GROUP PROJECT - REVISED

Churn Analytics for Credit Card Customers

Applied Analytic Modelling (BA706)

Tutor: David Parent

10TH JULY 2023

Group 3	
DARREN CARVALHO	301251637
SUDISH BASNET	301250603
JAGJEET SINGH MALHOTRA	301263211
OMPRAKASH BISTA	301251726

TABLE OF CONTENTS

1 Introduction

1.1 Problem Statement

In the world of finance, Banks have been facing a lot of issues which are mostly associated with the financial services provided by the respective financial institution. One of the major problems with financial services is credit card churning. This is a situation where a credit card company experiences a high rate of customers withdrawing their credit cards and switching to another company's card or taking other possible actions. This act of customers can lead to loss in high revenue from the cancelled cards, and it also incurs the cost of acquiring new customers to replace the ones who have churned. There are various reasons that can influence a user to withdraw the credit card services. One of the reasons can be customers using the service bonuses and getting rid of it before the end of the year to avoid the fee and later the same process is repeated with the other card. Other reasons associated with this can be client affordability, credit card points or credit limit, card category, and client backgrounds like gender, education, marital status, dependents, and livelihood.

1.2 Objective

The purpose of this analysis is to know the reason behind churn and recommend solutions according to the reasons. To do this analysis, a credit card company can use data mining and machine learning techniques to analyze customer behavior and identify potential churners by predicting the target variable associated with different entities from the dataset of customer’s information. So, with the identification of customers and the reasons with churn analytics, the company can take proactive steps to retain them, such as offering special promotions or incentives. This can help reduce the churn rate and increase customer loyalties.

2. Business Understanding for Churn Analytics

2.1 Finance

In the finance industry credit card churning is one of the common practices seen among the customers as the customer tends to switch service providers or quit the services after using all the special offers and bonuses.

2.2 Banking

In banking, Institutes use advertising as a tool to create hope among new and long-term clients. So, they prevent churning by having regulations to reduce the churning practice among the customers.

2.3 Stakeholder analysis

There is a churning which happens among internal stakeholder for e.g.: The employee working in the bank can involve in churning to get extra commission without knowing the fact that the customer will be finding that credit card services in his favor or not. The external stakeholders are the customers who are the most important aspect of the business, so their trust in accepting the services provided by the bank is a win-win situation in both ends.

3. Data Description

Credit card customers dataset consists of 10,000 customers mentioning their information like age, marital status, salary, credit card limit, credit card category, etc. In total, we have 24 attributes, out of which we will use the attrition flag as a target variable so that we can predict the churn. Simultaneously, the remaining attributes will be used as features to predict the target variable. However, after some exploratory data analysis, some features may be rejected as per their importance to our churn analytics.

In detail, our data description is explained below in the table as per the attributes. The dataset was taken from Kaggle (Kaggle, 2022).

Column	Description
CLIENTNUM	Client number. Unique identifier for the customer holding the account
Customer_Age	Demographic variable - Customer's Age in Years
Dependent_count	Demographic variable - Number of dependents
Marital_Status	Demographic variable - Married, Single, Divorced, Unknown
Card_Category	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Total_Relationship_count	Total no. of products held by the customer
Contacts_Count_12_mon	No. of Contacts in the last 12 months
Total_Revolving_Bal	Total Revolving Balance on the Credit Card
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Avg_Utilization_Ratio	Average Card Utilization Ratio
Attrition_Flag	Internal event (customer activity) variable - if the account is closed then 1 else 0
Gender	Demographic variable - M=Male, F=Female
Education_Level	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Income_Category	Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, >
Months_on_book	Period of relationship with bank
Months_Inactive_12_mon	No. of months inactive in the last 12 months
Credit_Limit	Credit Limit on the Credit Card
Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)
Naive_Bayes_Classifier_attribution	Naive Bayes
Naive_Bayes_Classifier_attribution	Naive Bayes

4. Data Preparation – ETL

The data appears to be highly imbalanced with only 16.06 % of the target as churned customers. To create a balanced target, a random sample of the size of churned customers was extracted from the existing customers and these were merged to form our data for modelling. The data contained ~3000 records after the down sampling which is still appropriate for modelling.

```
# Down sampling Existing customers

df2 = df[df['Attrition_Flag']=='Attrited Customer']

df1 = df[df['Attrition_Flag']=='Existing Customer']
df1 = df1.sample(1627)
df1.shape

df3 = pd.concat([df1, df2])
df3.shape
```

Since the target variable is binary, Attrition Flag is set to binary. ID, NBS1 and NBS2 are the variables which would be rejected. ID is a unique variable, and the target does not have any dependency on ID. NBS1 and NBS2 are variables created by the author which are results from a naïve bayes classification model. Explaining models can be difficult as it is not clear how and why these two variables were created.

Screenshot of the Data categories / rejections

Name	Partition Role	Role	Level
Attrition_Flag	**Default**	Target	Binary
Avg_Open_To_E	**Default**	Input	Interval
Avg_Utilization_I	**Default**	Input	Interval
Card_Category	**Default**	Input	Nominal
Contacts_Count	**Default**	Input	Interval
Credit_Limit	**Default**	Input	Interval
Customer_Age	**Default**	Input	Interval
Dependent_cour	**Default**	Input	Interval
Education_F	**Default**	Input	Interval
Education_Level	**Default**	Rejected	Nominal
Gender	**Default**	Input	Nominal
ID	**Default**	ID	Nominal
Income_Categor	**Default**	Rejected	Nominal
Income_FK	**Default**	Input	Interval
Marital_Status	**Default**	Input	Nominal
Months_Inactive	**Default**	Input	Interval
Months_on_bool	**Default**	Input	Interval
NB1	**Default**	Rejected	Interval
NB2	**Default**	Rejected	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
VAR1	**Default**	Rejected	Interval

During initial understanding of the data, two variables occur to be ordinal and can be ranked rather than creating dummy variables. Education Level and Income categories which have 7 and 6 unique values respectively. This encoding is expected to affect the regression and neural networks model due to the model requirements. A separate model evaluation with adding these rank encoded variables will be tested and the impact on performance will be observed. The ranking will be numeric with highest given to the most educated “Doctorate” in our case and maximum earning category which is “$120k +” in our case.

```python
[ ] # Ordinal Encoding for Education

    educ = {'High School' : 2,
     'Graduate' : 4,
     'Uneducated' : 1,
     'Unknown' :0,
     'College' : 3,
     'Post-Graduate' : 5,
     'Doctorate' : 6}


    df['Education_F'] = df['Education_Level'].map(educ)


    # Ordinal Encoding for Income Category

    df['Income_Category'].unique()

    inc_enc = {
        '$60K - $80K' : 70,
        'Less than $40K' : 40,
        '$80K - $120K' : 100,
        '$40K - $60K' : 50,
        '$120K +' : 120,
        'Unknown' : 0
    }

    df['Income_FK'] = df['Income_Category'].map(inc_enc)
    df.head(2)
```

t	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio	NB1	NB2	Education_F	Income_FK
4	42	1.625	0.061	0.000093	0.99991	2	70
1	33	3.714	0.105	0.000057	0.99994	4	40

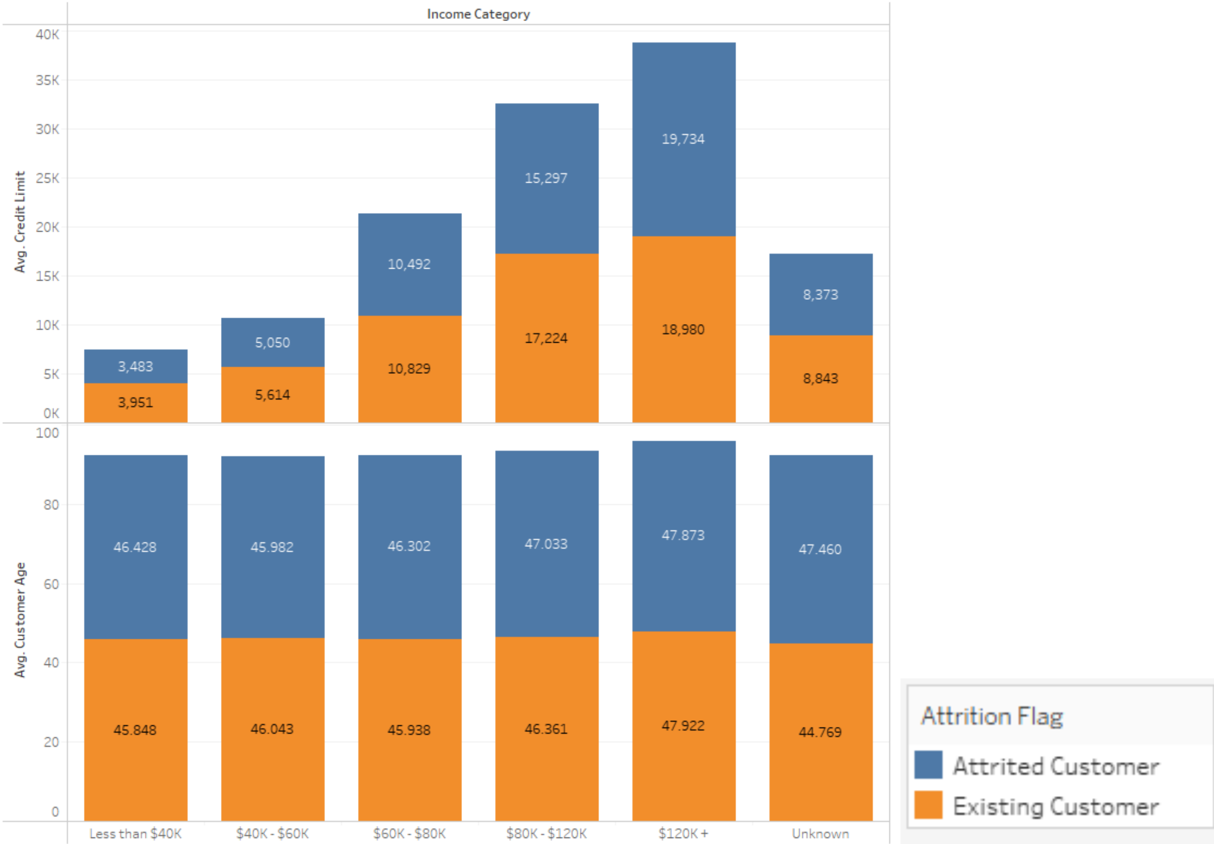
5. Summary statistics for Numerical / Categorical Variables

Interval Variables

	Target Level	Variable	Missing	Mean	Median	Skewness	Non Missing
ag	Existing Customer	Total_Trans_Amt	0	4645.641	4045	2.064733	814
ag	Existing Customer	Total_Amt_Chng_Q4_Q1	0	0.779862	0.746	1.895965	814
ag	Attrited Customer	Credit_Limit	0	7921.207	4028	1.858515	814
ag	Attrited Customer	Avg_Open_To_Buy	0	7226.508	3353	1.857117	814
ag	Existing Customer	Credit_Limit	0	8351.25	4798	1.772146	814
ag	Existing Customer	Avg_Open_To_Buy	0	7096.364	3682	1.751231	814
ag	Attrited Customer	Total_Trans_Amt	0	3066.125	2355	1.745838	814
ag	Attrited Customer	Avg_Utilization_Ratio	0	0.169102	0	1.590636	814
ag	Attrited Customer	Total_Ct_Chng_Q4_Q1	0	0.546996	0.52	1.503822	814
ag	Existing Customer	Total_Ct_Chng_Q4_Q1	0	0.741806	0.724	1.376499	814
ag	Attrited Customer	Total_Revolving_Bal	0	694.699	0	1.001164	814
ag	Existing Customer	Months_Inactive_12_mon	0	2.243243	2	0.708119	814
ag	Existing Customer	Avg_Utilization_Ratio	0	0.292096	0.206	0.615147	814
ag	Attrited Customer	Months_Inactive_12_mon	0	2.691646	3	0.578309	814
ag	Attrited Customer	Total_Trans_Ct	0	44.75184	43	0.406961	814
ag	Attrited Customer	Contacts_Count_12_mon	0	2.948403	3	0.404256	814
ag	Attrited Customer	Total_Relationship_Count	0	3.324324	3	0.233428	814
ag	Existing Customer	Dependent_count	0	2.269042	2	0.062744	814
ag	Existing Customer	Customer_Age	0	46.41032	46	-0.0055	814
ag	Existing Customer	Total_Trans_Ct	0	68.74079	71	-0.05018	814
ag	Existing Customer	Months_on_book	0	36.0688	36	-0.06122	814
ag	Attrited Customer	Customer_Age	0	46.60197	47	-0.06198	814
ag	Existing Customer	Contacts_Count_12_mon	0	2.358722	2	-0.0635	814
ag	Attrited Customer	Months_on_book	0	35.92752	36	-0.0864	814
ag	Attrited Customer	Dependent_count	0	2.366093	2	-0.11859	814
ag	Existing Customer	Total_Relationship_Count	0	3.941032	4	-0.27222	814
ag	Existing Customer	Total_Revolving_Bal	0	1254.886	1336	-0.30494	814
ag	Attrited Customer	Total_Amt_Chng_Q4_Q1	0	0.702741	0.712	-0.33815	814

Numerical variables did not contain any missing values but there were a few variables which were skewed. Three categorical variables (Education level, Marital Status, and Income Category) contained fields labelled unknown which have been considered as missing.

5.1 Customer Age, Credit limit in each income group for both the target classes



Average customer age and Average credit limit in each income category for both the classes.

As expected, the credit limit increases with an increase in the income. The average age for all income categories was around 44-45 years, which is probably the target group for the survey. There were no extreme or unique values for either of the classes as they showed similar trends for each income group; credit limit and age.

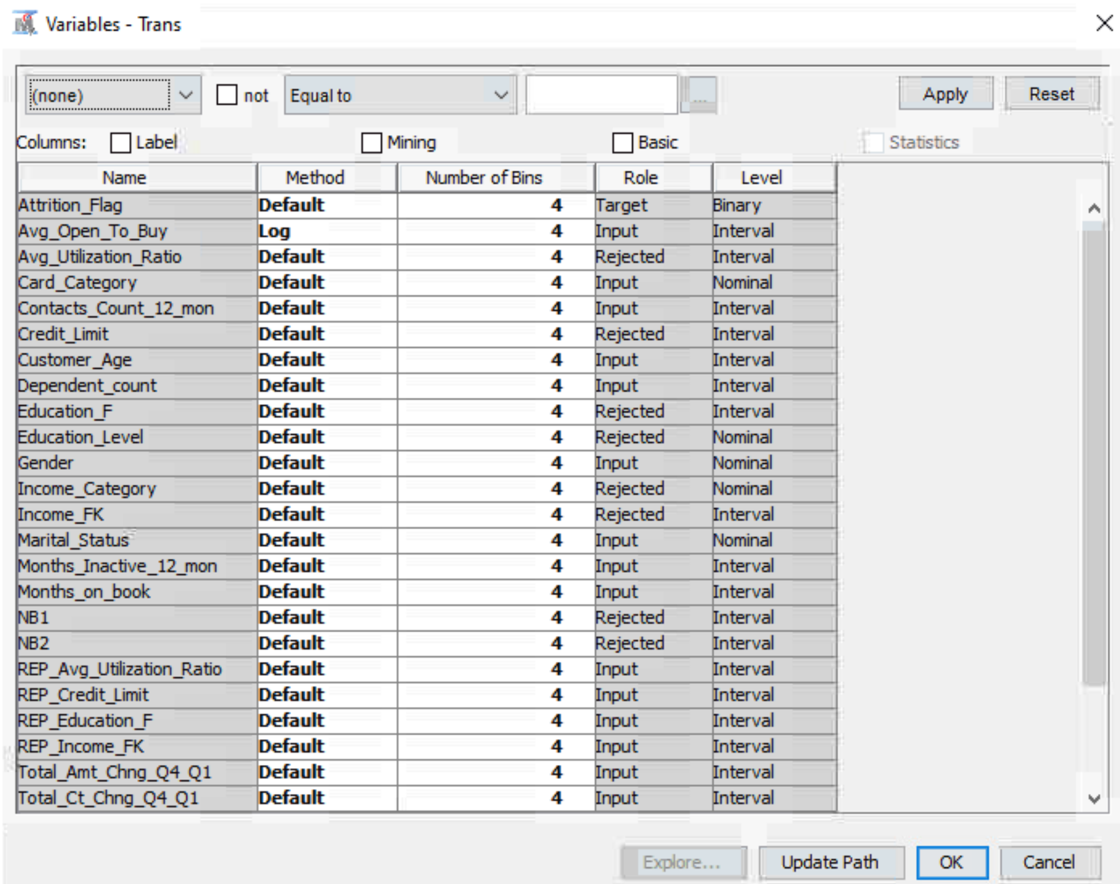
5.2 Gender wise proportion of card categories

	Card Category				
Gender	Blue	Gold	Platinum	Silver	Grand To..
F	1,686	17	4	72	1,779
M	1,337	26	4	108	1,475
Grand Total	3,023	43	8	180	3,254

As expected, most of the customers use the blue card category as the rest may have some annual fee or other restrictions upon it. The proportion of females is highest in the blue category while for other cards, the proportion of male is higher. The platinum users are only 8 among 3000 which would not be significant enough to model, as a result creating a model with blue card category as a class against all other card categories in one class may provide significant results.

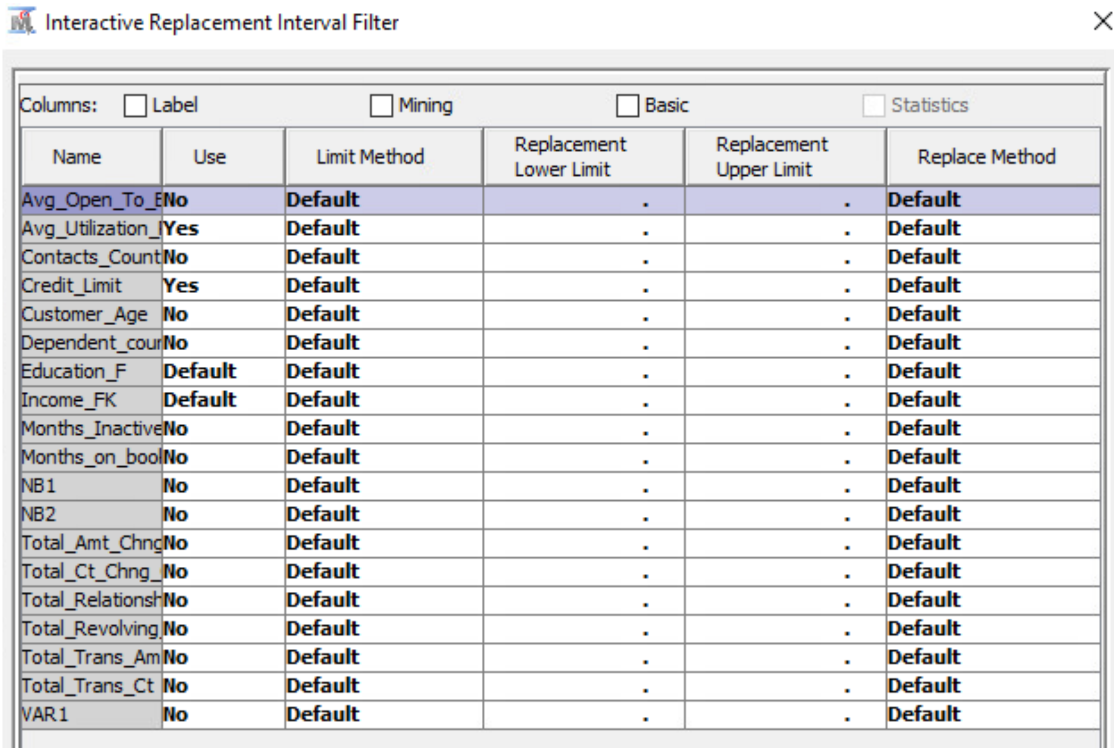
6. Transformations / Cap and Floor / Replacing Unknown with Mode

There are 4 variables which were skewed. Average utilization Ratio, Credit Limit, Total Transaction Amount, Average open to buy. These 4 variables were skewed for both the classes Existing and Attired customers. There were other variables as well with skewness but those were with respect to one class only which could be an impacting factor on the target and the skewness value was not very significant. As a result, we decided to perform skewness on these 4 variables. Our rational statistical technique to resolve the skewness was that the ones with higher skewness statistic were log transformed and the ones with lower skewness statistic were handled using cap and floor.



Name	Method	Number of Bins	Role	Level
Attrition_Flag	Default	4	Target	Binary
Avg_Open_To_Buy	Log	4	Input	Interval
Avg_Utilization_Ratio	Default	4	Rejected	Interval
Card_Category	Default	4	Input	Nominal
Contacts_Count_12_mon	Default	4	Input	Interval
Credit_Limit	Default	4	Rejected	Interval
Customer_Age	Default	4	Input	Interval
Dependent_count	Default	4	Input	Interval
Education_F	Default	4	Rejected	Interval
Education_Level	Default	4	Rejected	Nominal
Gender	Default	4	Input	Nominal
Income_Category	Default	4	Rejected	Nominal
Income_FK	Default	4	Rejected	Interval
Marital_Status	Default	4	Input	Nominal
Months_Inactive_12_mon	Default	4	Input	Interval
Months_on_book	Default	4	Input	Interval
NB1	Default	4	Rejected	Interval
NB2	Default	4	Rejected	Interval
REP_Avg_Utilization_Ratio	Default	4	Input	Interval
REP_Credit_Limit	Default	4	Input	Interval
REP_Education_F	Default	4	Input	Interval
REP_Income_FK	Default	4	Input	Interval
Total_Amt_Chng_Q4_Q1	Default	4	Input	Interval
Total_Ct_Chng_Q4_Q1	Default	4	Input	Interval

6.1 Log Transformation

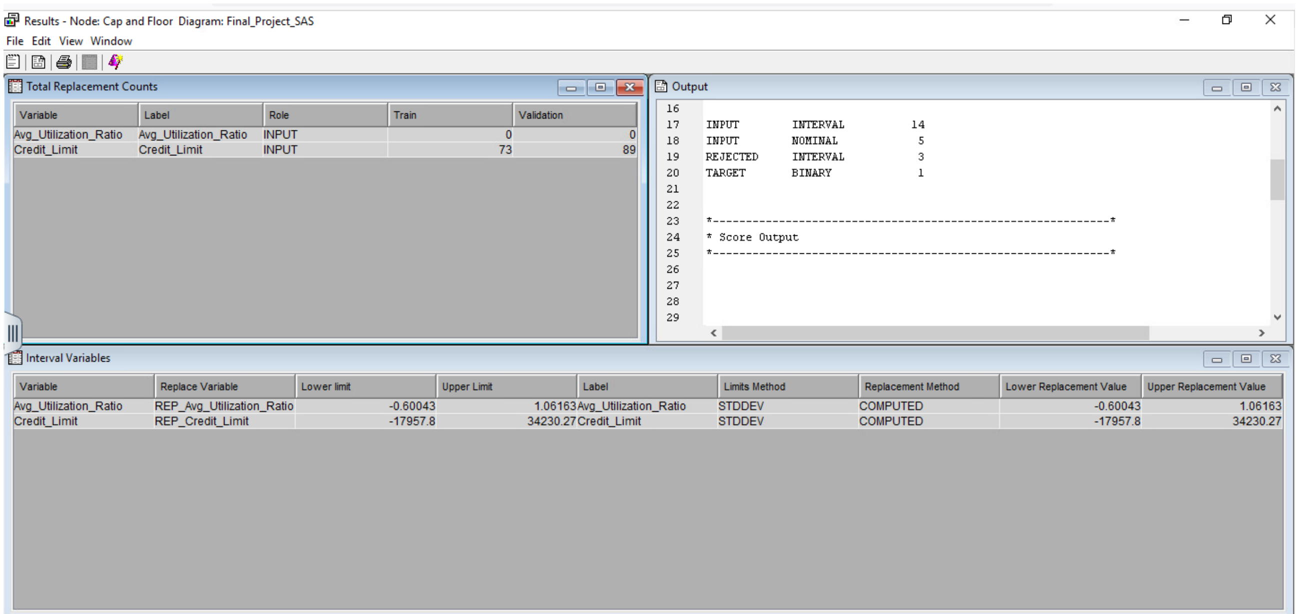


Interactive Replacement Interval Filter

Columns: Label Mining Basic Statistics

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method
Avg_Open_To_E	No	Default	.	.	Default
Avg_Utilization_	Yes	Default	.	.	Default
Contacts_Count	No	Default	.	.	Default
Credit_Limit	Yes	Default	.	.	Default
Customer_Age	No	Default	.	.	Default
Dependent_cour	No	Default	.	.	Default
Education_F	Default	Default	.	.	Default
Income_FK	Default	Default	.	.	Default
Months_Inactive	No	Default	.	.	Default
Months_on_bool	No	Default	.	.	Default
NB1	No	Default	.	.	Default
NB2	No	Default	.	.	Default
Total_Amt_Chng	No	Default	.	.	Default
Total_Ct_Chng_	No	Default	.	.	Default
Total_Relationsh	No	Default	.	.	Default
Total_Revolving	No	Default	.	.	Default
Total_Trans_Am	No	Default	.	.	Default
Total_Trans_Ct	No	Default	.	.	Default
VAR1	No	Default	.	.	Default

6.2 Cap and Floor



Results - Node: Cap and Floor Diagram: Final_Project_SAS

File Edit View Window

Total Replacement Counts

Variable	Label	Role	Train	Validation
Avg_Utilization_Ratio	Avg_Utilization_Ratio	INPUT	0	0
Credit_Limit	Credit_Limit	INPUT	73	89

Output

```
16
17  INPUT      INTERVAL      14
18  INPUT      NOMINAL        5
19  REJECTED   INTERVAL       3
20  TARGET     BINARY         1
21
22
23  *------------------------------------------------------------*
24  * Score Output
25  *------------------------------------------------------------*
26
27
28
29
```

Interval Variables

Variable	Replace Variable	Lower limit	Upper Limit	Label	Limits Method	Replacement Method	Lower Replacement Value	Upper Replacement Value
Avg_Utilization_Ratio	REP_Avg_Utilization_Ratio	-0.60043	1.06163	Avg_Utilization_Ratio	STDDEV	COMPUTED	-0.60043	1.06163
Credit_Limit	REP_Credit_Limit	-17957.8	34230.27	Credit_Limit	STDDEV	COMPUTED	-17957.8	34230.27

Interval variables did not contain any missing values. While there were 3 class variables Education level, Marital Status and Income Category which contained a class unknown. We considered unknown to be a missing field and the probable reason for unknown is that the information may seem confidential and people undertaking the survey may not have wanted to answer this question. We could create a model with these variables as it is not necessary that customers want to share such details and we could have a custom class for such input. But the unknown class could contain a vast number of customer categories which may or may not include the predefined classes among the variables. For instance, unknown in marital status could mean married but do not want to share details, divorced, or even single and this could impact model performance. As a result, we decided to impute the missing values with the mode from each variable.

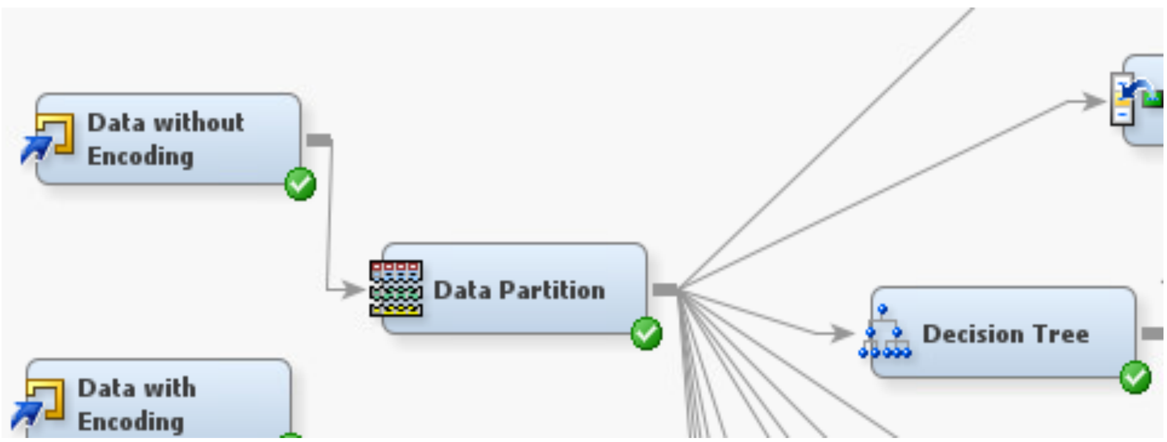
7. Modelling

7.1 Decision Tree Models

Tree was used as the first predictive model, as Tree does not require any data handling or such; we can pass dirty data to it and it can give us the important variables. So, to understand the data more we used Tree as the first model.

7.1.1 Maximal Tree

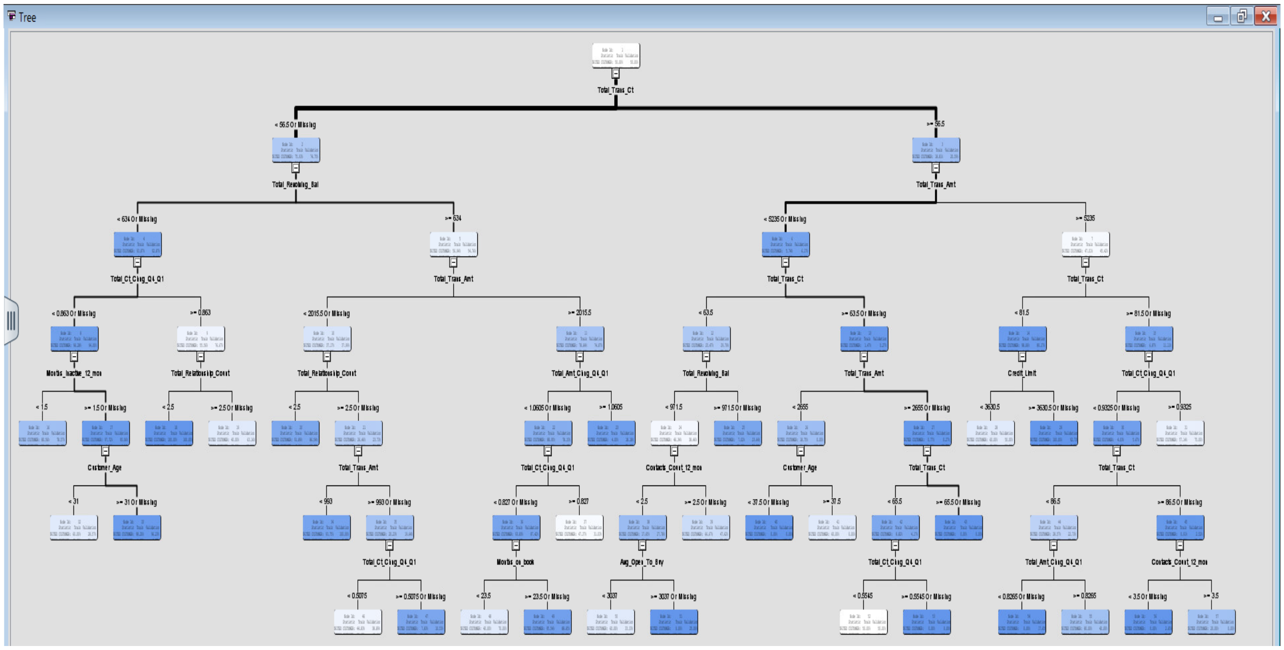
Data was imported and connected to the data partition node; the data partition node split the data into 50:50. That is 50% for training and 50% for validation. After this, a decision Tree node was brought in and Maximal tree was produced.



And the following changes were made to the properties:

ubtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	

The results for this tree were as follows:



So, we got a maximal tree after splitting was done for all the leaves having a logit value greater than 0.07. So, we got a tree with 29 leaves.

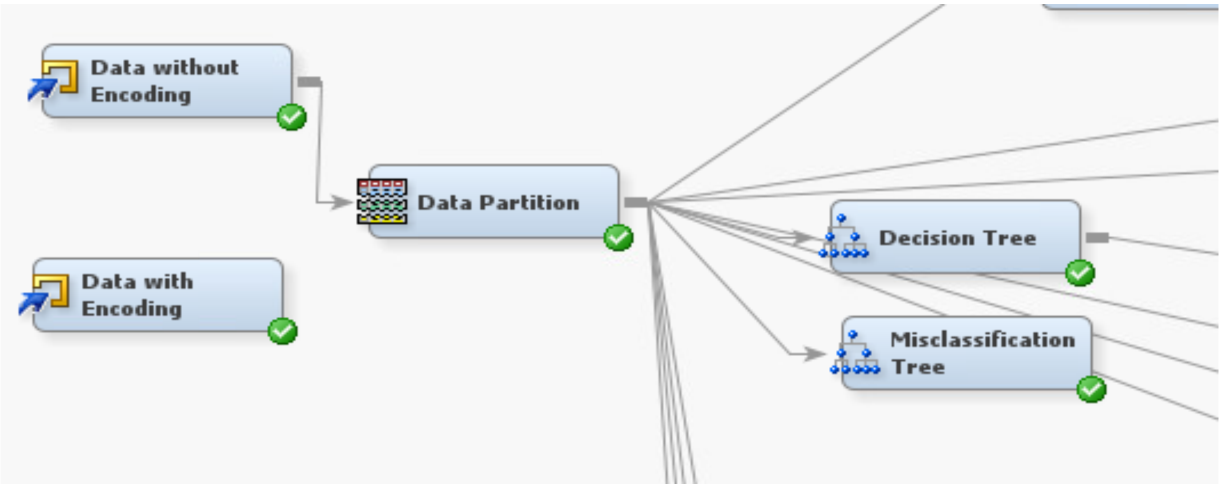
The misclassification rate for this tree was found to be: 0.105166.

With this Tree we can find out that, the maximum number of attired customers are caused by the following variables: Total_Trans_Cnt plays an important role and is the first split in the tree, followed by Total_Revolving Balance.

So, we can see that at node 17, the validation score for 405 customers is 95%, which shows that our model has a good build quality. So, we can conclude that the Customers having transaction counts less than 56.5 and a revolving balance less than 634 with inactive years close to 1.5 have a higher chance of getting churned, so we need to formulate our marketing strategy keeping these people in mind.

7.1.2 Misclassification Tree

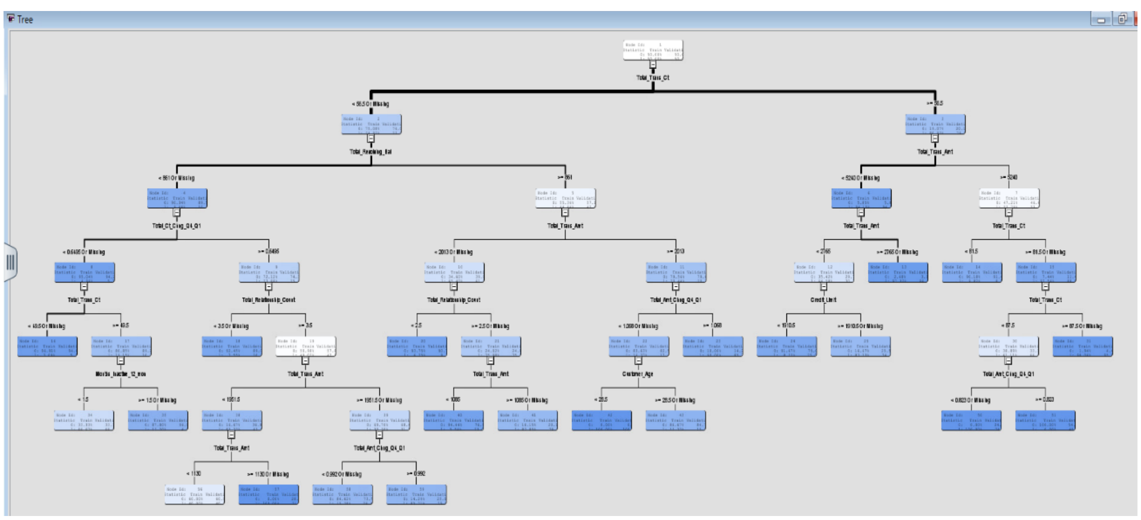
Now to reduce the curse of dimensionality we start pruning the tree and keeping the assessment value to be misclassification rate. So, we bring in another decision tree node and connect it to the data partition node.



The following changes were made to the properties panel:

Method	Assessment
umber of Leaves	1
ssessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	

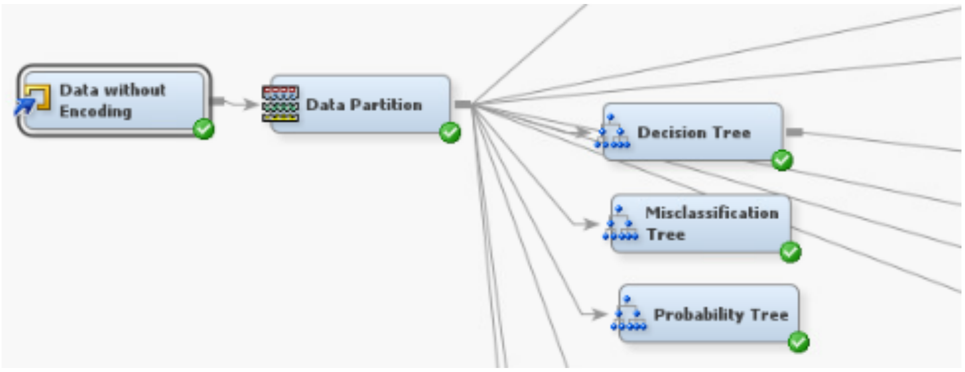
Now we run the node and find the pruned tree, the results were as follows:



The number of leaves was reduced from 29 to 21 and the value for misclassification rate was also reduced to 0.096556.

7.1.3 Probability Tree

Now we changed the assessment value to Average square error in the decision tree node and connected it to the data partition.



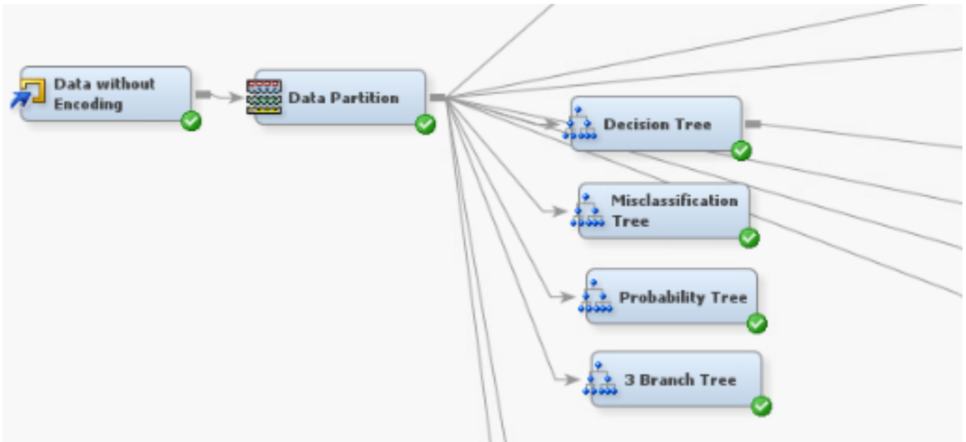
Subtree	
Method	Assessment
Number of Leaves	1
ssessment Measure	Average Square Error
ssessment Fraction	0.25
ross Validation	

We can find that the misclassification rate increased to 0.099 so this model is not better than the previous model, so we do not select this model.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition_Flag		NOBS	Sum of Frequencies	1628	1626	
Attrition_Flag		MISC	Misclassification Rate	0.074939	0.099016	
Attrition_Flag		MAX	Maximum Absolute Error	0.990909	1	
Attrition_Flag		SSE	Sum of Squared Errors	176.4582	248.1521	
Attrition_Flag		ASE	Average Squared Error	0.054195	0.076308	
Attrition_Flag		RASE	Root Average Squared Error	0.232798	0.276238	
Attrition_Flag		DIV	Divisor for ASE	3256	3252	
Attrition_Flag		DFT	Total Degrees of Freedom	1628	.	

7.1.4 3 Leaves Tree

In our models above, we could see that the tree was splitting at same variables again and again. To overcome that, we set the number of branches to 3 and assessment method set to misclassification and check for the misclassification rate.



Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
gnificance Level	0.2
ssing Values	Use in search
se Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5

It does reduce the number of leaves from 29 to 20 but the misclassification rate does not improve and the value for it is 0.10 which is not better than the normal misclassification tree. So, we reject this model as well.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Attrition_Flag		NOBS	Sum of Frequencies	1628	1626
Attrition_Flag		MISC	Misclassification Rate	0.067568	0.107011
Attrition_Flag		MAX	Maximum Absolute Error	0.992157	1
Attrition_Flag		SSE	Sum of Squared Errors	175.9933	284.06
Attrition_Flag		ASE	Average Squared Error	0.054052	0.087349
Attrition_Flag		RASE	Root Average Squared Error	0.232491	0.295549
Attrition_Flag		DIV	Divisor for ASE	3256	3252
Attrition_Flag		DFT	Total Degrees of Freedom	1628	.

7.1.5 Summary

We can find that the misclassification tree is the best tree, with the minimum misclassification rate. The summary for misclassification rates can be found in the table below.

Sr. No	Decision Tree Model	Misclassification Rate (Validation Data)
1	Maximal Tree	0.105166

2	Misclassification Tree	0.095941
3	Probability Tree	0.099016
4	3 Branch Misclassification tree	0.107011

7.1.6 Inferences from this model

So, we found out Total transaction count, total revolving balance, and total transaction amount play an important role in deciding if a customer would churn or not. With our analysis, we found out that for a customer having total transaction count less than 56.5 and total revolving balance less than $634, there is a 93% chance of the customer churning. So, we need to focus on these customers and send some promotional messages or email to such customers so that they do not churn and are retained.

7.2 Logistic Regression

7.2.3 Full Regression

Results - Node: Full Regression Diagram: Final_Project_SAS

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition_Flag		_AIC_	Akaike's Information Crite...	1021.01	.	.
Attrition_Flag		_ASE_	Average Squared Error	0.09094	0.10232	.
Attrition_Flag		_AVERR_	Average Error Function	0.296379	0.325942	.
Attrition_Flag		_DFE_	Degrees of Freedom for E...	1600	.	.
Attrition_Flag		_DFM_	Model Degrees of Freedom	28	.	.
Attrition_Flag		_DFT_	Total Degrees of Freedom	1628	.	.
Attrition_Flag		_DIV_	Divisor for ASE	3256	3252	.
Attrition_Flag		_ERR_	Error Function	965.0101	1059.964	.
Attrition_Flag		_FPE_	Final Prediction Error	0.094122	.	.
Attrition_Flag		_MAX_	Maximum Absolute Error	0.991258	0.997972	.
Attrition_Flag		_MSE_	Mean Square Error	0.092531	0.10232	.
Attrition_Flag		_NOBS_	Sum of Frequencies	1628	1626	.
Attrition_Flag		_NW_	Number of Estimate Weig...	28	.	.
Attrition_Flag		_RASE_	Root Average Sum of Squ...	0.301562	0.319875	.
Attrition_Flag		_RFPE_	Root Final Prediction Error	0.306794	.	.
Attrition_Flag		_RMSE_	Root Mean Squared Error	0.304189	0.319875	.
Attrition_Flag		_SBC_	Schwarz's Bayesian Criter...	1172.073	.	.
Attrition_Flag		_SSE_	Sum of Squared Errors	296.0994	332.7451	.
Attrition_Flag		_SUMW_	Sum of Case Weights Ti...	3256	3252	.
Attrition_Flag		_MISC_	Misclassification Rate	0.12715	0.143296	.

Misclassification rate for validation in Full regression is 0.143296.

Results - Node: Full Regression Diagram: Final_Project_SAS

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	13.3943	2.8285	22.42	<.0001		999.000
Contacts_Count_12_mon		1	-0.5476	0.0803	46.48	<.0001	-0.3411	0.578
Customer_Age		1	-0.0117	0.0175	0.45	0.5039	-0.0504	0.988
Dependent_count		1	-0.1392	0.0679	4.19	0.0405	-0.0987	0.870
Gender	F	1	-0.4427	0.1626	7.41	0.0065		0.642
Months_Inactive_12_mon		1	-0.5884	0.0882	44.53	<.0001	-0.3218	0.555
Months_on_book		1	0.0309	0.0178	3.02	0.0824	0.1334	1.031
REP_Avg_Utilization_Ratio		1	3.6213	0.9818	13.61	0.0002	0.5531	37.385
REP_Card_Category	Blue	1	0.0527	0.1996	0.07	0.7918		1.054
REP_Credit_Limit		1	-0.00008	0.000023	12.06	0.0005	-0.3878	1.000
REP_Education_Level	College	1	-0.00375	0.2448	0.00	0.9878		0.996
REP_Education_Level	Doctorate	1	-0.1744	0.3266	0.29	0.5933		0.840
REP_Education_Level	Graduate	1	0.2466	0.1489	2.74	0.0976		1.280
REP_Education_Level	High School	1	0.1271	0.1872	0.46	0.4969		1.136
REP_Education_Level	Post-Graduate	1	-0.4278	0.3026	2.00	0.1575		0.652
REP_Income_Category	$120K +	1	0.1310	0.2804	0.22	0.6404		1.140
REP_Income_Category	$40K - $60K	1	0.1610	0.2101	0.59	0.4436		1.175
REP_Income_Category	$60K - $80K	1	-0.2961	0.2218	1.78	0.1817		0.744
REP_Income_Category	$80K - $120K	1	-0.0544	0.2141	0.06	0.7993		0.947
REP_LOG_Avg_Open_To_Buy		1	1.0319	0.2473	17.41	<.0001	0.7156	2.806
REP_LOG_Total_Trans_Amt		1	-4.4252	0.3439	165.57	<.0001	-1.6127	0.012
REP_Marital_Status	Divorced	1	-0.1344	0.2122	0.40	0.5265		0.874
REP_Marital_Status	Married	1	0.3334	0.1367	5.95	0.0147		1.396
Total_Amt_Chng_Q4_Q1		1	1.8875	0.4351	18.82	<.0001	0.2321	6.603
Total_Ct_Chng_Q4_Q1		1	2.5975	0.3858	45.33	<.0001	0.3465	13.431
Total_Relationship_Count		1	0.4153	0.0568	53.54	<.0001	0.3629	1.515
Total_Revolving_Bal		1	0.000351	0.000171	4.19	0.0406	0.1755	1.000
Total_Trans_Ct		1	0.1942	0.0119	267.26	<.0001	2.4335	1.214

Full regression analysis Based on valid chi-square.

As the highest valid chi-square for full regression model is with the parameter Total_Trans_Ct, that is 267.26. As a result, it is the most important variable based on valid chi-square followed by REP_LOG_Total_Trans_Amt which is 165.57, it is second important variable to be look after. Third important variable is Total_Relationship_Count which is 53.54 and least valid chi square variable is REP_Education_Level college which is 0.00.

File Edit View Window

Output

		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Total_Revolving_Bal		1	0.000351	0.000171	4.19	0.0406	0.1755	1.000
Total_Trans_Ct		1	0.1942	0.0119	267.26	<.0001	2.4335	1.214

Odds Ratio Estimates

Effect		Point Estimate
Contacts_Count_12_mon		0.578
Customer_Age		0.988
Dependent_count		0.870
Gender	F vs M	0.413
Months_Inactive_12_mon		0.555
Months_on_book		1.031
REP_Avg_Utilization_Ratio		37.385
REP_Card_Category	Blue vs GSP	1.111
REP_Credit_Limit		1.000
REP_Education_Level	College vs Uneducated	0.790
REP_Education_Level	Doctorate vs Uneducated	0.666
REP_Education_Level	Graduate vs Uneducated	1.015
REP_Education_Level	High School vs Uneducated	0.900
REP_Education_Level	Post-Graduate vs Uneducated	0.517
REP_Income_Category	$120K + vs Less than $40K	1.075
REP_Income_Category	$40K - $60K vs Less than $40K	1.108
REP_Income_Category	$60K - $80K vs Less than $40K	0.701
REP_Income_Category	$80K - $120K vs Less than $40K	0.893
REP_LOG_Avg_Open_To_Buy		2.806
REP_LOG_Total_Trans_Amt		0.012
REP_Marital_Status	Divorced vs Single	1.067
REP_Marital_Status	Married vs Single	1.703
Total_Amt_Chng_Q4_Q1		6.603
Total_Ct_Chng_Q4_Q1		13.431
Total_Relationship_Count		1.515
Total_Revolving_Bal		1.000
Total_Trans_Ct		1.214

The odds ratio estimate for full regression is as follows.

- If we look at the most important variables from above and try to interpret the odds ratio for that we can come up with the following inferences for them. For Total_Trans_Ct if there is a 1 unit increase that is for every increase in the quantity of Trans_Ct the chance that they might churn would increase by 21.4%. So the more the number of transactions the more likely is the person to churn.
- For the next important variable that is Log_Total_Trans_Amt the odds ratio is 0.012, that is for every exponential increase in total transaction amount the chance of person churning would decrease by 98%, that is a quite significant increase and tells us a lot about the importance of this variable.
- Similarly, if we look at other factors and their odds ratio we can make the following inferences :
- Increase in customer age has 1.2% less chance of churning.
- Male has 59% chance of churning compared to female.
- People who are educated are 21% less likely to churn against people who are uneducated.
- A person having an income of more than $120K is less likely to churn compared to individuals having an income of less than $40k.
- People who are divorced are 6.7% less likely of getting churned compared to people who are single.
- Customers having blue cards have 11% less likely of being churned while compared to GSP card category.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Contacts_Count_12_mon	1	46.4761	<.0001
Customer_Age	1	0.4468	0.5039
Dependent_count	1	4.1947	0.0405
Gender	1	7.4133	0.0065
Months_Inactive_12_mon	1	44.5264	<.0001
Months_on_book	1	3.0171	0.0824
REP_Avg_Utilization_Ratio	1	13.6050	0.0002
REP_Card_Category	1	0.0697	0.7918
REP_Credit_Limit	1	12.0622	0.0005
REP_Education_Level	5	4.6055	0.4659
REP_Income_Category	4	2.2559	0.6888
REP_LOG_Avg_Open_To_Buy	1	17.4142	<.0001
REP_LOG_Total_Trans_Amt	1	165.5661	<.0001
REP_Marital_Status	2	9.3814	0.0092
Total_Amt_Chng_Q4_Q1	1	18.8235	<.0001
Total_Ct_Chng_Q4_Q1	1	45.3260	<.0001
Total_Relationship_Count	1	53.5444	<.0001
Total_Revolving_Bal	1	4.1943	0.0406
Total_Trans_Ct	1	267.2606	<.0001

7.2.2 Forward Regression

Results - Node: Forward Regression Diagram: Final_Project_SAS

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition_Flag		_AIC_	Akaike's Information Crite...	1025.497	.	.
Attrition_Flag		_ASE_	Average Squared Error	0.094531	0.101492	.
Attrition_Flag		_AVERR_	Average Error Function	0.307585	0.324382	.
Attrition_Flag		_DFE_	Degrees of Freedom for E...	1616	.	.
Attrition_Flag		_DFM_	Model Degrees of Freedom	12	.	.
Attrition_Flag		_DFT_	Total Degrees of Freedom	1628	.	.
Attrition_Flag		_DIV_	Divisor for ASE	3256	3252	.
Attrition_Flag		_ERR_	Error Function	1001.497	1054.892	.
Attrition_Flag		_FPE_	Final Prediction Error	0.095935	.	.
Attrition_Flag		_MAX_	Maximum Absolute Error	0.9914	0.995829	.
Attrition_Flag		_MSE_	Mean Square Error	0.095233	0.101492	.
Attrition_Flag		_NOBS_	Sum of Frequencies	1628	1626	.
Attrition_Flag		_NW_	Number of Estimate Weig...	12	.	.
Attrition_Flag		_RASE_	Root Average Sum of Squ...	0.307458	0.318578	.
Attrition_Flag		_RFPE_	Root Final Prediction Error	0.309733	.	.
Attrition_Flag		_RMSE_	Root Mean Squared Error	0.308598	0.318578	.
Attrition_Flag		_SBC_	Schwarz's Bayesian Criter...	1090.238	.	.
Attrition_Flag		_SSE_	Sum of Squared Errors	307.7919	330.0524	.
Attrition_Flag		_SUMW_	Sum of Case Weights Ti...	3256	3252	.
Attrition_Flag		_MISC_	Misclassification Rate	0.132064	0.140221	.

The validation for Target variable Attrition Flag has a misclassification Rate of 0.140221.

Results - Node: Forward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	Total_Trans_Ct	1	1	453.4686	<.0001	1721.6
2	REP_LOG_Total_Trans_Amt	1	2	249.1509	<.0001	1449.2
3	Total_Ct_Chng_04_01	1	3	128.8265	<.0001	1315.9
4	Total_Revolving_Bal	1	4	109.7476	<.0001	1210.2
5	Total_Relationship_Count	1	5	59.5112	<.0001	1135.8
6	Contacts_Count_12_mon	1	6	49.2130	<.0001	1082.3
7	Months_Inactive_12_mon	1	7	48.9575	<.0001	1060.7
8	Gender	1	8	24.5786	<.0001	1071.0
9	Total_Amt_Chng_04_01	1	9	22.5296	<.0001	1062.2
10	REP_Marital_Status	2	10	10.0492	0.0066	1054.9
11	Months_on_book	1	11	6.4238	0.0113	1065.4

The selected model, based on the error rate for the validation data, is the model trained in Step 10. It consists of the following effects:

Intercept Contacts_Count_12_mon Gender Months_Inactive_12_mon REP_LOG_Total_Trans_Amt REP_Marital_Status Total_Amt_Chng_04_01 Total_Ct_Chng_04_01 Total_Relationship_Count Total_Revolvir

And the intercept which was not visible in screenshot was Total_Trans_Ct.

Results - Node: Forward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	21.6511	2.0428	112.33	<.0001		999.000
Contacts_Count_12_mon		1	-0.5252	0.0775	45.98	<.0001	-0.3271	0.591
Gender	F	1	-0.3920	0.0824	22.64	<.0001		0.676
Months_Inactive_12_mon		1	-0.5681	0.0853	44.33	<.0001	-0.3107	0.567
REP_LOG_Total_Trans_Amt		1	-4.3674	0.3306	174.48	<.0001	-1.5916	0.013
REP_Marital_Status	Divorced	1	-0.1816	0.2061	0.78	0.3784		0.834
REP_Marital_Status	Married	1	0.3510	0.1328	6.98	0.0082		1.421
Total_Amt_Chng_Q4_Q1		1	1.8857	0.4206	20.10	<.0001	0.2319	6.591
Total_Ct_Chng_Q4_Q1		1	2.5393	0.3751	45.83	<.0001	0.3388	12.671
Total_Relationship_Count		1	0.3954	0.0550	51.60	<.0001	0.3455	1.485
Total_Revolving_Bal		1	0.000803	0.000090	79.56	<.0001	0.4016	1.001
Total_Trans_Ct		1	0.1925	0.0115	280.96	<.0001	2.4126	1.212

Forward regression analysis Based on valid chi-square.

As the highest valid chi-square for forward regression model is with the parameter Total_Trans_Ct that is 280.96. As a result, it is the most important variable based on valid chi-square followed by REP_LOG_Total_Trans_Amt which is 174.48 (Second important variable to be look after). Third important variable is Intercept which is 112.33 and least valid chi square variable is REP_Marital_Status divorced which is 0.78

Results - Node: Forward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Contacts_Count_12_mon	1	45.9774	<.0001
Gender	1	22.6425	<.0001
Months_Inactive_12_mon	1	44.3268	<.0001
REP_LOG_Total_Trans_Amt	1	174.4793	<.0001
REP_Marital_Status	2	9.9435	0.0069
Total_Amt_Chng_Q4_Q1	1	20.0980	<.0001
Total_Ct_Chng_Q4_Q1	1	45.8341	<.0001
Total_Relationship_Count	1	51.5996	<.0001
Total_Revolving_Bal	1	79.5593	<.0001
Total_Trans_Ct	1	280.9578	<.0001

Results - Node: Forward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

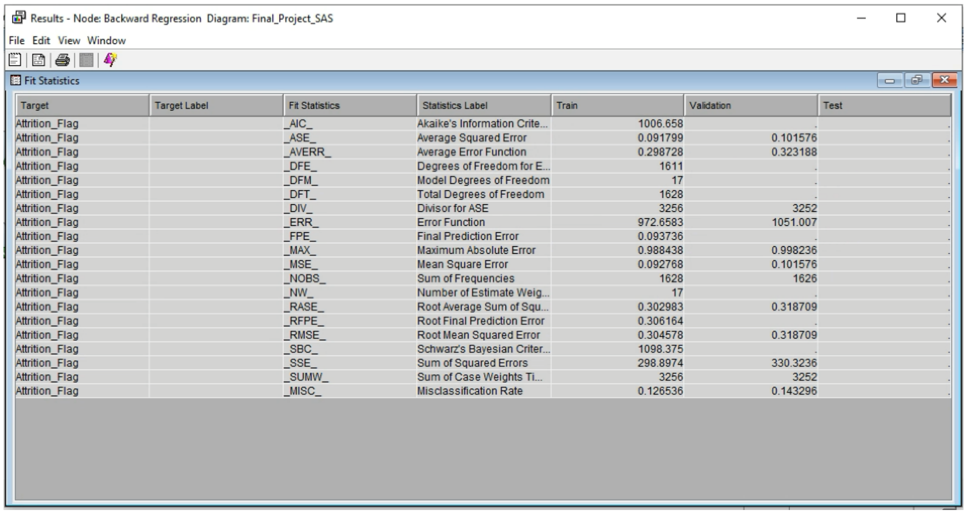
Odds Ratio Estimates

Effect		Point Estimate
Contacts_Count_12_mon		0.591
Gender	F vs M	0.457
Months_Inactive_12_mon		0.567
REP_LOG_Total_Trans_Amt		0.013
REP_Marital_Status	Divorced vs Single	0.988
REP_Marital_Status	Married vs Single	1.683
Total_Amt_Chng_Q4_Q1		6.591
Total_Ct_Chng_Q4_Q1		12.671
Total_Relationship_Count		1.485
Total_Revolving_Bal		1.001
Total_Trans_Ct		1.212

Odds ratio for forward regression is as follows.

- Male has 54% chance of churning compared to female.
- Contacts having count 12 months has 41% less chance of churning.
- People who are divorced are more likely to churn by 1.2% compared to single people.
- Total amounts change in Q4-Q1 increases with 559% of churning.
- With every 1 unit increase in relationship counts, the chance of person churning increases by 48.5%.

7.2.1 Backward Regression



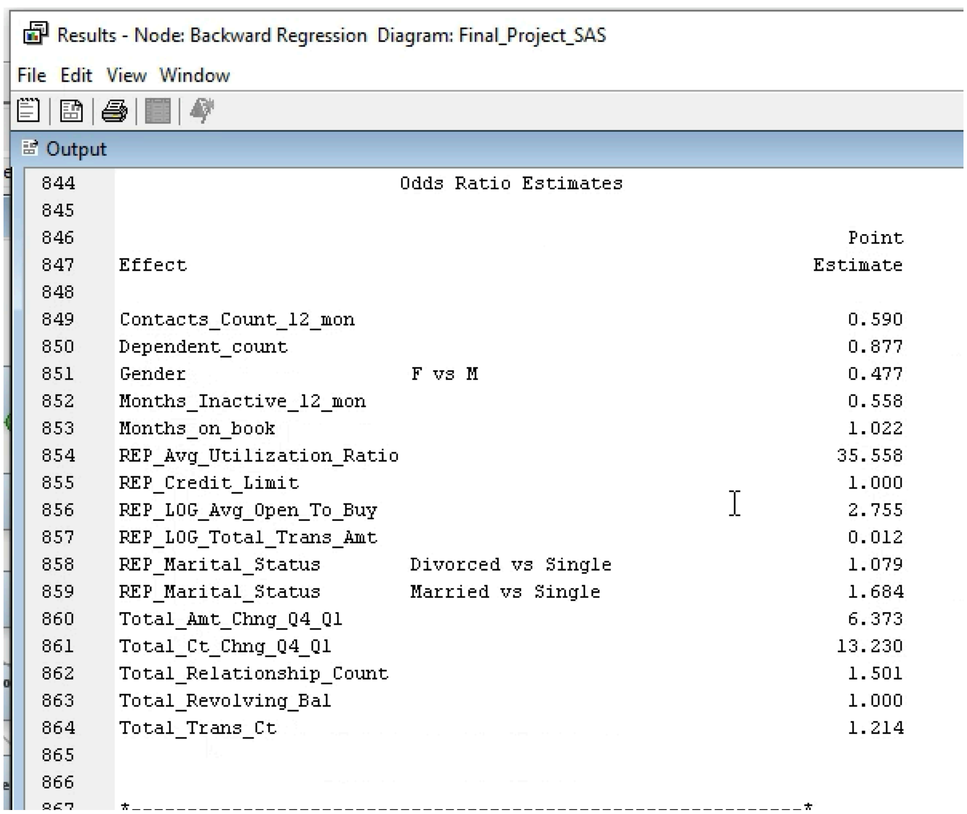
Results - Node: Backward Regression Diagram: Final_Project_SAS

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition_Flag		_AIC_	Akaike's Information Crite...	1006.658	.	.
Attrition_Flag		_ASE_	Average Squared Error	0.091799	0.101576	.
Attrition_Flag		_AVERR_	Average Error Function	0.298728	0.323188	.
Attrition_Flag		_DFE_	Degrees of Freedom for E...	1611	.	.
Attrition_Flag		_DFM_	Model Degrees of Freedom	17	.	.
Attrition_Flag		_DFT_	Total Degrees of Freedom	1628	.	.
Attrition_Flag		_DIV_	Divisor for ASE	3256	3252	.
Attrition_Flag		_ERR_	Error Function	972.6583	1051.007	.
Attrition_Flag		_FPE_	Final Prediction Error	0.093736	.	.
Attrition_Flag		_MAX_	Maximum Absolute Error	0.988438	0.998236	.
Attrition_Flag		_MSE_	Mean Square Error	0.092768	0.101576	.
Attrition_Flag		_NOBS_	Sum of Frequencies	1628	1626	.
Attrition_Flag		_NW_	Number of Estimate Weig...	17	.	.
Attrition_Flag		_RASE_	Root Average Sum of Squ...	0.302983	0.318709	.
Attrition_Flag		_RFPE_	Root Final Prediction Error	0.306164	.	.
Attrition_Flag		_RMSE_	Root Mean Squared Error	0.304578	0.318709	.
Attrition_Flag		_SBC_	Schwarz's Bayesian Criter...	1098.375	.	.
Attrition_Flag		_SSE_	Sum of Squared Errors	298.8974	330.3236	.
Attrition_Flag		_SUMW_	Sum of Case Weights Ti...	3256	3252	.
Attrition_Flag		_MISC_	Misclassification Rate	0.126536	0.143296	.

Misclassification rate for validation in Backward regression is 0.143296.



Results - Node: Backward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

```
844                         Odds Ratio Estimates
845
846                                                                    Point
847   Effect                                                        Estimate
848
849   Contacts_Count_12_mon                                            0.590
850   Dependent_count                                                  0.877
851   Gender                     F vs M                                0.477
852   Months_Inactive_12_mon                                           0.558
853   Months_on_book                                                   1.022
854   REP_Avg_Utilization_Ratio                                       35.558
855   REP_Credit_Limit                                                 1.000
856   REP_LOG_Avg_Open_To_Buy                                          2.755
857   REP_LOG_Total_Trans_Amt                                          0.012
858   REP_Marital_Status         Divorced vs Single                    1.079
859   REP_Marital_Status         Married vs Single                     1.684
860   Total_Amt_Chng_Q4_Q1                                             6.373
861   Total_Ct_Chng_Q4_Q1                                             13.230
862   Total_Relationship_Count                                         1.501
863   Total_Revolving_Bal                                              1.000
864   Total_Trans_Ct                                                   1.214
865
866
```

The results are Point estimates for the Odds Ratio Estimates for Backward Regression as follows:

- In gender, male are 52.3% likely to churn compared to female.
- People who are inactive for 12 months have 44% less chance of being churned.
- Average utilization ratio of card has 3,555% chance of churning.
- People who are married have 68.4% of churning compared to single people.
- Dependent count is 12.3% less likely to be churned.
- People who are on books for months increases by 2.2% to churn.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Contacts_Count_12_mon	1	44.4344	<.0001
Dependent_count	1	3.9151	0.0479
Gender	1	15.6982	<.0001
Months_Inactive_12_mon	1	45.2819	<.0001
Months_on_book	1	3.9326	0.0474
REP_Avg_Utilization_Ratio	1	13.3715	0.0003
REP_Credit_Limit	1	14.1632	0.0002
REP_LOG_Avg_Open_To_Buy	1	17.0804	<.0001
REP_LOG_Total_Trans_Amt	1	169.8443	<.0001
REP_Marital_Status	2	9.1124	0.0105
Total_Amt_Chng_Q4_Q1	1	18.8528	<.0001
Total_Ct_Chng_Q4_Q1	1	45.6194	<.0001
Total_Relationship_Count	1	52.4223	<.0001
Total_Revolving_Bal	1	4.4935	0.0340
Total_Trans_Ct	1	273.2680	<.0001

Results - Node: Backward Regression Diagram: Final_Project_SAS

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	13.4595	2.7776	23.48	<.0001		999.000
Contacts_Count_12_mon		1	-0.5270	0.0791	44.43	<.0001	-0.3283	0.590
Dependent_count		1	-0.1315	0.0664	3.92	0.0479	-0.0932	0.877
Gender	F	1	-0.3702	0.0934	15.70	<.0001		0.691
Months_Inactive_12_mon		1	-0.5841	0.0868	45.28	<.0001	-0.3195	0.558
Months_on_book		1	0.0222	0.0112	3.93	0.0474	0.0954	1.022
REP_Avg_Utilization_Ratio		1	3.5712	0.9766	13.37	0.0003	0.5454	35.558
REP_Credit_Limit		1	-0.00008	0.000021	14.16	0.0002	-0.3804	1.000
REP_LOG_Avg_Open_To_Buy		1	1.0134	0.2452	17.08	<.0001	0.7027	2.755
REP_LOG_Total_Trans_Amt		1	-4.4162	0.3389	169.84	<.0001	-1.6094	0.012
REP_Marital_Status	Divorced	1	-0.1231	0.2087	0.35	0.5553		0.884
REP_Marital_Status	Married	1	0.3222	0.1349	5.71	0.0169		1.380
Total_Amt_Chng_Q4_Q1		1	1.8520	0.4265	18.85	<.0001	0.2277	6.373
Total_Ct_Chng_Q4_Q1		1	2.5825	0.3824	45.62	<.0001	0.3445	13.230
Total_Relationship_Count		1	0.4059	0.0561	52.42	<.0001	0.3547	1.501
Total_Revolving_Bal		1	0.000361	0.000170	4.49	0.0340	0.1804	1.000
Total_Trans_Ct		1	0.1936	0.0117	273.27	<.0001	2.4260	1.214

Backward regression analysis Based on valid chi-square.

As the highest valid chi-square for backward regression model is with the parameter Total_Trans_Ct that is 273.27. As a result, it is the most important variable based on valid chi-square followed by REP_LOG_Total_Trans_Amt which is 169.84 (Second important variable to be look after). Third important variable is Total_Relationship_Count which is 52.42 and least valid chi square variable is REP_Marital_Status Divorced which is 0.35.

Results - Node: Backward Regression Diagram: Final_Project_SAS

Edit View Window

Output

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	REP_Card_Category	1	18	0.0697	0.7918	1060.5
2	REP_Income_Category	4	17	2.2539	0.6892	1057.3
3	Customer_Age	1	16	0.4339	0.5101	1057.3
4	REP_Education_Level	5	15	4.8681	0.4322	1051.0

The selected model, based on the error rate for the validation data, is the model trained in Step 4. It consists of the following effects:

Intercept Contacts_Count_12_mon Dependent_count Gender Months_Inactive_12_mon Months_on_book REP_Avg_Utilization_Ratio REP_Credit_Limit REP_LOG_Avg_Open_To_Buy REP_LOG_Total_Trans_Amt REP_Ma
Total_Relationship_Count Total_Revolving_Bal Total_Trans_Ct

7.2.4 Stepwise Regression

Results - Node: Stepwise Regression Diagram: Final_Project_SAS

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Attrition_Flag		_AIC_	Akaike's Information Crite...	1025.497	.	.
Attrition_Flag		_ASE_	Average Squared Error	0.094531	0.101492	.
Attrition_Flag		_AVERR_	Average Error Function	0.307585	0.324382	.
Attrition_Flag		_DFE_	Degrees of Freedom for E...	1616	.	.
Attrition_Flag		_DFM_	Model Degrees of Freedom	12	.	.
Attrition_Flag		_DFT_	Total Degrees of Freedom	1628	.	.
Attrition_Flag		_DIV_	Divisor for ASE	3256	3252	.
Attrition_Flag		_ERR_	Error Function	1001.497	1054.892	.
Attrition_Flag		_FPE_	Final Prediction Error	0.095935	.	.
Attrition_Flag		_MAX_	Maximum Absolute Error	0.9914	0.995829	.
Attrition_Flag		_MSE_	Mean Square Error	0.095233	0.101492	.
Attrition_Flag		_NOBS_	Sum of Frequencies	1628	1626	.
Attrition_Flag		_NW_	Number of Estimate Weig...	12	.	.
Attrition_Flag		_RASE_	Root Average Sum of Squ...	0.307458	0.318578	.
Attrition_Flag		_RFPE_	Root Final Prediction Error	0.309733	.	.
Attrition_Flag		_RMSE_	Root Mean Squared Error	0.308598	0.318578	.
Attrition_Flag		_SBC_	Schwarz's Bayesian Criter...	1090.238	.	.
Attrition_Flag		_SSE_	Sum of Squared Errors	307.7919	330.0524	.
Attrition_Flag		_SUMW_	Sum of Case Weights Ti...	3256	3252	.
Attrition_Flag		_MISC_	Misclassification Rate	0.132064	0.140221	.

Misclassification rate for validation in Stepwise regression is 0.140221.

Results - Node: Stepwise Regression Diagram: Final_Project_SAS

File Edit View Window

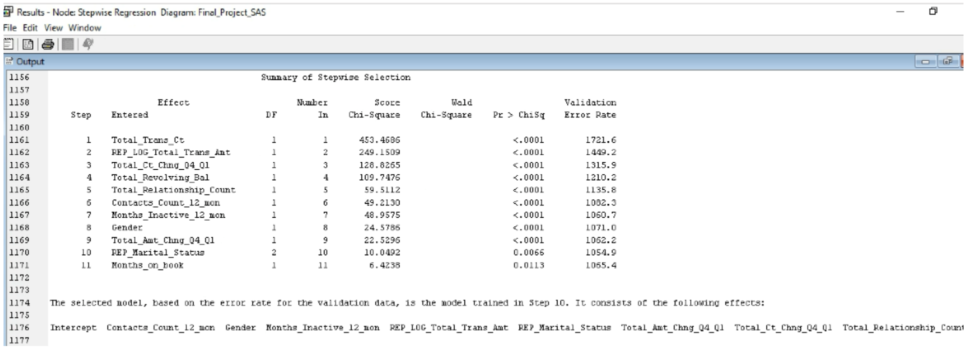
Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	21.6511	2.0428	112.33	<.0001		999.000
Contacts_Count_12_mon		1	-0.5252	0.0775	45.98	<.0001	-0.3271	0.591
Gender	F	1	-0.3920	0.0824	22.64	<.0001		0.676
Months_Inactive_12_mon		1	-0.5681	0.0853	44.33	<.0001	-0.3107	0.567
REP_LOG_Total_Trans_Amt		1	-4.3674	0.3306	174.48	<.0001	-1.5916	0.013
REP_Marital_Status	Divorced	1	-0.1816	0.2061	0.78	0.3784		0.834
REP_Marital_Status	Married	1	0.3510	0.1328	6.98	0.0082		1.421
Total_Amt_Chng_Q4_Q1		1	1.8857	0.4206	20.10	<.0001	0.2319	6.591
Total_Ct_Chng_Q4_Q1		1	2.5393	0.3751	45.83	<.0001	0.3388	12.671
Total_Relationship_Count		1	0.3954	0.0550	51.60	<.0001	0.3455	1.485
Total_Revolving_Bal		1	0.000803	0.000090	79.56	<.0001	0.4016	1.001
Total_Trans_Ct		1	0.1925	0.0115	280.96	<.0001	2.4126	1.212

Stepwise regression analysis Based on valid chi-square.

As the highest valid chi-square for Stepwise model is with the parameter Total_Trans_Ct that is 280.96; it is the most important variable based on valid chi-square followed by REP_LOG_Total_Trans_Amt which is 174.48 (Second important variable to be look after). Third important variable is Intercept which is 112.33 and least valid chi square variable is REP_Marital_Status Divorced which is 0.78.



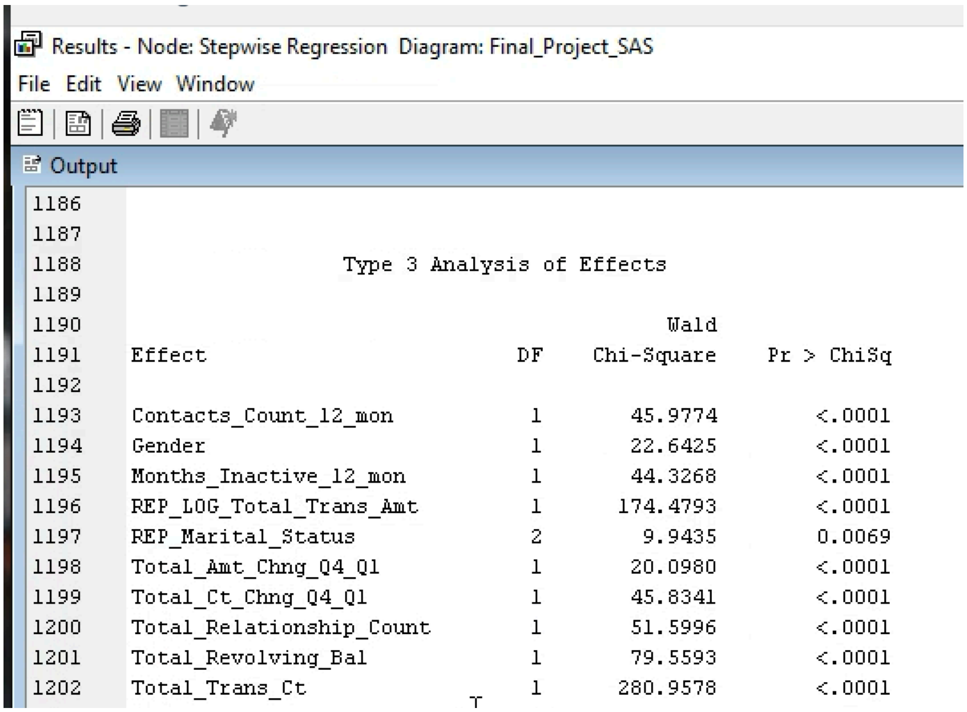
Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	Total_Trans_Ct	1	1	453.4686		<.0001	1721.6
2	REP_LOG_Total_Trans_Amt	1	2	249.1509		<.0001	1449.2
3	Total_Ct_Chng_Q4_Q1	1	3	128.8265		<.0001	1315.9
4	Total_Revolving_Bal	1	4	109.7476		<.0001	1210.2
5	Total_Relationship_Count	1	5	59.5112		<.0001	1135.8
6	Contacts_Count_12_mon	1	6	49.2130		<.0001	1082.3
7	Months_Inactive_12_mon	1	7	48.9575		<.0001	1060.7
8	Gender	1	8	24.5786		<.0001	1071.0
9	Total_Amt_Chng_Q4_Q1	1	9	22.5296		<.0001	1062.2
10	REP_Marital_Status	2	10	10.0492		0.0066	1054.9
11	Months_on_book	1	11	6.4238		0.0113	1065.4

The selected model, based on the error rate for the validation data, is the model trained in Step 10. It consists of the following effects:

Intercept Contacts_Count_12_mon Gender Months_Inactive_12_mon REP_LOG_Total_Trans_Amt REP_Marital_Status Total_Amt_Chng_Q4_Q1 Total_Ct_Chng_Q4_Q1 Total_Relationship_Count

And the intercept which was not visible in screenshot are Total_Revolving_Bal and Total_Trans_Ct.



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Contacts_Count_12_mon	1	45.9774	<.0001
Gender	1	22.6425	<.0001
Months_Inactive_12_mon	1	44.3268	<.0001
REP_LOG_Total_Trans_Amt	1	174.4793	<.0001
REP_Marital_Status	2	9.9435	0.0069
Total_Amt_Chng_Q4_Q1	1	20.0980	<.0001
Total_Ct_Chng_Q4_Q1	1	45.8341	<.0001
Total_Relationship_Count	1	51.5996	<.0001
Total_Revolving_Bal	1	79.5593	<.0001
Total_Trans_Ct	1	280.9578	<.0001

Odds Ratio Estimates

Effect		Point Estimate
Contacts_Count_12_mon		0.591
Gender	F vs M	0.457
Months_Inactive_12_mon		0.567
REP_LOG_Total_Trans_Amt		0.013
REP_Marital_Status	Divorced vs Single	0.988
REP_Marital_Status	Married vs Single	1.683
Total_Amt_Chng_Q4_Q1		6.591
Total_Ct_Chng_Q4_Q1		12.671
Total_Relationship_Count		1.485
Total_Revolving_Bal		1.001
Total_Trans_Ct		1.212

Odds ratios estimate for Stepwise regression is as follows:

- Female has 54% chance of being churned compared to male.
- Customers who are inactive for 12 months on books have 43% chance of churn.
- People who are single have a churn rate of 1.2% than people who are divorced.
- Contacts count in 12 months has 41% less chance of churning.
- Total relationship counts increases churning by 49%.

Internal event (customer activity) variable - if the account is closed then 1 else 0.

Overall forward regression and stepwise regression has the best misclassification rate of 0.140221 in the regression model compared to other models.

7.2.5 Comparison of regression models without transformation

Instead of performing log transformation on variables with high skewness statistic, the cap and floor technique was used on them. The variables included in cap and floor were Avg_Open_To_Buy and Total_Trans_Amt. All regression models were executed with these changes and the performance was reported respectively. Full regression and backward regression models showed an improvement in the misclassification rate while the other models did not improve. The misclassification rate performance is as follows.

Sr. No	Regression model	Misclassification Rate (Log Transformation)(Validation Data)	Misclassification Rate (Cap n Floor)(Validation Data)
1	Full regression	0.143296	0.13899
2	Forward Regression	0.140221.	0.14760
3	Backward Regression	0.143296	0.13899
4	Stepwise Regression	0.140221	0.14760

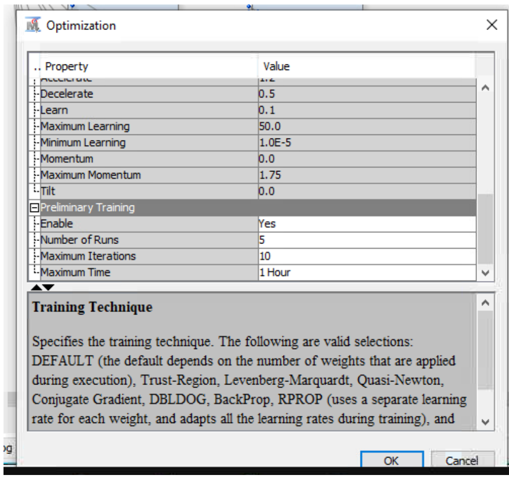
The chi-square values for respective variables remained the same in cap n floor as well since the variables that had been transformed did not create significant impact on the target variable. The same thing goes with odds ratio comparison which was similar for the full regression and forward regression as mentioned in the above sections.

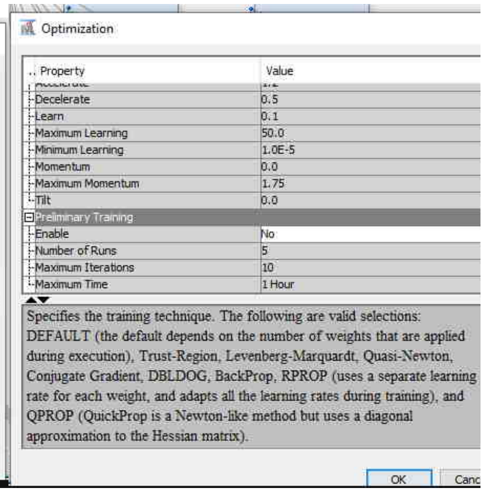
7.3 Neural Networks

A Neural Network model is as similar as regression but with flexible addition to model virtually any association between input and target variables. Our dataset consists of consistent data, so we do not have to worry about imputing or transforming variables for the requirement of models. To approximate any continuous association between inputs and the target, 9 neural network models with different configurations will be applied to our data. The best model will be filtered out by comparing the MISC (Misclassification rate) and the model selection criteria will be as default profit and loss. We will initiate the analysis by comparing the model to find either preliminary training can be helpful to achieve high accuracy and apply the configurations throughout the later models. Maximum iterations will be 50 for all the models expect one with 100 to compare the result.

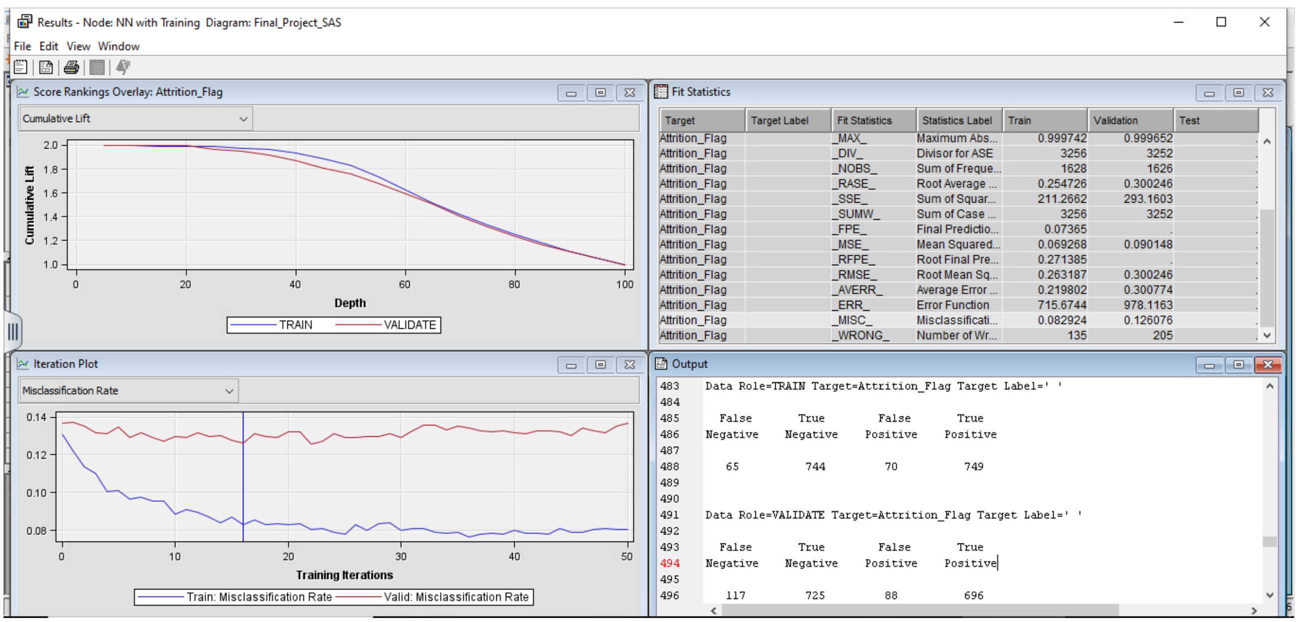
7.3.1 Model with preliminary training enable vs disable

Applying settings:





Results:



Outcome: After the comparison of above models, model with training enabled has the MISC of 0.126076 while the model with training disabled has MISC of 0.108856. Thus, we should disable the training for further models to get better accuracy.

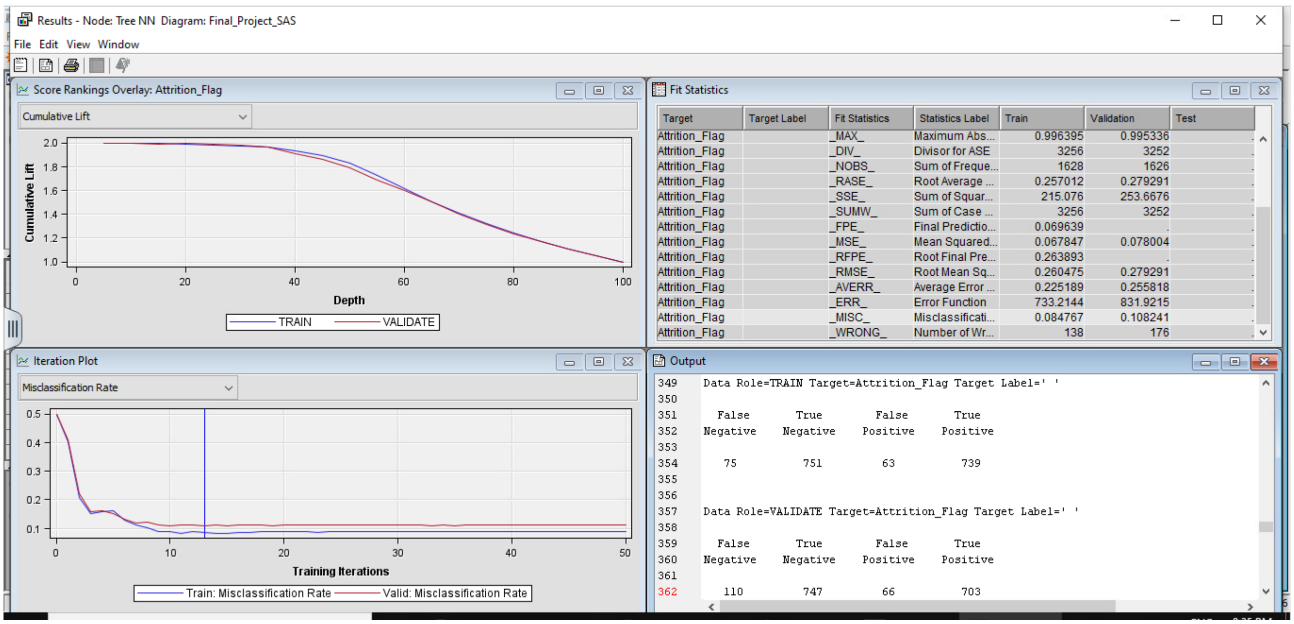
First model Misclassification rate for negative class: 0.138954

Second model Misclassification rate for negative class: 0.117788

7.3.2 Decision tree neural network

Since we already have our decision tree, we can now connect our node to the best tree we have and apply the settings according to above outcomes.

Result:



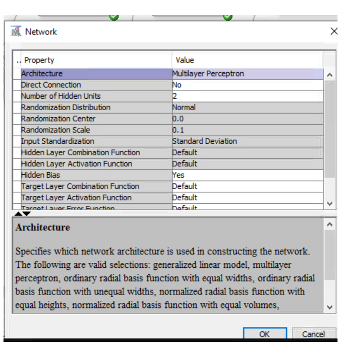
Outcome: The new ASE for this model is 0.108241. By far the best neural network model.

Misclassification rate for negative class: 0.128354

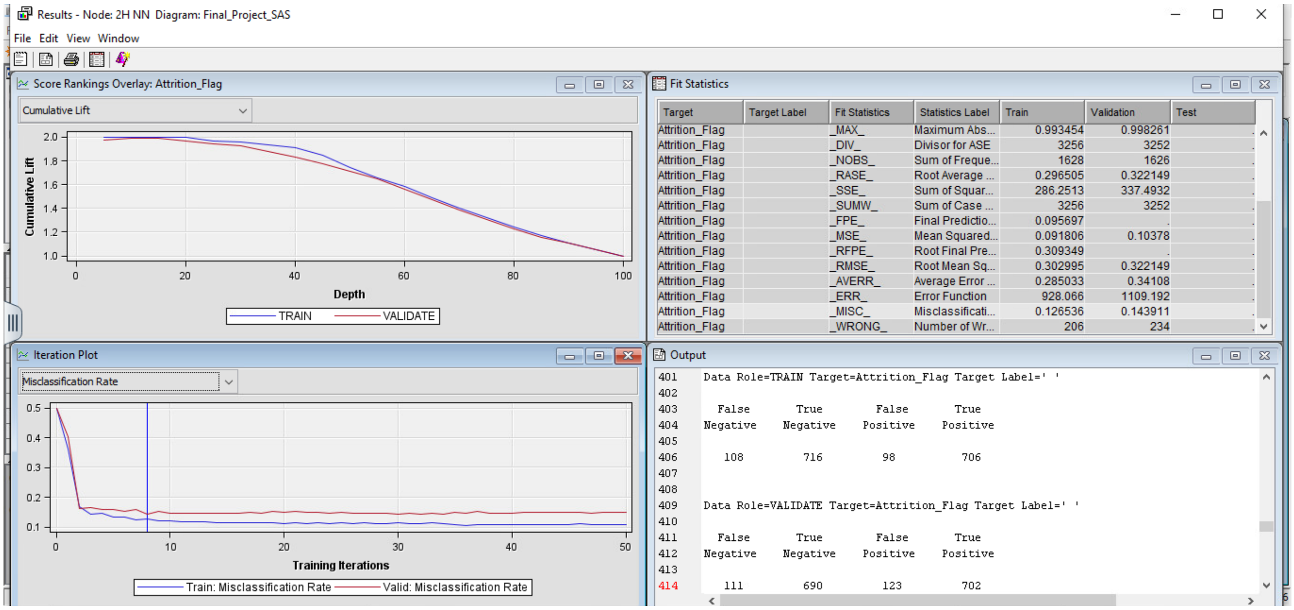
7.3.3 2 Hidden unit's neural network

As neural networks work better due to hidden layers, to get a better result, we will start by assigning 2 hidden units to the data partition.

Applying setting:



Result:

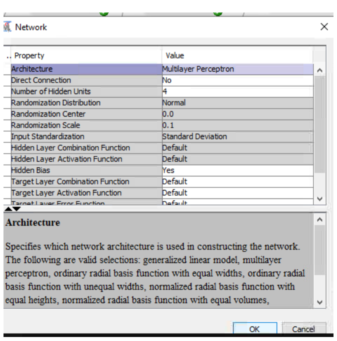


Outcome: The ASE for our model with 2 hidden units is 0.143911. Not better than previous models so we will just add up the hidden units until it gives the constant result, or the ASE rise again.

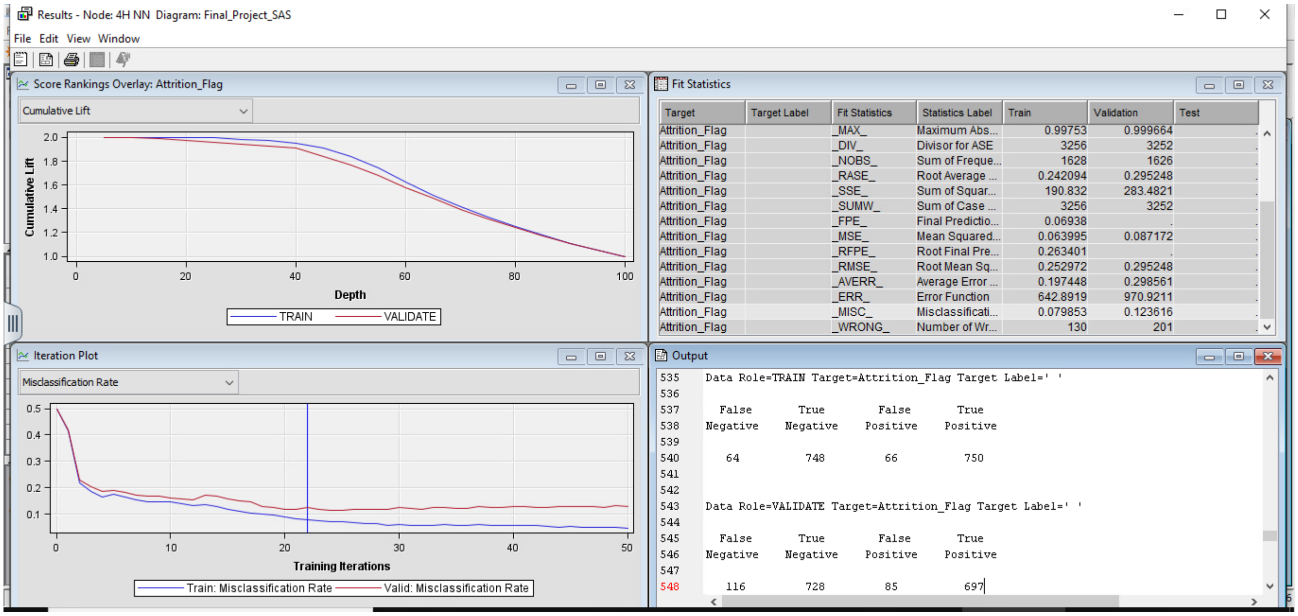
Misclassification rate for negative class: 0.138576

7.3.4 4 Hidden unit's neural network

Applying setting for 4 hidden unit:



Result:

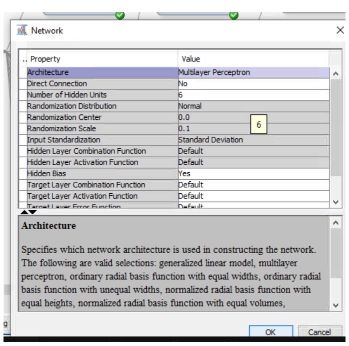


Outcome: For this model with 4 hidden units, the ASE is 0.123616 - Better than 2 hidden units but not better than model connected with decision tree. Thus, we will add more hidden units.

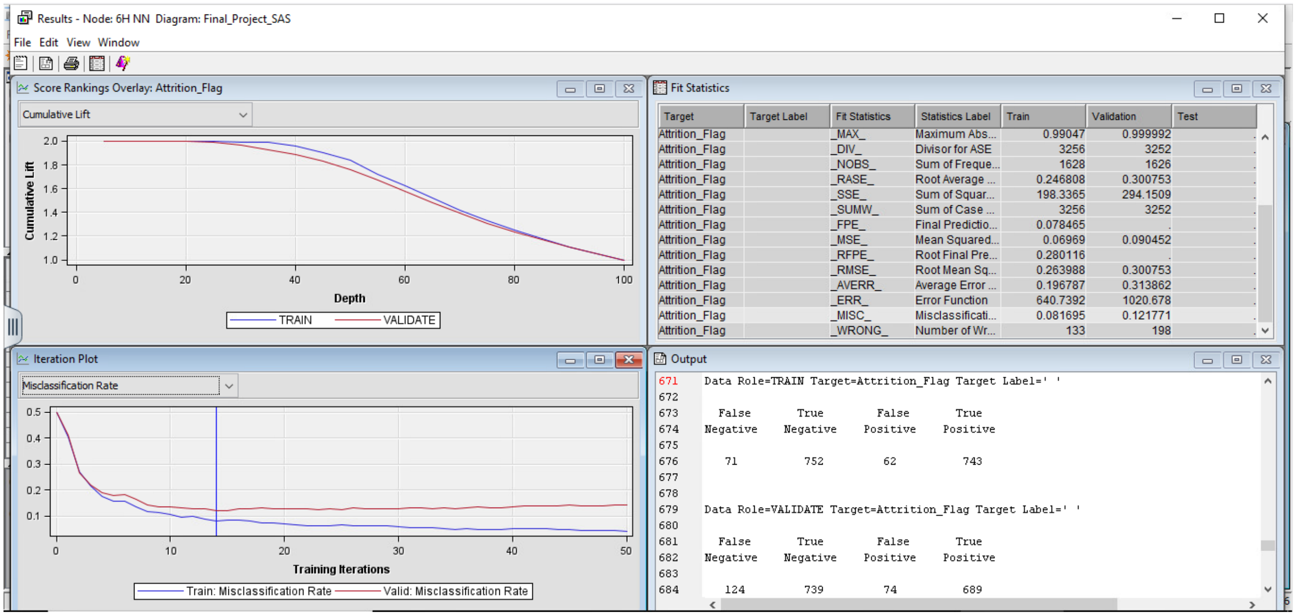
Misclassification rate for negative class: 0.137440

7.3.5 6 Hidden unit's neural network

Applying setting for 6 hidden units:



Result:



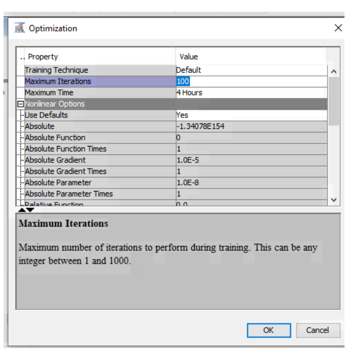
Outcome: This model, with 6 hidden units, has an ASE of 0.121771, which is lesser than the previous model with lower units. Thus, we will settle down at model with 6 hidden units because after this model the MISC stays constant for every interval.

Misclassification rate for negative class: 0.143684

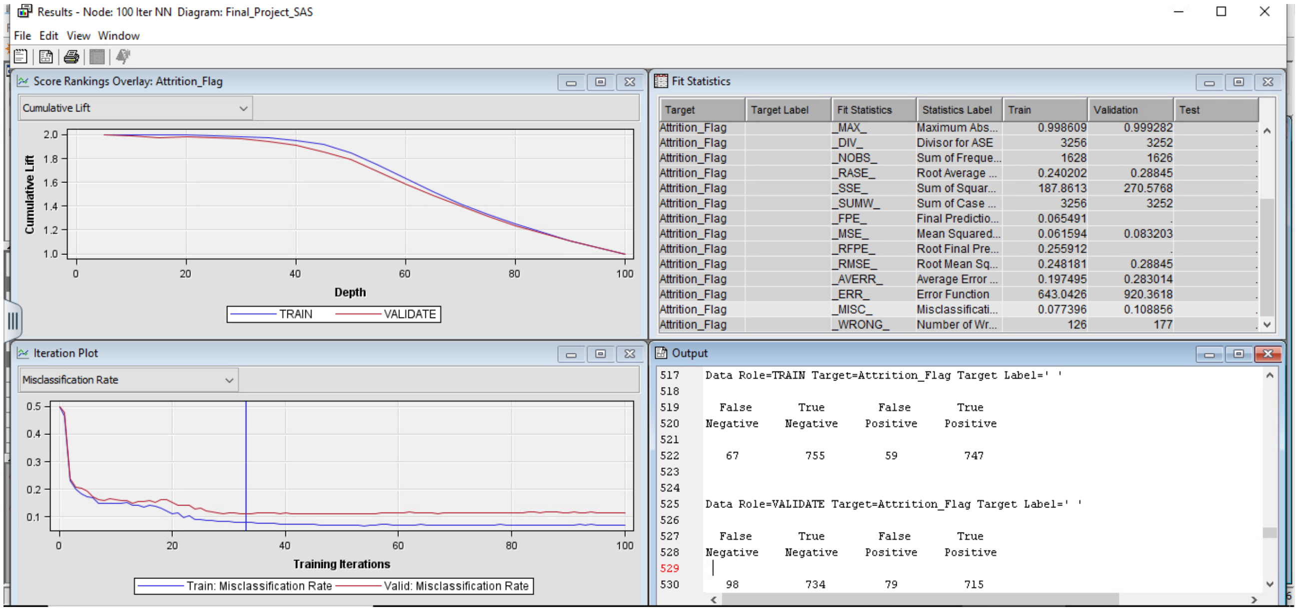
7.3.6 100 Iteration neural network

To check if the accuracy increases with increase in iteration value, we set it to 100 from 50 with default hidden units.

Applying setting for maximum iteration to be 100:



Result:

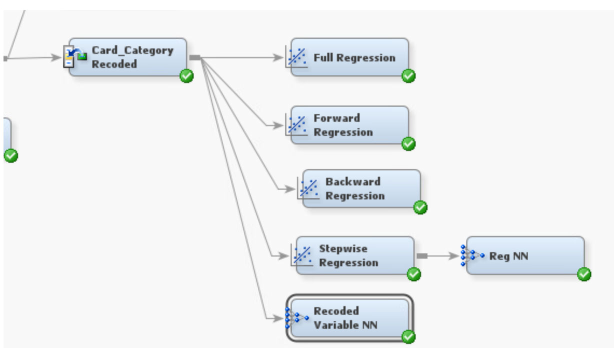


Outcome: Our new model with maximum iteration has an MISC of 0.108856. Thus, the same as the initial model with no training but better than the model with hidden units higher than 3 with 50 iterations.

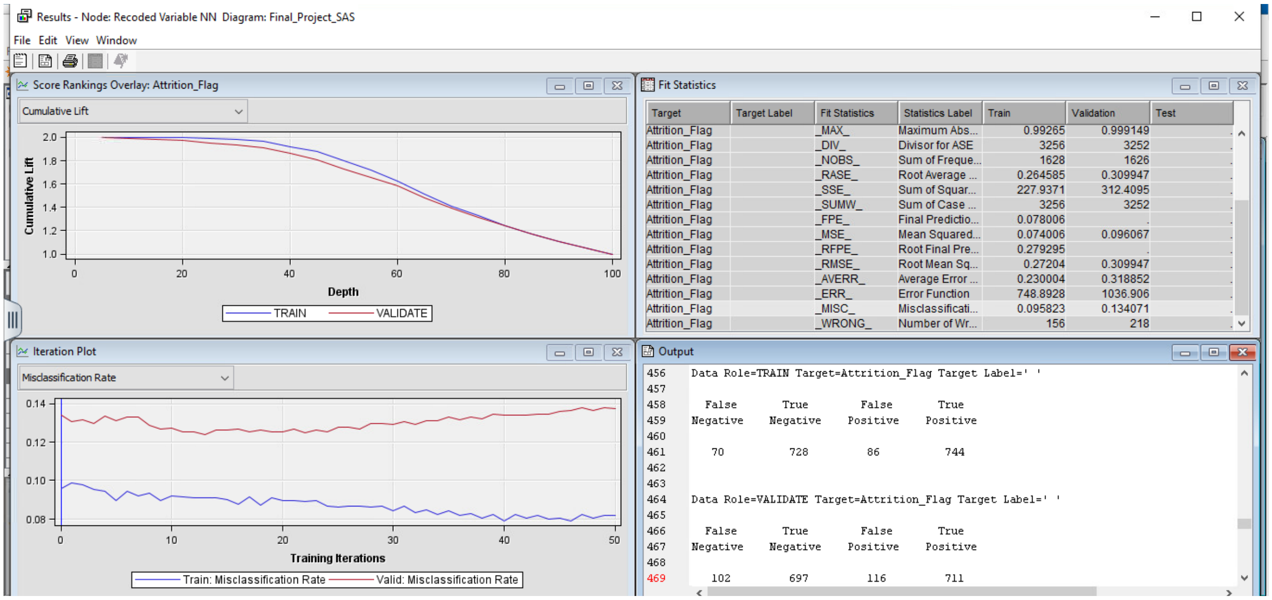
Misclassification rate for negative class: 0.117788

7.3.7 Recoded variable neural network

Connecting model to existing recoded card category:



Result:



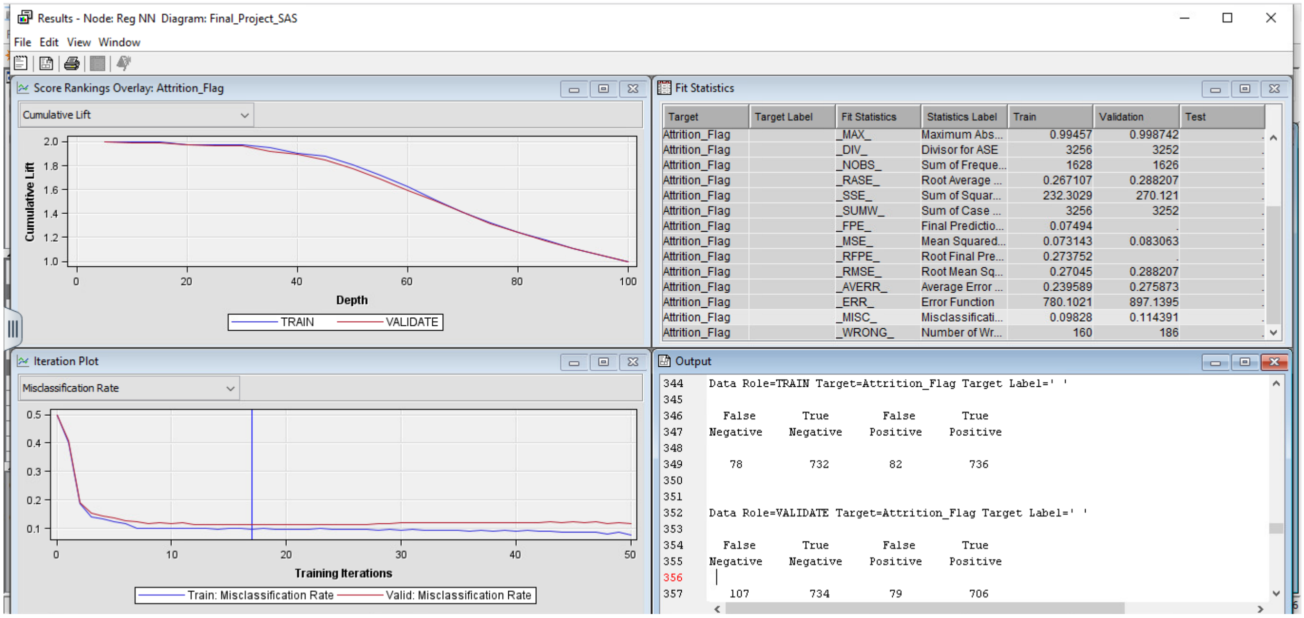
Outcome: Our new neural network model connected with the recoded card category has an MISC of 0.134071. Thus, the second highest MISC till now after model with 2 hidden units.

Misclassification rate for negative class: 0.127659

7.3.8 Stepwise regression neural network

In search of better accuracy, we then connected the neural network model with the best regression model which is stepwise (Can be seen in figure above).

Result:



Outcome: This model with node connected to stepwise regression has the MISC of 0.114391. Thus, the second-best model out of all the neural network models in assessment of MISC.

Misclassification rate for negative class: 0.127229

7.3.9 Summary table for neural network model

Neural Network Models	Iterations	Hidden units	MISC
NN with Training	50	3	0.126076
NN without Training	50	3	0.108856
Tree NN	50	3	0.108241
2H- NN	50	2	0.143911
4H-NN	50	4	0.123616
6H-NN	50	6	0.121771
100 Iter NN	100	3	0.108856
Recoded variable NN	50	3	0.134071

Reg NN	50	3	0.114391

Overall, the best neural network model is the model connected with decision tree with an MISC of 0.108241, having iteration and hidden units of 50 and 3. And the model with the highest MISC is the model with 2 hidden units.

7.3.10 Model Assessment

Does ordinal encoding help improve performance of the best Regression / Neural Networks model?

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

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree4	3 Branch Tree	0.09779	0.05878	0.07310	0.07987
	Neural5	4H NN	0.10578	0.06615	0.08968	0.08354
	Tree3	Misclassification Tree	0.10640	0.08435	0.09951	0.08959
	Tree2	Probability Tree	0.11009	0.07142	0.08968	0.08647
	Neural3	Reg NN	0.11132	0.07710	0.10258	0.08359
	Neural2	NN with Training	0.11132	0.06367	0.08722	0.08503
	Neural4	Tree NN	0.11501	0.07162	0.09275	0.08207
	Tree	Decision Tree	0.11747	0.05837	0.07740	0.08864
	Neural6	6H NN	0.11870	0.07157	0.10074	0.08826
	Neural	NN without Training	0.11993	0.07217	0.09337	0.08983
	Neural9	100 Iter NN	0.11993	0.07217	0.09337	0.08983
	Neural7	Recoded Variable NN	0.12669	0.06764	0.09275	0.09249
	Reg2	Stepwise Regression	0.14084	0.10135	0.14066	0.09982
	Reg4	Forward Regression	0.14084	0.10135	0.14066	0.09982
	Neural8	2H NN	0.14330	0.08521	0.11364	0.09915
	Reg3	Backward Regression	0.14822	0.09934	0.13636	0.10102

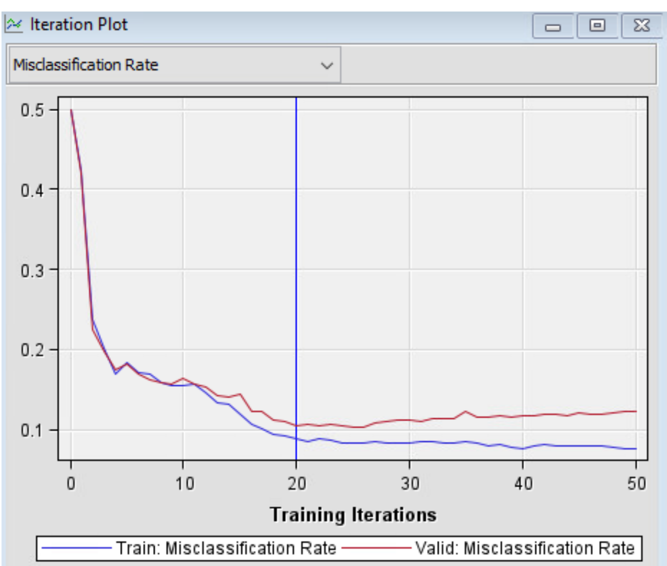
Among the Neural Networks and Regression models, the neural networks model with 4 Hidden layers provided the lowest misclassification rate and hence we will compare this model with the same configuration for ordinal encoded variables and one hot encoded variable and compare results respectively.

Neural Networks model with 4 Hidden Layers – Model without ordinal Encoding:

Fit Statistics

Target=Attrition_Flag Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	1628.00	.
DFE	Degrees of Freedom for Error	1491.00	.
DFM	Model Degrees of Freedom	137.00	.
_NW_	Number of Estimated Weights	137.00	.
AIC	Akaike's Information Criterion	981.84	.
SBC	Schwarz's Bayesian Criterion	1720.97	.
ASE	Average Squared Error	0.07	0.08
MAX	Maximum Absolute Error	0.99	1.00
DIV	Divisor for ASE	3256.00	3252.00
NOBS	Sum of Frequencies	1628.00	1626.00
RASE	Root Average Squared Error	0.26	0.29
SSE	Sum of Squared Errors	215.39	271.66
SUMW	Sum of Case Weights Times Freq	3256.00	3252.00
FPE	Final Prediction Error	0.08	.
MSE	Mean Squared Error	0.07	0.08
RFPE	Root Final Prediction Error	0.28	.
RMSE	Root Mean Squared Error	0.27	0.29
AVERR	Average Error Function	0.22	0.28
ERR	Error Function	707.84	905.26
MISC	Misclassification Rate	0.09	0.11
WRONG	Number of Wrong Classifications	146.00	172.00



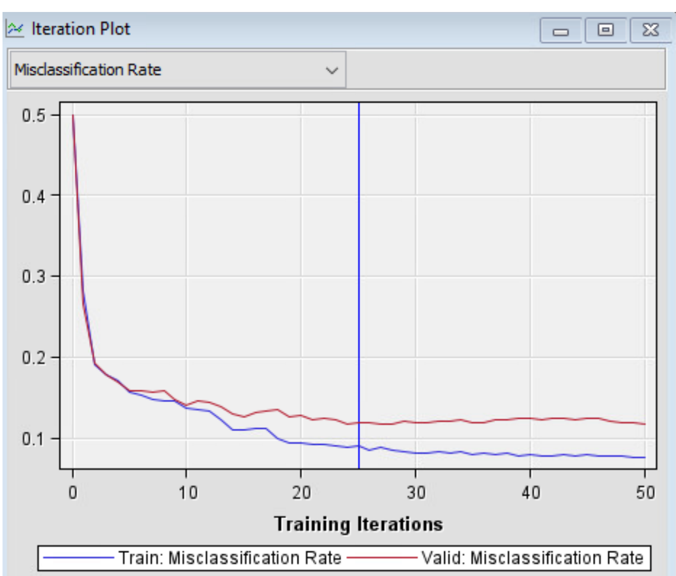
The neural networks model with 4 hidden units was the best model among the regression and neural networks models. With a misclassification rate slightly higher than the overall best model. Although neural networks lack explain ability, they can be very powerful machine learning approaches. In our case, the misclassification rate converged to the lowest at around 20 iterations after which it was overfitting. All categorical variables had been one hot encoded which resulted in a model with a lot of variables. This can be computationally expensive as well.

Neural Networks model with 4 Hidden Layers – Model with ordinal Encoding:

Fit Statistics

Target=Attrition_Flag Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	1628.00	.
DFE	Degrees of Freedom for Error	1527.00	.
DFM	Model Degrees of Freedom	101.00	.
_NW_	Number of Estimated Weights	101.00	.
AIC	Akaike's Information Criterion	946.39	.
SBC	Schwarz's Bayesian Criterion	1491.29	.
ASE	Average Squared Error	0.07	0.09
MAX	Maximum Absolute Error	0.99	1.00
DIV	Divisor for ASE	3256.00	3252.00
NOBS	Sum of Frequencies	1628.00	1626.00
RASE	Root Average Squared Error	0.26	0.30
SSE	Sum of Squared Errors	221.24	295.11
SUMW	Sum of Case Weights Times Freq	3256.00	3252.00
FPE	Final Prediction Error	0.08	.
MSE	Mean Squared Error	0.07	0.09
RFPE	Root Final Prediction Error	0.28	.
RMSE	Root Mean Squared Error	0.27	0.30
AVERR	Average Error Function	0.23	0.31
ERR	Error Function	744.39	995.64
MISC	Misclassification Rate	0.09	0.12
WRONG	Number of Wrong Classifications	149.00	193.00



Although, this model included relatively lower number of variables by a factor of 10 which reduced the dimensionality drastically, performance of the model was lower as compared to the one without ordinal encoding. Ordinal encoding may sound logical from a domain perspective, but while modelling, independent flag variables for each class created more impact on the target rather than ranking the respective variables. The model converged to the lowest misclassification rate at 25 iterations post which there were signs of overfitting. But in this case the extent of validation misclassification rate after reaching the minima was not as drastic as in the case of the previous model.

Results - Node: Model Comparison Diagram: Final_Project_SAS

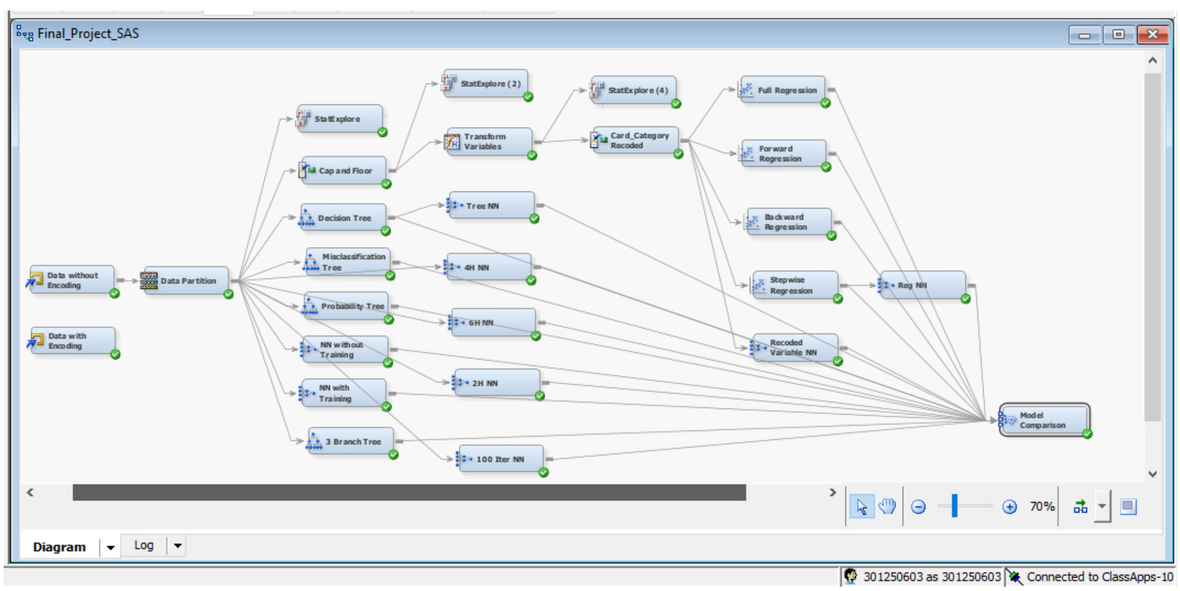
File Edit View Window

Fit Statistics

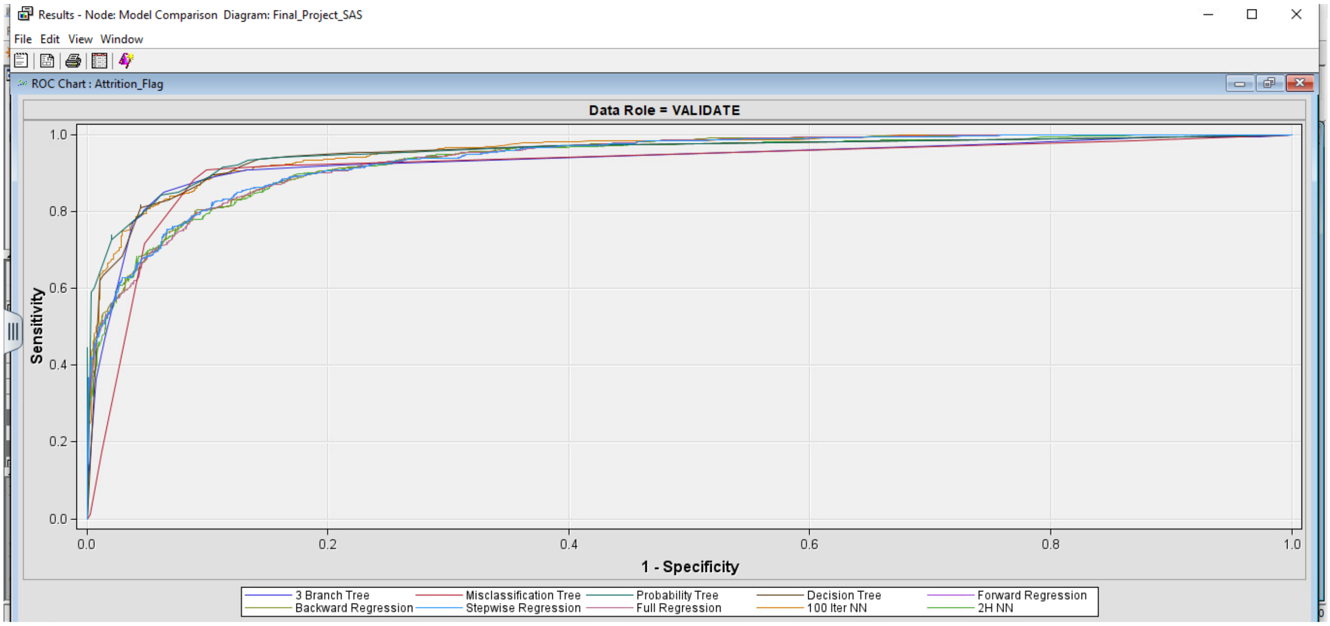
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Valid: Roc Index ▼	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Val Max Abs Err
	Neural4	Neural4	Tree NN	Attrition_Flag		0.084767	1628	0.084767	0.996395	215.076	0.96	0.066055	0.257012	3256	1628	1626	0.108241	
Y	Neural6	Neural6	6H NN	Attrition_Flag		0.064496	1628	0.064496	0.994522	151.8726	0.956	0.046644	0.215972	3256	1628	1626	0.115621	
	Neural	Neural	NN without ...	Attrition_Flag		0.077396	1628	0.077396	0.998609	187.8613	0.955	0.057697	0.240202	3256	1628	1626	0.108856	
	Neural9	Neural9	100 Iter NN	Attrition_Flag		0.077396	1628	0.077396	0.998609	187.8613	0.955	0.057697	0.240202	3256	1628	1626	0.108856	
	Tree2	Tree2	Probability ...	Attrition_Flag		0.074939	1628	0.074939	0.990909	176.4582	0.954	0.054195	0.232798	3256	1628	1626	0.099016	
	Neural3	Neural3	Reg NN	Attrition_Flag		0.09828	1628	0.09828	0.99457	232.3029	0.954	0.071346	0.267107	3256	1628	1626	0.114391	
	Neural5	Neural5	4H NN	Attrition_Flag		0.079853	1628	0.079853	0.99753	190.832	0.95	0.058609	0.242094	3256	1628	1626	0.123616	
	Tree	Tree	Decision Tr...	Attrition_Flag		0.066339	1628	0.066339	0.982005	151.7986	0.949	0.046621	0.215919	3256	1628	1626	0.105166	
	Neural2	Neural2	NN with Tra...	Attrition_Flag		0.082924	1628	0.082924	0.999742	211.2662	0.949	0.064885	0.254726	3256	1628	1626	0.126076	
	Neural7	Neural7	Recoded V...	Attrition_Flag		0.095823	1628	0.095823	0.99265	227.9371	0.942	0.070005	0.264585	3256	1628	1626	0.134071	
	Reg3	Reg3	Backward ...	Attrition_Flag		0.126536	1628	0.126536	0.988438	298.8974	0.936	0.091799	0.302983	3256	1628	1626	0.143296	
	Reg2	Reg2	Stepwise R...	Attrition_Flag		0.132064	1628	0.132064	0.9914	307.7919	0.936	0.094531	0.307458	3256	1628	1626	0.140221	
	Reg4	Reg4	Forward Re...	Attrition_Flag		0.132064	1628	0.132064	0.9914	307.7919	0.936	0.094531	0.307458	3256	1628	1626	0.140221	
	Reg	Reg	Full Regres...	Attrition_Flag		0.12715	1628	0.12715	0.991258	296.0994	0.935	0.09094	0.301562	3256	1628	1626	0.143296	
	Neural8	Neural8	2H NN	Attrition_Flag		0.126536	1628	0.126536	0.993454	286.2513	0.931	0.087915	0.296505	3256	1628	1626	0.143911	
	Tree4	Tree4	3 Branch Tr...	Attrition_Flag		0.067568	1628	0.067568	0.992157	175.9933	0.93	0.054052	0.232491	3256	1628	1626	0.107011	
	Tree3	Tree3	Misclassific...	Attrition_Flag		0.077396	1628	0.077396	0.980583	217.5717	0.922	0.066822	0.258499	3256	1628	1626	0.095941	

8. Models Comparison

All the models for this churn analytics which includes decision trees, regression and neural networks are further connected to the model comparison node to draw the best performance model. Comparison model is set up to use Misclassification for the selection statistic. The best performer will be filtered by viewing roc result and ranking according to other model assessment metrics.



ROC curve for the models



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Misclassification Rate	Valid: Roc Index ▼	Valid: Sum of Frequenc
	Neural4	Neural4	Tree NN	Attrition_Flag		0.084767	1628	0.084767	0.996395	215.076	0.066055	0.257012	3256	1628	0.108241	0.96	1
Y	Neural6	Neural6	6H NN	Attrition_Flag		0.064496	1628	0.064496	0.994522	151.8726	0.046644	0.215972	3256	1628	0.115621	0.956	1
	Neural	Neural	NN without ...	Attrition_Flag		0.077396	1628	0.077396	0.998609	187.8613	0.057697	0.240202	3256	1628	0.108856	0.955	1
	Neural9	Neural9	100 Iter NN	Attrition_Flag		0.077396	1628	0.077396	0.998609	187.8613	0.057697	0.240202	3256	1628	0.108856	0.955	1
	Tree2	Tree2	Probability ...	Attrition_Flag		0.074939	1628	0.074939	0.990909	176.4582	0.054195	0.232798	3256	1628	0.099016	0.954	1
	Neural3	Neural3	Reg NN	Attrition_Flag		0.09828	1628	0.09828	0.99457	232.3029	0.071346	0.267107	3256	1628	0.114391	0.954	1
	Neural5	Neural5	4H NN	Attrition_Flag		0.079853	1628	0.079853	0.99753	190.832	0.058609	0.242094	3256	1628	0.123616	0.95	1
	Tree	Tree	Decision Tr...	Attrition_Flag		0.066339	1628	0.066339	0.982005	151.7986	0.046621	0.215919	3256	1628	0.105166	0.949	1
	Neural2	Neural2	NN with Tra...	Attrition_Flag		0.082924	1628	0.082924	0.999742	211.2662	0.064885	0.254726	3256	1628	0.126076	0.949	1
	Neural7	Neural7	Recoded V...	Attrition_Flag		0.095823	1628	0.095823	0.99265	227.9371	0.070005	0.264585	3256	1628	0.134071	0.942	1
	Reg3	Reg3	Backward ...	Attrition_Flag		0.126536	1628	0.126536	0.988438	298.8974	0.091799	0.302983	3256	1628	0.143296	0.936	1
	Reg2	Reg2	Stepwise R...	Attrition_Flag		0.132064	1628	0.132064	0.9914	307.7919	0.094531	0.307458	3256	1628	0.140221	0.936	1
	Reg4	Reg4	Forward Re...	Attrition_Flag		0.132064	1628	0.132064	0.9914	307.7919	0.094531	0.307458	3256	1628	0.140221	0.936	1
	Reg	Reg	Full Regres...	Attrition_Flag		0.12715	1628	0.12715	0.991258	296.0994	0.09094	0.301562	3256	1628	0.143296	0.935	1
	Neural8	Neural8	2H NN	Attrition_Flag		0.126536	1628	0.126536	0.993454	286.2513	0.087915	0.296505	3256	1628	0.143911	0.931	1
	Tree4	Tree4	3 Branch Tr...	Attrition_Flag		0.067568	1628	0.067568	0.992157	175.9933	0.054052	0.232491	3256	1628	0.107011	0.93	1
	Tree3	Tree3	Misclassific...	Attrition_Flag		0.077396	1628	0.077396	0.980583	217.5717	0.066822	0.258499	3256	1628	0.095941	0.922	1

Outcome: According to the ROC, the best model is set to be misclassification tree (Tree 3) with ROC index 0.922 on validation. It also has the best misclassification rate of 0.095941. With the selection criteria of training data for misclassification rate with selection depth 10, neural network with 6 hidden units is better with a rate of 0.064496. Overall, backward, forward, and stepwise regression has the same ROC index of 0.936.

9. Conclusion

For the churn analytics of credit card users, we have built 17 models in total which consists of 4 decision tree models, 4 regression models and 9 neural network models. Besides a few skewed variables, our dataset did not have any missing values. Three categorical variables (Education level, Marital Status, and Income Category) where the fields labelled were unknown and considered as missing were imputed with mode. 4 Variables which were skewed includes Average utilization Ratio, Credit Limit, Total Transaction Amount, and Average open to buy. We followed our rational statistical technique to resolve the skewness which was that the ones with higher skewness statistic were log transformed and the ones with lower skewness statistic were handled using cap and floor. While connecting the node for each model, we came to conclusion that the best models with higher accuracy were the decision trees, followed by neural network and regression.

According to misclassification rate for model assessment, out of 4 decision trees, misclassification is found to be the best one, and for neural network model connected with best tree is the most accurate one. For regression, forward and stepwise are competing. In consideration of ROC index, misclassification tree is the best one and the worst one is tree neural network having index 0.9222 and 0.96 respectively. Based on the result from the decision tree and odds ratio from regression, Total transaction count, total revolving balance, and total transaction amount play an important role in deciding if a customer would churn or not. With our analysis we found out that for a customer having total transaction count less than 56.5 and total revolving balance less than $634, there is a 93% chance of the customer churning. Beside this, male has 59% chance of churning compared to female and people who are educated are 21% less likely to churn against people who are uneducated.

We need to minimize the false positive for our model as we cannot afford a customer to be attributed as churned when our model says that the customer will not churn; this might be a loss to the company. Also, for our business purposes we need to focus on the customer's number. So, the model with lowest false positive is selected as the attired person is labeled as 1.

Model Selection based on Train: Misclassification Rate (_MISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	Attrition_Flag		45	751	63	769
Tree	Decision Tree	VALIDATE	Attrition_Flag		85	727	86	728
Tree3	Misclassification Tree	TRAIN	Attrition_Flag		47	735	79	767
Tree3	Misclassification Tree	VALIDATE	Attrition_Flag		75	732	81	738
Tree2	Probability Tree	TRAIN	Attrition_Flag		35	727	87	779
Tree2	Probability Tree	VALIDATE	Attrition_Flag		68	720	93	745
Neural	NN without Training	TRAIN	Attrition_Flag		67	755	59	747
Neural	NN without Training	VALIDATE	Attrition_Flag		98	734	79	715
Neural2	NN with Training	TRAIN	Attrition_Flag		65	744	70	749
Neural2	NN with Training	VALIDATE	Attrition_Flag		117	725	88	696
Tree4	3 Branch Tree	TRAIN	Attrition_Flag		54	758	56	760
Tree4	3 Branch Tree	VALIDATE	Attrition_Flag		93	732	81	720
Neural4	Tree NN	TRAIN	Attrition_Flag		75	751	63	739
Neural4	Tree NN	VALIDATE	Attrition_Flag		110	747	66	703
Neural5	4H NN	TRAIN	Attrition_Flag		64	748	66	750
Neural5	4H NN	VALIDATE	Attrition_Flag		116	728	85	697
Neural6	6H NN	TRAIN	Attrition_Flag		60	769	45	754
Neural6	6H NN	VALIDATE	Attrition_Flag		115	740	73	698
Neural8	2H NN	TRAIN	Attrition_Flag		108	716	98	706
Neural8	2H NN	VALIDATE	Attrition_Flag		111	690	123	702
Neural9	100 Iter NN	TRAIN	Attrition_Flag		67	755	59	747
Neural9	100 Iter NN	VALIDATE	Attrition_Flag		98	734	79	715
Reg	Full Regression	TRAIN	Attrition_Flag		102	709	105	712
Reg	Full Regression	VALIDATE	Attrition_Flag		120	700	113	693
Reg4	Forward Regression	TRAIN	Attrition_Flag		109	708	106	705
Reg4	Forward Regression	VALIDATE	Attrition_Flag		127	712	101	686

521	Reg4	Forward Regression	VALIDATE	Attrition_Flag	127	712	101	686
522	Reg3	Backward Regression	TRAIN	Attrition_Flag	101	709	105	713
523	Reg3	Backward Regression	VALIDATE	Attrition_Flag	121	701	112	692
524	Reg2	Stepwise Regression	TRAIN	Attrition_Flag	109	708	106	705
525	Reg2	Stepwise Regression	VALIDATE	Attrition_Flag	127	712	101	686
526	Neural7	Recoded Variable NN	TRAIN	Attrition_Flag	70	728	86	744
527	Neural7	Recoded Variable NN	VALIDATE	Attrition_Flag	102	697	116	711
528	Neural3	Reg NN	TRAIN	Attrition_Flag	78	732	82	736
529	Neural3	Reg NN	VALIDATE	Attrition_Flag	107	734	79	706
530								

10. Business Report

To minimize the number of customers who are withdrawing the use of credit cards with respective financial institutions, officials should focus on major customers who have revolving balance less than $634 and simultaneously the customer whose count of transactions per month is below 57. The marketing department can send some promotional messages or email to such customers so that they do not churn and are retained. This overall saves the time and money to be invested in all the credit card users and be effective as well.

Besides this, if the institution wants then they can focus on other groups of people based on gender, education, marital status, and card type but these are just the minor groups for prioritizing than customers having low revolving balance and transaction.

11. References

- *Credit Card Customers Prediction*. (2022, October 30). Kaggle. https://www.kaggle.com/datasets/whenamthancodes/credit-card-customers-prediction (DATA SET)
- Logan, H. (2022, July 15). *What You Need to Know About Credit Card Churning in Canada*. NerdWallet Canada. https://www.nerdwallet.com/ca/credit-cards/what-is-credit-card-churning
- Kusumowibowo, T. S. (2022, October 7). *Credit Card Customer Churn Predictive Analytics - Dev Genius*. Medium. https://blog.devgenius.io/credit-card-customer-churn-predictive-analytics-b012ff8c385d
- *Churning: Definition and Types in Finance*. (2022, January 31). Investopedia. https://www.investopedia.com/terms/c/churning.asp