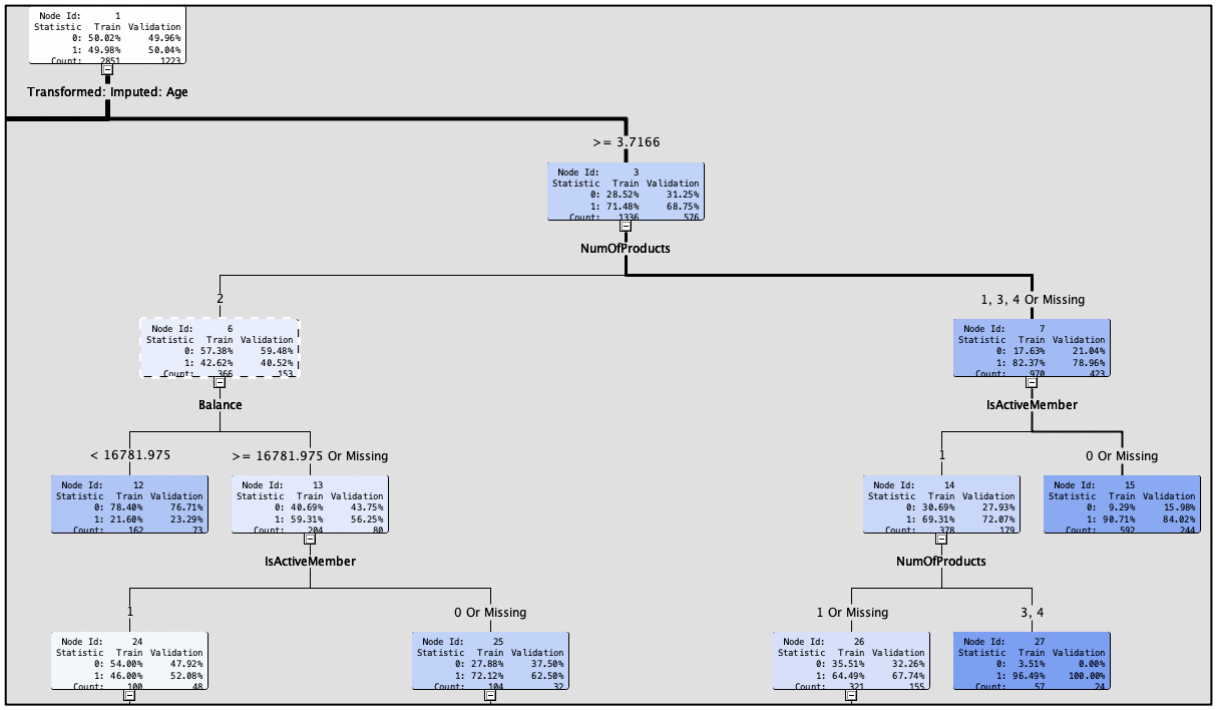
The decision tree model is being pruned using the N method with 15 leaves. After pruning, the complexity of the tree should be reduced and there should be less overfitting in our model.



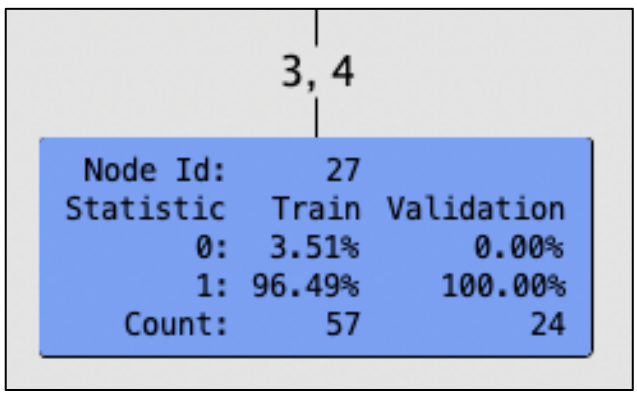
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited	Exited	_NOBS_	Sum of Frequencies	2851	1223	.
Exited	Exited	_MISC_	Misclassification Rate	0.211505	0.234669	.
Exited	Exited	_MAX_	Maximum Absolute ...	0.964912	0.907095	.
Exited	Exited	_SSE_	Sum of Squared Err...	878.4171	420.591	.
Exited	Exited	_ASE_	Average Squared Err...	0.154054	0.171951	.
Exited	Exited	_RASE_	Root Average Squar...	0.392497	0.414669	.
Exited	Exited	_DIV_	Divisor for ASE	5702	2446	.
Exited	Exited	_DFT_	Total Degrees of Fre...	2851	.	.

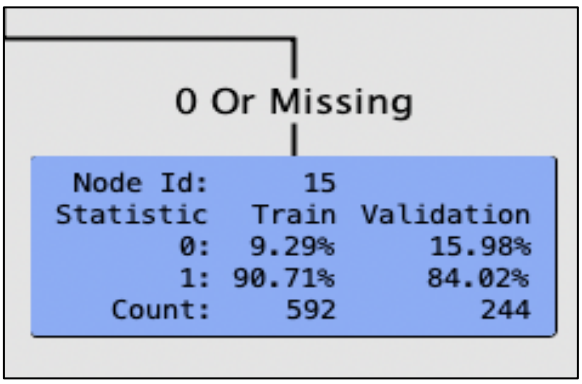
Based on the optimized decision tree model, it is observed that the misclassification rate of the model on the validation data is 0.23. This indicates that our decision tree model has an accuracy of 77%.

7.1.3 Interpretation of the model

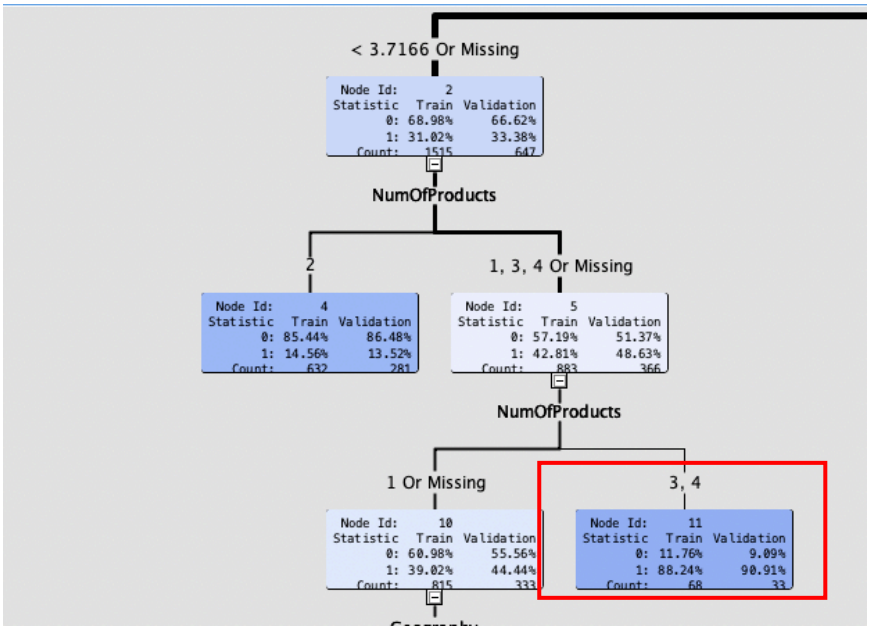


By glancing through the decision tree diagram, it is easily noticed that node 27 and node 15 are being highlighted with a darker colour compared to other nodes. This usually indicates that the node is particularly having a high probability of outcome. Also, it is noted that transformed "Age" variable would need to be converted to a normal value by exponential function. (e^3.7166 = 41)



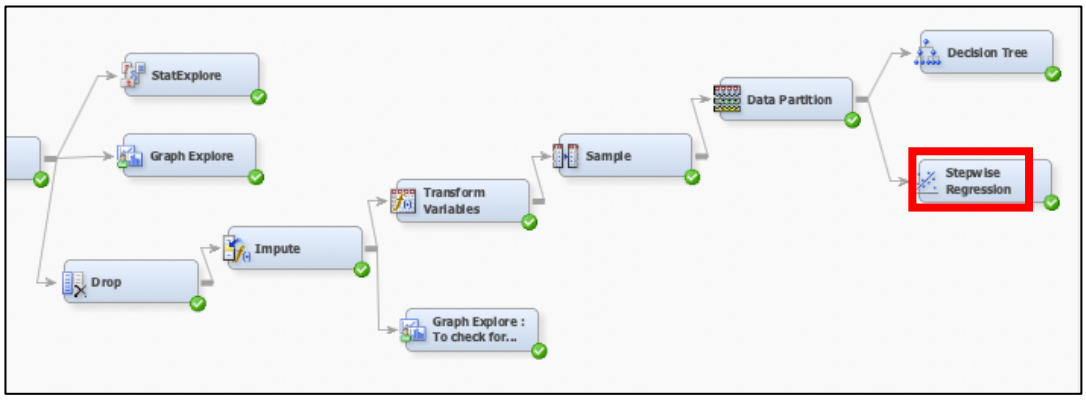


By diving deeper into both the nodes, it can be seen that if a customer is an active member, with age more than or equals to 41, and has purchased 3 or 4 products, the customer has a 100% chance of churning. In node 15, if a customer is not an active member, with age more than more than or equals to 41 years old and has purchased 1, 3 or 4 products, the customer has a 84% chance of churning.



Interestingly, for customers who are less than 41 years old, there is a 91% probability that the customer will churn if he/she purchased 3 or 4 products from the bank.

7.2 Logistic Regression Model



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

The stepwise logistic regression model is selected as the next prediction model with "Validation Misclassification" being selected as the selection criterion.

7.2.1 Interpretation of the Stepwise Logistic Regression Model

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	NumOfProducts	3	1	478.0420		<.0001	0.3083
2	LOG_IMP_Age	1	2	328.3824		<.0001	0.2674
3	IsActiveMember	1	3	142.7955		<.0001	0.2698
4	Geography	2	4	91.2186		<.0001	0.2617
5	IMP_Gender	1	5	34.8445		<.0001	0.2592
6	CreditScore	1	6	5.6056		0.0179	0.2551

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 6. It consists of the following effects:

Intercept CreditScore Geography IMP_Gender IsActiveMember LOG_IMP_Age NumOfProducts

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
3952.325	2799.788	1152.5373	9	<.0001

In the summary of stepwise selection table, it is noticed that the selected logistic model is step 6. The p-value of the overall model is less than 0.05, which indicates that the overall model is adequate.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
CreditScore	1	5.5924	0.0180
Geography	2	90.1821	<.0001
IMP_Gender	1	34.9602	<.0001
IsActiveMember	1	133.2851	<.0001
LOG_IMP_Age	1	287.0630	<.0001
NumOfProducts	3	276.7610	<.0001

Looking at the “Type 3 analysis of effects”, it is known that all six variables are important in the model due to their p-values less than 0.05.

Analysis of Maximum Likelihood Estimates

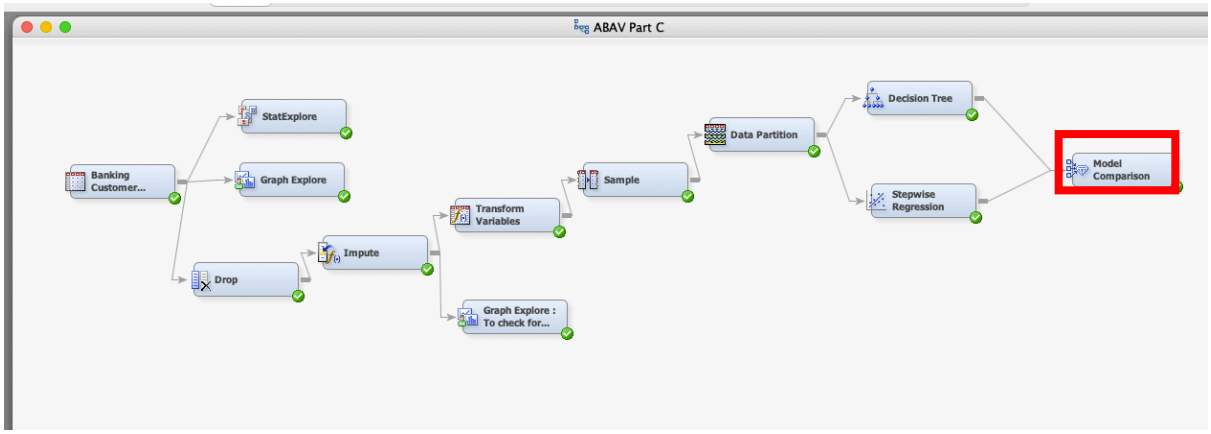
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-8.9679	9.5389	0.88	0.3471		0.000
CreditScore		1	-0.00113	0.000477	5.59	0.0180	-0.0612	0.999
Geography	France	1	-0.3834	0.0625	37.61	<.0001		0.682
Geography	Germany	1	0.6401	0.0703	82.80	<.0001		1.897
IMP_Gender	Female	1	0.2760	0.0467	34.96	<.0001		1.318
IsActiveMember	0	1	0.5487	0.0475	133.29	<.0001		1.731
LOG_IMP_Age		1	3.5006	0.2066	287.06	<.0001	0.4862	33.136
NumOfProducts	1	1	-2.9114	9.5042	0.09	0.7594		0.054
NumOfProducts	2	1	-4.3696	9.5043	0.21	0.6457		0.013
NumOfProducts	3	1	-0.5081	9.5073	0.00	0.9574		0.602

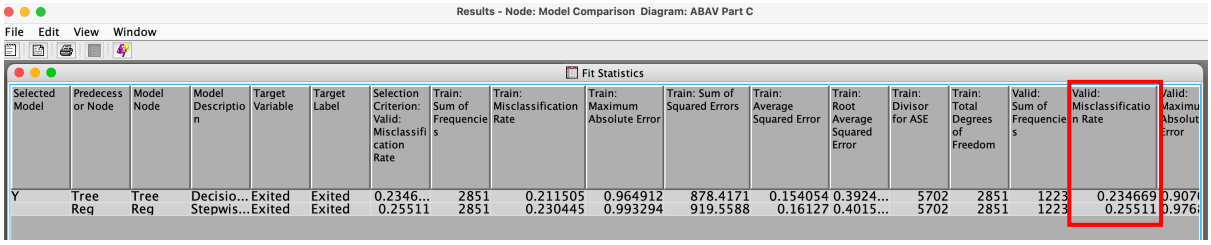
Odds Ratio Estimates

Effect		Point Estimate
CreditScore		0.999
Geography	France vs Spain	0.881
Geography	Germany vs Spain	2.452
IMP_Gender	Female vs Male	1.737
IsActiveMember	0 vs 1	2.996
LOG_IMP_Age		33.136
NumOfProducts	1 vs 4	<0.001
NumOfProducts	2 vs 4	<0.001
NumOfProducts	3 vs 4	<0.001

As for the odds ratio estimates, it is known that the variable "Age" has the highest point estimate of 33.136 and highest positive estimate value of 3.5006. This means that age has the strongest influence on the target variable. Since the estimate value is positive, it means that for an additional increase in the age of customer, the odds ratio of the customer leaving compared to the customer staying with the bank is 33.136. In short, the higher the age of the customer, the higher the chances of the customer churning.

8. Critical Interpretation Of Outcomes





From the model comparison table, decision model is identified as the better model due to its lower validation misclassification rate of 0.2347 compared to logistic regression model which has a misclassification rate of 0.25511.

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
Tree	Decision Tree	TRAIN	Exited	Exited	276	1099	327	1149
Tree	Decision Tree	VALIDATE	Exited	Exited	135	459	152	477
Reg	Stepwise Regression	TRAIN	Exited	Exited	337	1106	320	1088
Reg	Stepwise Regression	VALIDATE	Exited	Exited	165	464	147	447

Decision Tree	0 = Stay	1 = Churn
0 = Stay	TP (477)	FP (152)
1 = Churn	FN (135)	TN (459)

Stepwise Logistic Regression	0 = Stay	1 = Churn
0 = Stay	TP (447)	FP (147)
1 = Churn	FN (165)	TN (464)

However, when looking at the event classification table, the decision tree model is having a higher false positive compared to false negative. This means that the decision tree model is more likely to label a customer that is loyal as a customer that is going to defect. On the contrary, the logistic regression model have more false negative compared to false positive. This means that the stepwise logistic model is more likely to label a customer who is going to defect as someone who is loyal. Instead of going for the model with the highest accuracy, the bank should choose a prediction model that has the least number of false negative and higher false positive as it is a less costly mistake. The bank would not want to miss out in identifying customers who are going to churn as it is going to cost them more when compared to wrongly classified a loyal customer as someone who is going to churn. In this case, decision tree model should be the winner as it fulfils both the needs of the bank with the lowest misclassification rate and lowest number of false negative.

The table below summarises both the properties of prediction models and their interpretations in business terms.

Model Name	Properties	Evaluation/ Validation/ Results	Interpretation of outcomes in business terms
Decision Tree	Optimization is done by pruning the tree using N method and selecting only 15 leaves	After pruning the tree, a misclassification rate of 0.2347 @ 15 leaves is achieved.	There is a 100% chance that a customer will churn if the following criteria are met: 1. Age 41 and above 2. Purchased 3 or 4 products 3. Is an active member

			There is a 91% chance that a customer will churn if the following criteria are met: 1. Age lesser than 41 years old 2. Purchased 3 or 4 products There is 84% chance that a customer will churn if the following criteria are met: 1. Age 41 and above 2. Purchased 1,3 or 4 products 3. Is not an active member
Logistic regression	A stepwise approach is selected where the optimum model is at Step 6	The validation misclassification rate of 0.25511 is achieved	The outcome of "Exited" is heavily influenced by age. It is suspected that older people are more demanding and have lesser tolerance level towards the service/products provided by the bank. Hence, the bank should pay more attention to their older aged customers.

9. Discussion and Conclusion

In summary, decision tree based model is a more appropriate model for the bank to identify which customer is likely to churned based on historical data. It provides the probability of all the outcomes given certain characteristics which makes it easier to interpret and understand compared to a logistic model. Furthermore, since banks would like to what are the characteristics that causes their customer to churn, it is more practical to use a tree based model as it is able to generate nodes for each of the characteristics which are also customizable based on different needs. With a logistic regression model, the bank would only know which variable is significant in predicting the outcome, but there is very limited information provided when the bank wants to know the exact probability of a customer churning under certain circumstances. Therefore, by utilizing a decision tree model, banks are able to fulfil their business needs of identifying who are the customers that are going to churn and take preventive measures based on the probability given to them.

Based on the outcomes from the prediction models, there are a few conclusions that can be formed:

1. Age is the most influential factor in affecting whether a customer is going to leave the bank or stay with the bank. (Customers older than 41 years old are more likely to churn)
2. Customers who are inactive are more likely to churn
3. Customers who purchased 3 or 4 products highly likely to churn

In order to better retain their customers, the bank can take several measures like initiating customer engagement via various communication channels such as website and social media with the purpose of collecting customer feedback in order to better understand their overall banking experience. The bank can understand why or what are the reasons that causes the customer to be inactive or what are the difficulties that are being encountered when dealing with the bank. This is crucial to allow the bank to formulate effective strategies to prevent their customers from churning in the future. Besides that, the bank can revamp their product offerings and try to be more competitive in terms of pricing and rates. Since customers who purchased 3 to 4 products have high probability of churning, it is highly likely that the products offered by the banks are causing dissatisfaction among the customers. Finally, the bank can leverage on big data to create predictive models and conduct insightful data analytics to predict or identify customers with high churn risk. By identifying customers with high churn risk, the bank can take appropriate actions early to ensure customer retention.