Final Project Report DSCI 5240

**Data Mining and Machine Learning for Business**

|  |  |
| --- | --- |
| **Group 4:** |  |
| Nishith Reddy Mannuru | 11467417 |

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **TOPIC** | **Page No.** |
| **1** | **Introduction** | **3** |
| **2** | **Motivation** | **3** |
| **3** | **Executive Summary** | **3** |
| **4** | **Dataset Information** | **4** |
| **5** | **Attribute Information** | **4** |
| **6** | **Data Preparation** | **5** |
| **7** | **Exploratory Data Analysis** | **6** |
| **8** | **Model Analysis** | **14** |
| **9** | **Model Comparison and Results** | **28** |
| **10** | **Final Diagram** | **31** |
| **11** | **Conclusion** | **32** |
| **12** | **References** | **32** |

# INTRODUCTION

The rising defaults will cause lending institutions, such as commercial banks, to lose a large amount of money. As a result, having a risk prediction model and being able to define the most critical variables that are indicative of people who are more likely to default on credit card loans is critical for banks. A robust model is not only a beneficial tool for lending institutions to make credit card application decisions, but it may also assist consumers in recognizing habits that can harm their credit ratings.A large number of credit applications are received by commercial banks. Many of them are turned down for a variety of reasons, including large loan balances, insufficient income, or too many inquiries on a person's credit report, for example. Manually analyzing these applications is mundane, error-prone, and

time-consuming. You'll be charged late fees, your interest rate will rise, and your credit score will suffer if you don't pay your credit card bill on time. If you continue to miss payments, your card may be suspended, your debt may be transferred to a collection agency, and the debt collector may file a lawsuit against you and garnish your wages. If you have a debt - that is, if you do not pay the whole amount owed you will be charged interest on your purchases for the billing cycle. A credit score is a numerical expression that represents a person's credit worthiness based on a level of study of their credit files. A credit score is essentially determined by the information contained in a credit report, which is normally obtained from credit agencies.

Machine learning algorithms are now the most widely used method for creating models. The term "machine learning" was created by Arthur Samuel in 1959. Machine learning's main goal is to teach computers how to examine past data or samples that we offer and automatically learn from them by detecting patterns in order to make better decisions in the future. There are numerous factors to consider when selecting the most appropriate model. Machine learning algorithms are of two types: supervised and unsupervised learning. Unsupervised learning is utilized when the given training data set is neither labelled nor categorised, whereas supervised learning is used to predict output values in the test dataset.Instead than finding out the correct answers, unsupervised learning deals with data that has unknown possible outputs and seeks for hidden structures in the unlabeled data.

# MOTIVATION

Is it possible to anticipate who is most likely to default? If so, the bank may be able to prevent the loss by offering the consumer other options such as debt consolidation and forbearance. As the financial industry has evolved the past few decades, financial threats are presenting a pattern regarding commercial bank credit risk. One of the most serious threats to commercial banks is credit customer risk prediction. Our analysis could help the banking sector identify who is more susceptible to defaulting on credit card payments and take calculated action to combat this.

# EXECUTIVE SUMMARY

The objective is to compare different data mining techniques to find out what are the factors which are influencing the default payment next month in the credit card clients the most. It is important to find out so that it can help verify the causes that are contributing to various age groups defaulting on credit

cards. The end-goal is to build a more complete picture of how and to classify them in a variety of meaningful ways to predict who is more likely to default on credit card loans.

The following models are used to classify the defaulters of credit card loan:

1. Logistic Regression with all variables.
2. Logistic Regression with only significant variables.
3. Stepwise Logistic Regression
4. Decision Tree
5. HP Forest
6. Auto Neural with only TanH activation
7. Auto Neural with only Sine activation

In the final step, all the models are compared to find the best model that classifies the defaulters of credit card loans.

# DATASET INFORMATION

This project uses second-hand data retrieved from the UCI Machine Learning repository website; [default](https://archive.ics.uci.edu/ml/datasets/default%2Bof%2Bcredit%2Bcard%2Bclients) [of credit card clients [1].](https://archive.ics.uci.edu/ml/datasets/default%2Bof%2Bcredit%2Bcard%2Bclients) It encapsulates a varied set of features about the real cardholders' credit risk data in Taiwan and compares the predictive accuracy of the probability of default customers.

# ATTRIBUTE INFORMATION

ID: unique identification number assigned to each customer LIMIT BAL: amount of given credit access line

SEX: gender (1 = male; 2 = female)

EDUCATION: highest degree obtained (1 = graduate school; 2 = university; 3 = high school; 4 = others; 5 = unknown; 6 = unknown)

MARRIAGE: marital status (1 = married; 2 = single; 3 = others) AGE: age in year

PAY 0: monthly payment record in September PAY 2: monthly payment record in August PAY 3: monthly payment record in July

PAY 4: monthly payment record in June PAY 5: monthly payment record in May PAY 6: monthly payment record in April

BILL\_AMT1 : total amount owed in september BILL\_AMT2 : total amount owed in August BILL\_AMT3 : total amount owed in July BILL\_AMT4 : total amount owed in June BILL\_AMT5 : total amount owed in May BILL\_AMT6 : total amount owed in April

PAY\_AMT1: amount of previous payment in september PAY\_AMT2: amount of previous payment in August

PAY\_AMT3: amount of previous payment in July PAY\_AMT4: amount of previous payment in June PAY\_AMT5: amount of previous payment in May PAY\_AMT6: amount of previous payment in April

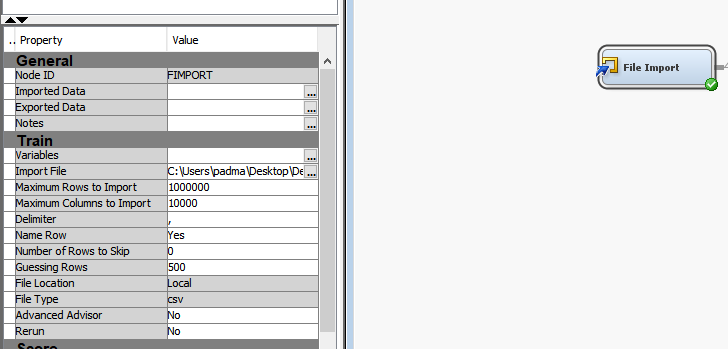
default payment next month: whether a customer is defaulted on next months payment or not (1 = defaulter; 0 = non-defaulter)

The columns from PAY0 to PAY6 represent the repayment status in each month from April to September:

-1 indicates paying duly for one month; -2 indicates paying duly for two months; The positive number shows how many months the payment has been delayed. This multivariate dataset has 30,000 instances and covers 25 attributes. The ID column of the dataset is not included for analysis. The 23 variables that are used as the explanatory variables are listed in the link to the [Excel Sheet](https://docs.google.com/spreadsheets/d/1hx4zPDp2ksy2_VOiNH_giv9ERbYJfWN-gAeCnO741Yo/edit#gid%3D0).

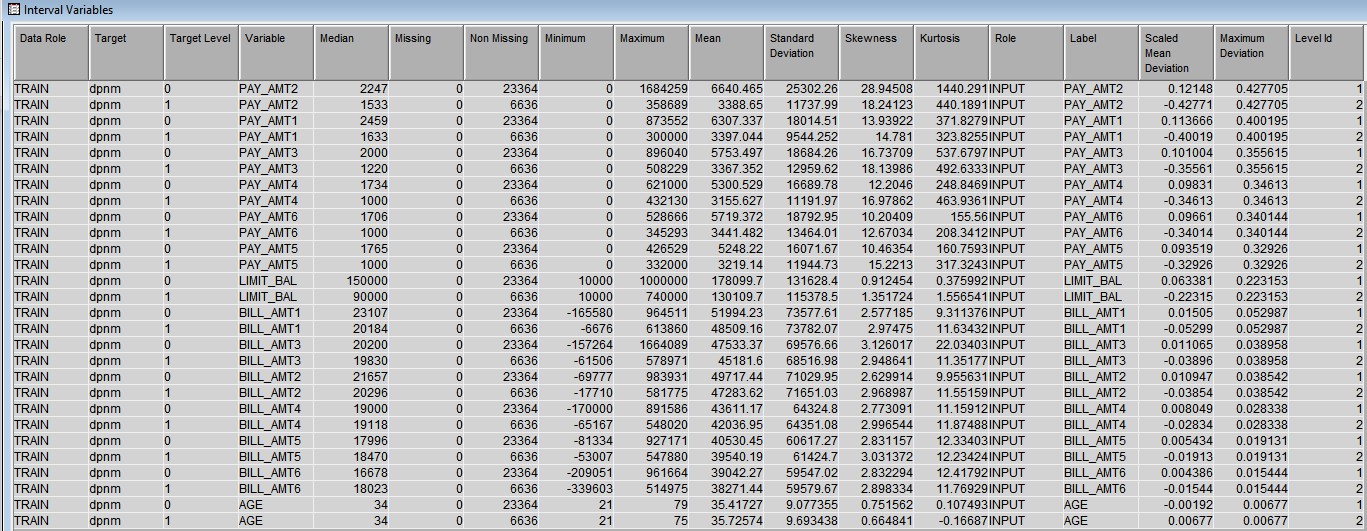
## LOADING THE DATASET

**DATA PREPARATION**

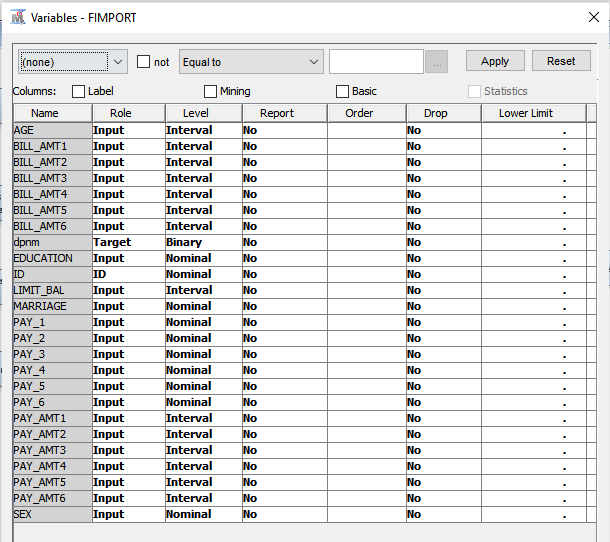
The dataset is loaded into SAS enterprise miner by creating a new project and then a new diagram. We used the file import node to import our data excel file.

# EXPLORATORY DATA ANALYSIS

The Summary statistics for the Interval variables in the dataset is shown below.

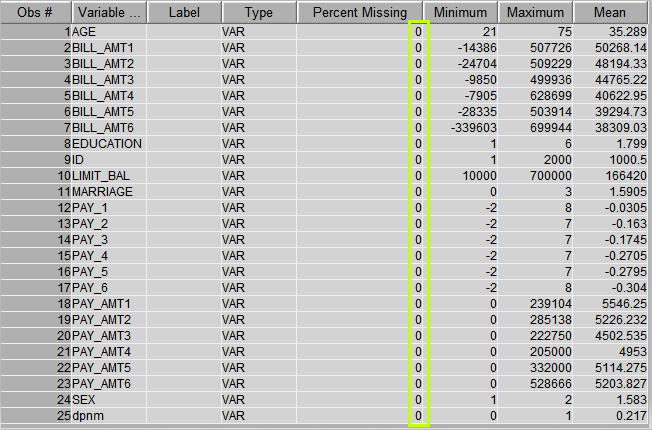


## Exploring the variables in the dataset:

We have determined the different independent variables and their levels, along with the target variable.

## MISSING VALUES

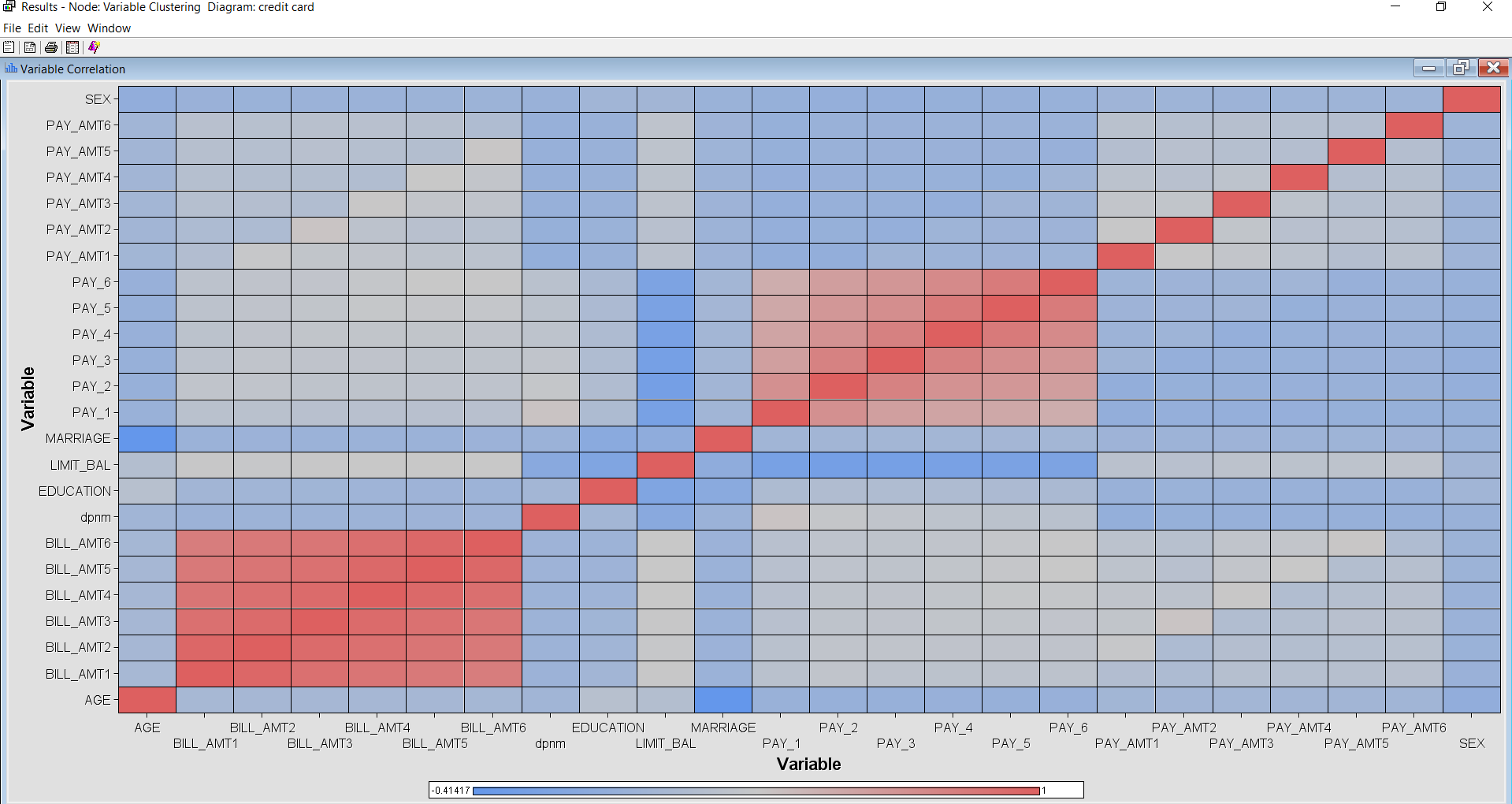
The dataset does not have any missing values.



## CORRELATION ANALYSIS

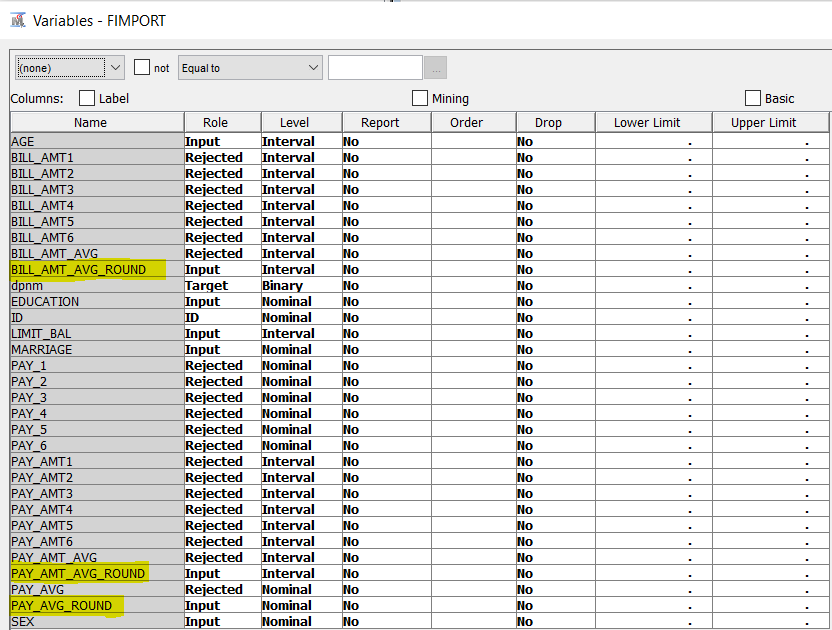
Correlation analysis is a statistical method to analyse the strength and the relationship between the variables. To know how strongly the variables are related to each other we plot a correlation matrix.

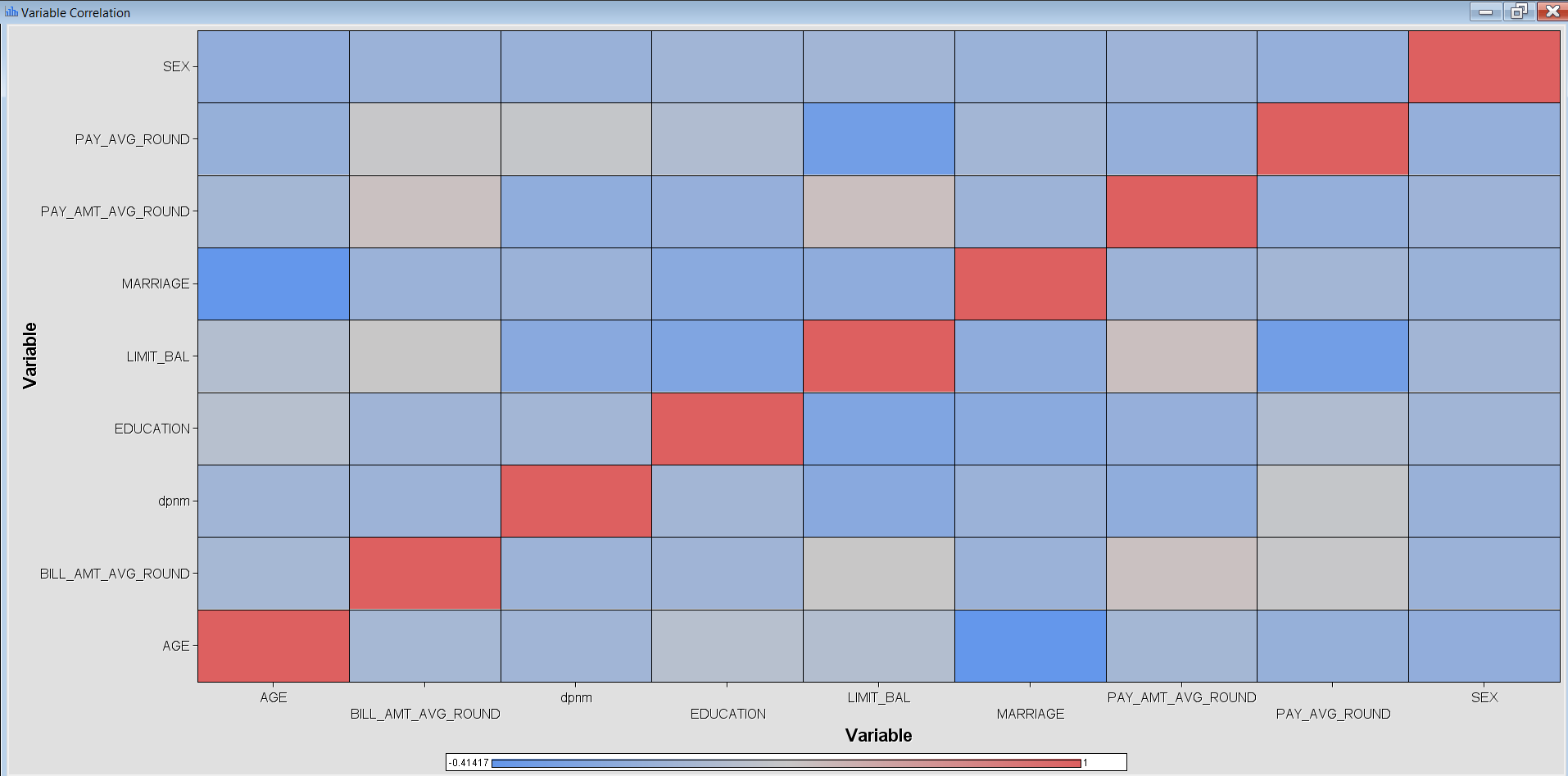
In our case all the BILL\_AMT variables and PAY variables are highly correlated with each other. We have to address the issue of correlation before we can perform any analysis, as correlated variables can affect the output of our analysis. In contrast to covariance, correlation indicates the strength of a link between two variables. The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will highly decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation..If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.



## RESOLVING THE ISSUE OF CORRELATION

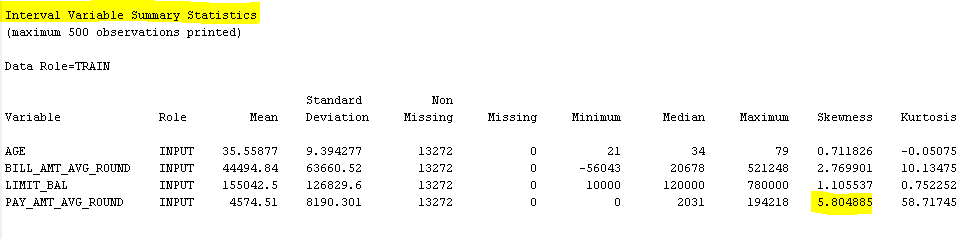
* + All the BILL\_AMT (1 to 6) variables are highly correlated with each other and Pay\_1 to 6 are relatively less correlated.
  + Since the BILL\_AMT variables represent the amount of bill statements generated in a particular month (i.e., for 6 months) , we decided to create a new column which represents the average amount of bill statements for 6 months. So that we can eliminate the correlation.
  + Since we are going to use the average for the bill amount, it makes sense to calculate the average for the PAY variables as well as the PAY\_AMT variables, because PAY variable represents the repayment status for 6 months and PAY\_AMT variables represent the amount of previous payment for 6 months.
  + We are doing this to put everything on the same scale.

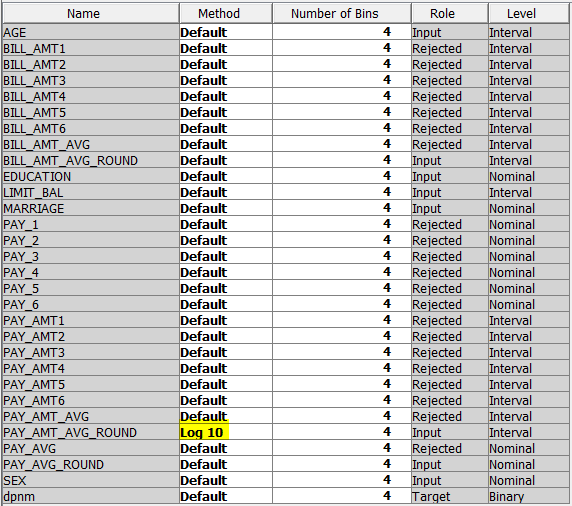


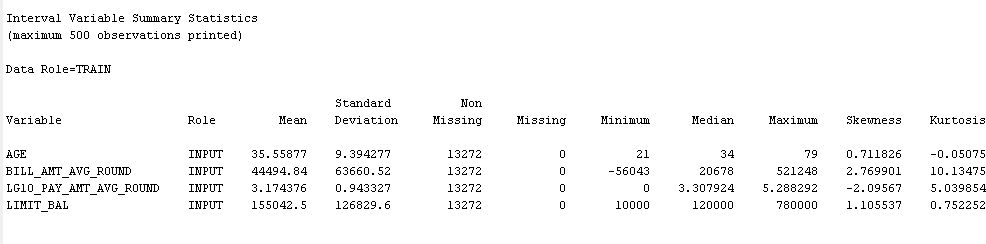


## NORMAL DISTRIBUTION

* Data cleaning is the most important step that is to be performed before any analysis.
* We used Stat explorer in SAS enterprise miner to check for missing values.
* And, as can be observed there are no missing values in our data set.
* We can also see that some variables, PAY\_AMT\_AVG\_ROUNG is highly skewed which may require transformation.
* Since the PAY\_AMT\_AVG\_ROUND variable is skewed, we performed Log10 transformation, to make sure it was in the range of -3 to +3 standard deviations away from the mean.



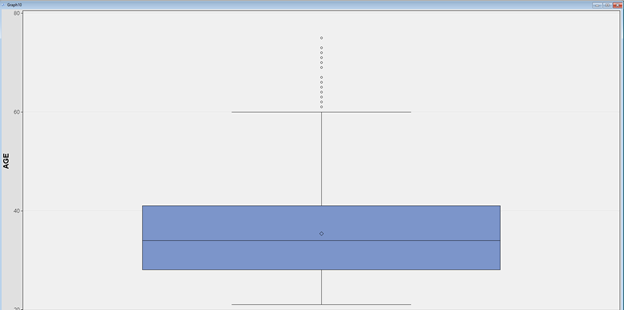


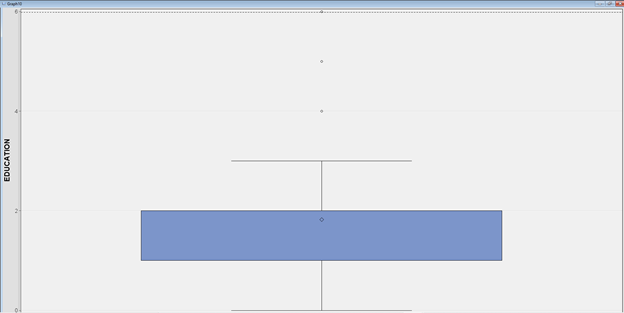


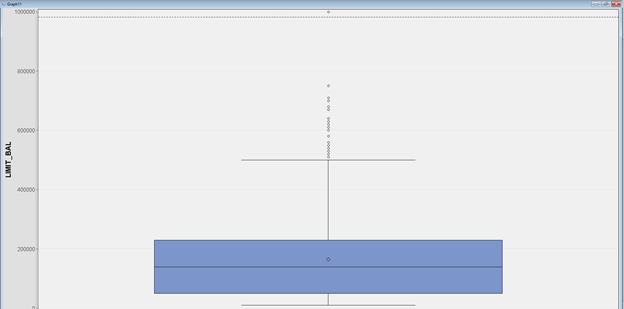
* No variables are highly skewed.

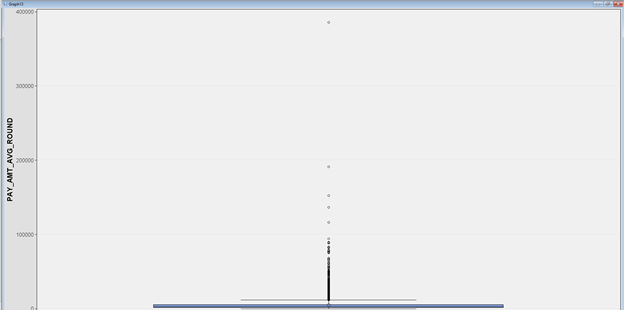
## OUTLIER DETECTION

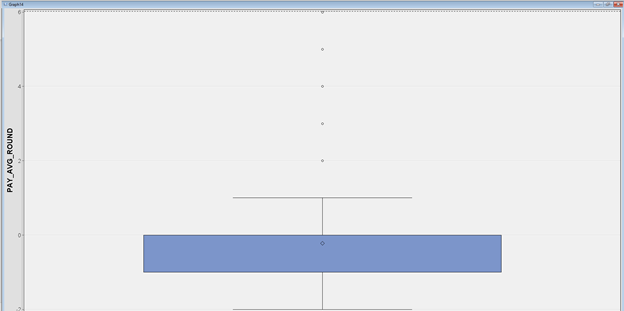
* An outlier is a lone observation that appears to diverge significantly from the rest of the sample.
* An outlier could be a sign of skewed data. For instance, data could have been wrongly coded, or an experiment could have been done incorrectly.
* In some circumstances, determining whether an outlying point is bad data may be impossible. Outliers could be the result of chance or could indicate something scientifically important.

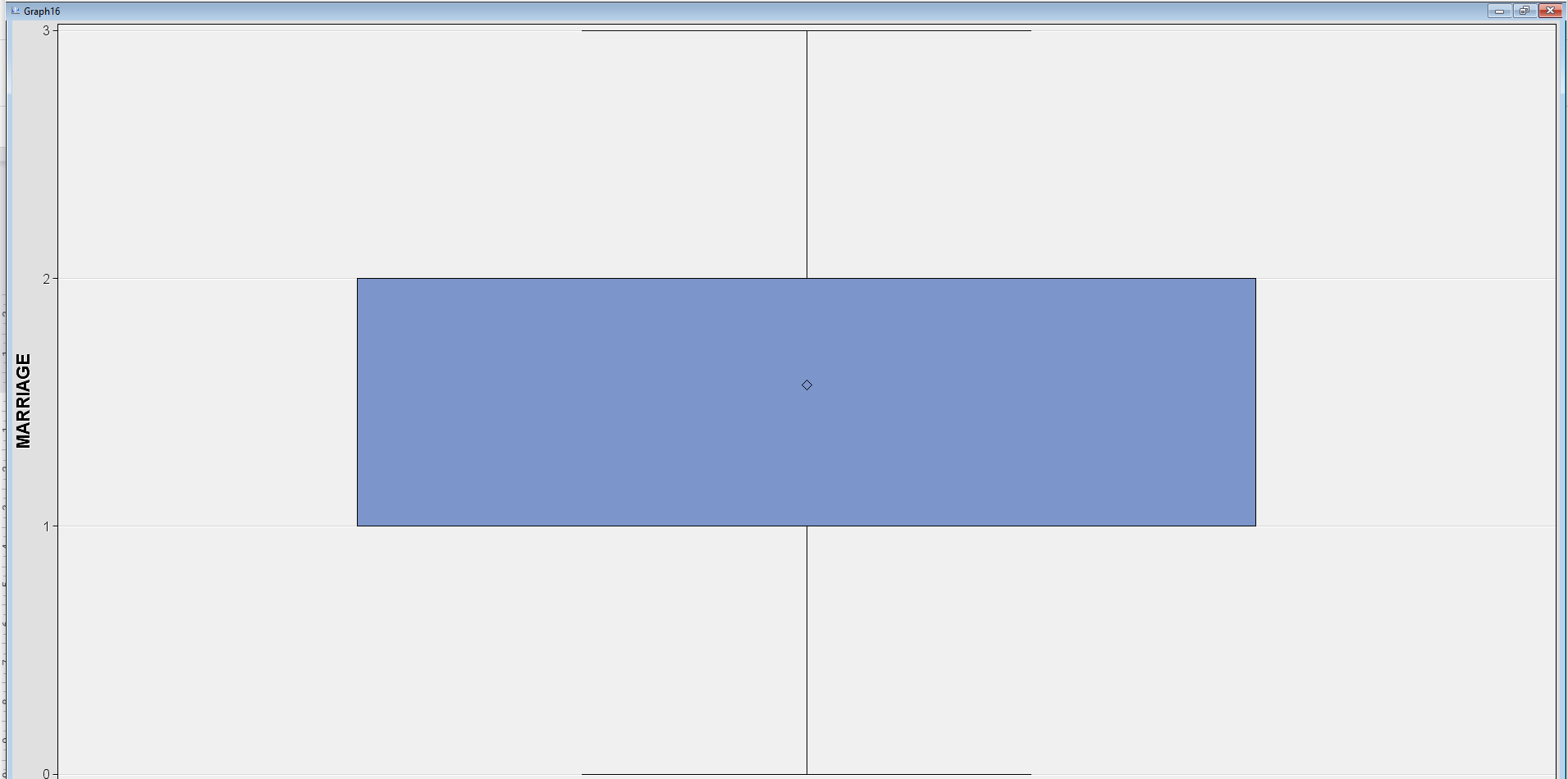






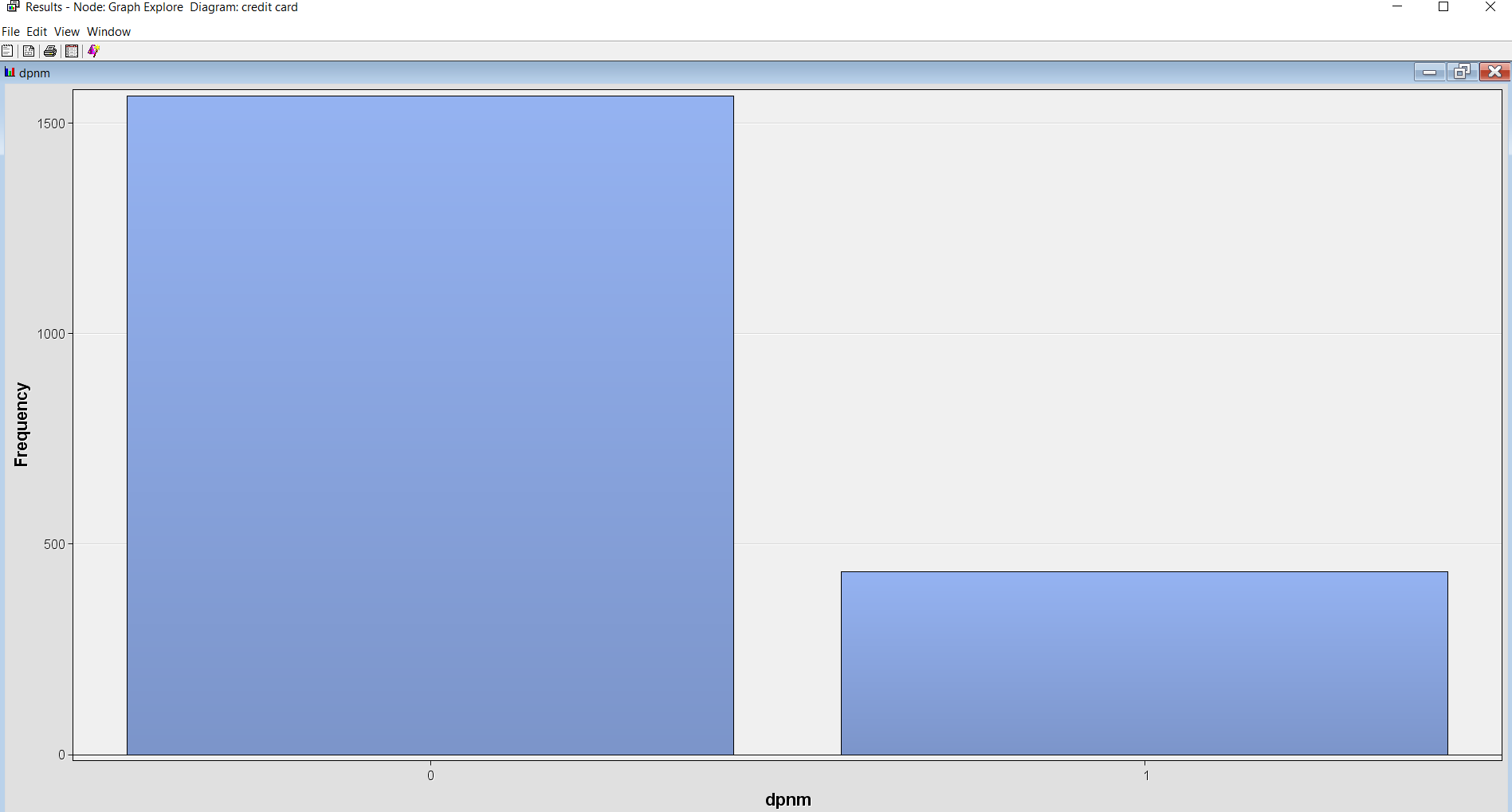




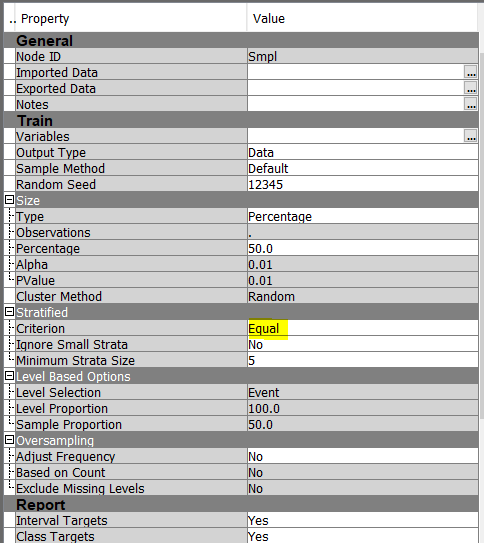


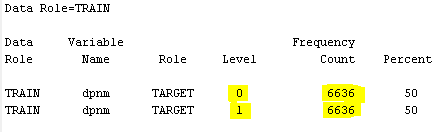
There are outliers, but they are not anomalies or data entry mistakes. These outliers are nothing but data points that usually fluctuate considerably from each other amongst all other observations. So, we are considering them to be real data.

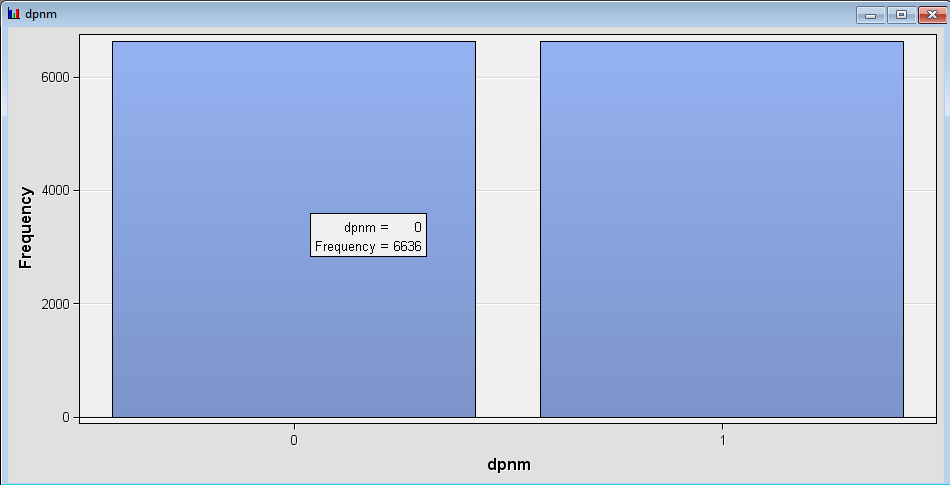
## TARGET VARIABLE (DPNM) BALANCING



* The target variable “dpnm” is highly unbalanced. It contains only 6636 cases of people who have defaulted on their credit card.
* So we are using undersampling to make our data target variable balanced. I.eWe are considering the exact number of “ 0’s “ to exactly match with our “ 1’s “ in our data, so that our analysis would not be skewed onto a single variable.





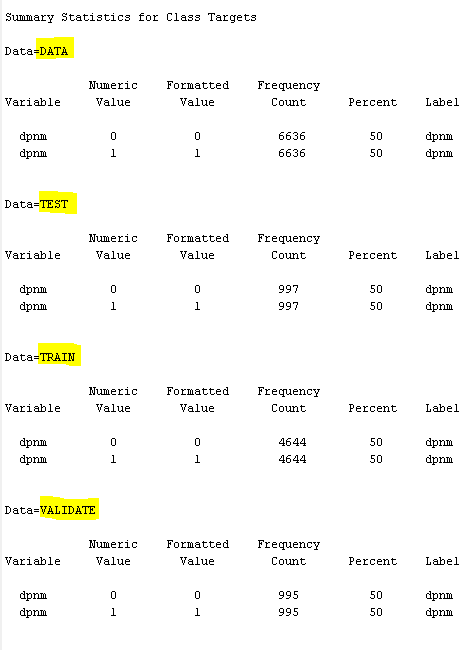
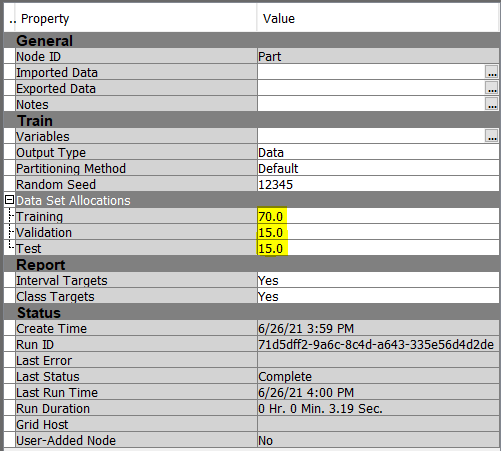


* Now, both our target variables look balanced and are perfect to undergo analysis.

# MODEL ANALYSIS

## DATA PARTITIONING

* Splitting the data into training, validation and testing is a really important step to check how our models are gonna perform with unseen data.
* So, we are splitting the data in the ratio of 70% training , 15% validation, and 15% testing.



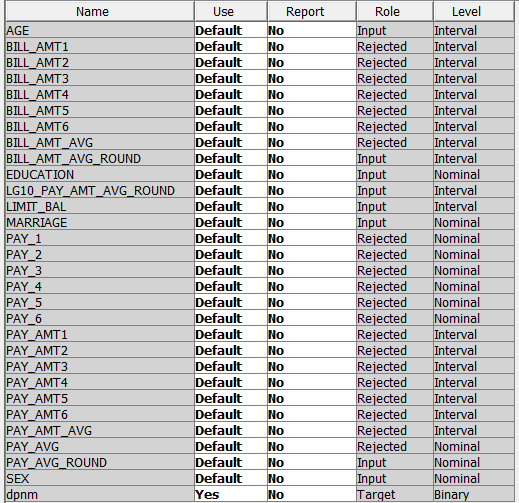
* It can be seen in the above image that the split is random,

but the proportionality is maintained. The % values are almost exact, and proportionality is maintained even when the frequency count fluctuates for each set. This result shows how the data partition has been accurately taken place without any errors.

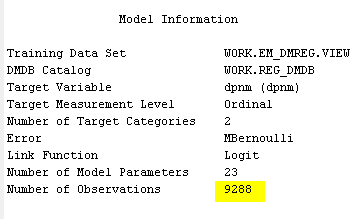
# LOGISTIC REGRESSION

Logistic regression is a statistical analysis approach for predicting a nominal value based on previous data set observations. A logistic regression model analyzes the relationship between one or more existing independent variables to predict a dependent nominal variable.

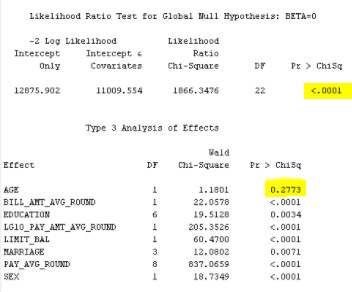
# Model 1- Using all the variables



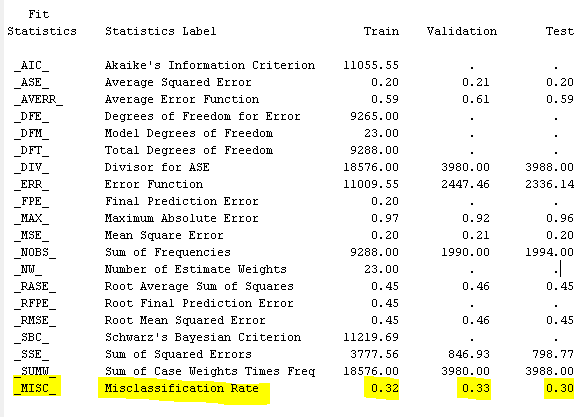
The above image shows the variables that we will be using for this logistic regression.



The number of observations the model is using is 9288, which is exactly the number of observations that we had in our training set.

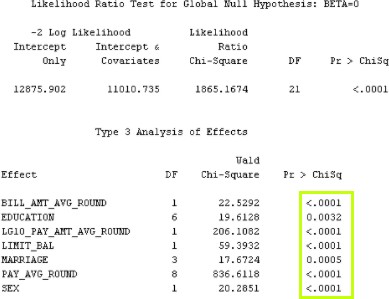
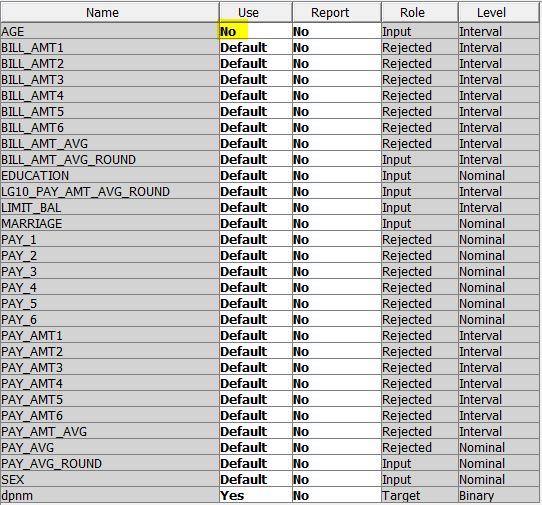


* From the image it is clear that the model is significant because the p-value is ( <0.001) which is less than 0.05.
* All the predictor variables are also significant except Age.
* The P-value of Age is > 0.05, so it's not significant.
* So, in our next logistic regression analysis we have to remove the age variable.

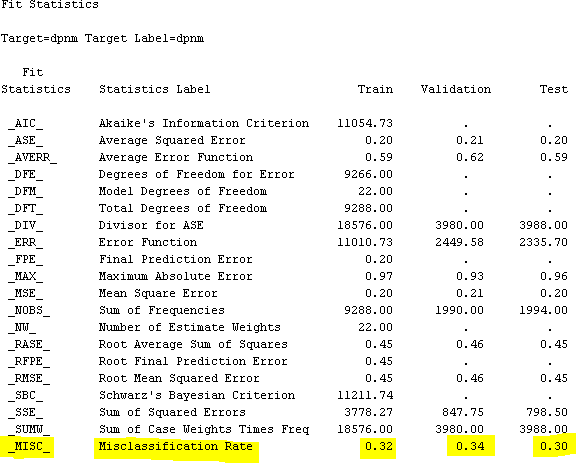


# Model 2- Using only significant variables

Then we ran the second logistic regression model with only the significant variables. The Age variable has been removed as it is insignificant.

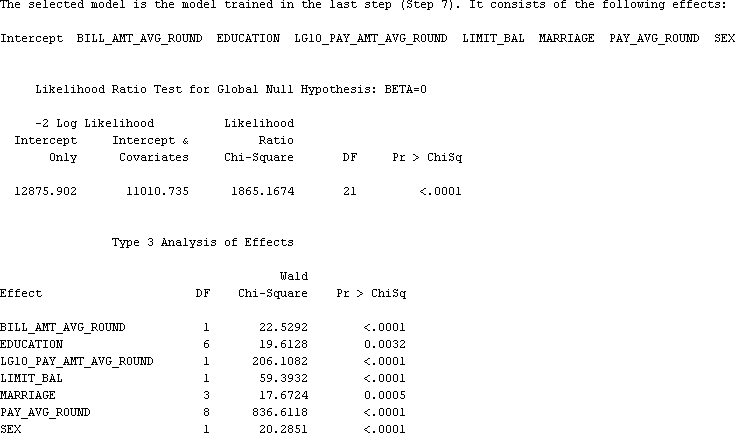


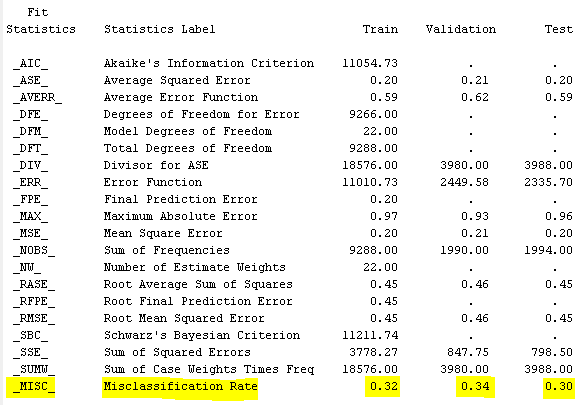
* + Now, all the variables are significant. The misclassification rate for validation data is 0.34.



# Model 3- Using stepwise regression

Stepwise regression is a statistical approach of fitting regression models in which the selection of predictive variables is done automatically. Each stage considers whether a variable should be added to or subtracted from the set of explanatory variables based on some predetermined parameter.

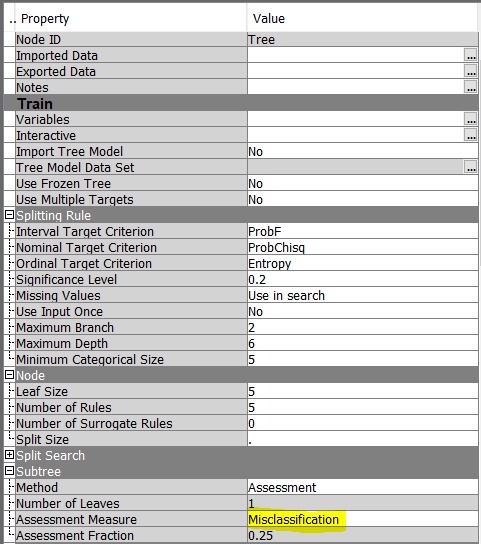




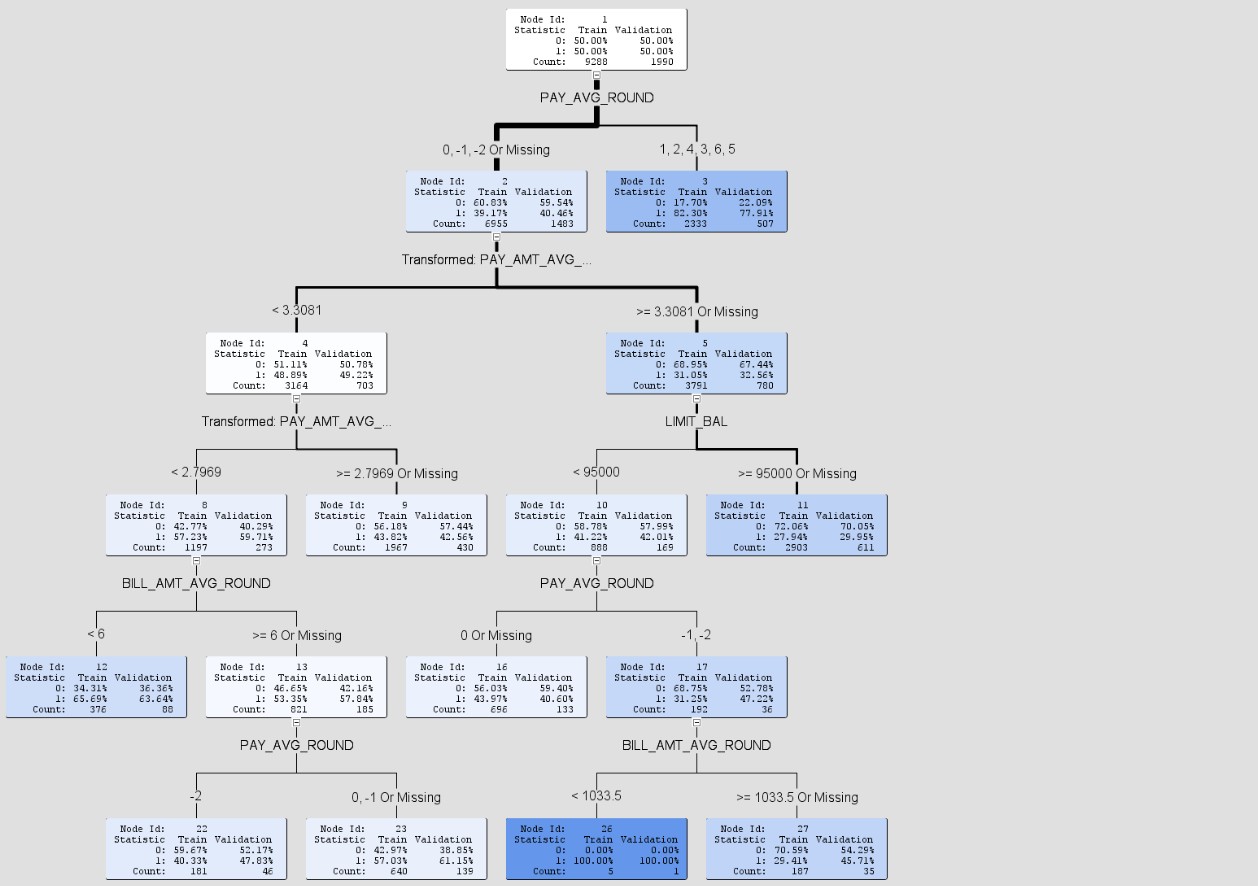
* + All the variables generated by stepwise regression are significant. The validation misclassification rate is 0.34.
  + The misclassification rate for the above 2 models is slightly higher than the logistic regression model with all variables which shows that the model is generalizing well with unseen data.

# DECISION TREES

We're utilizing a decision tree to mathematically derive the probable conclusion from a series of related choices. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable.

We are using Misclassification rate as the assessment measure, so that at the end we can determine the best model based on misclassification rate.

## TREE DIAGRAM

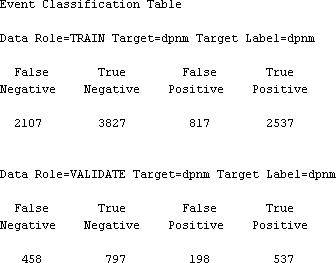
The below tree diagram shows PAY\_AVG\_ROUND is the root node that is used for the first split into 2 sub leaves. We can see that the intensity of the color corresponds to whether the person will default on the credit loan and from the tree diagram, we can see that people who have BILL\_AMT\_AVG\_ROUND with less than $1033.5 will default on their credit card loan.

### LEAF STATISTICS

* + - Based on the misclassification rate the number of leaves are 9.

### SUB-TREE ASSESSMENT PLOT

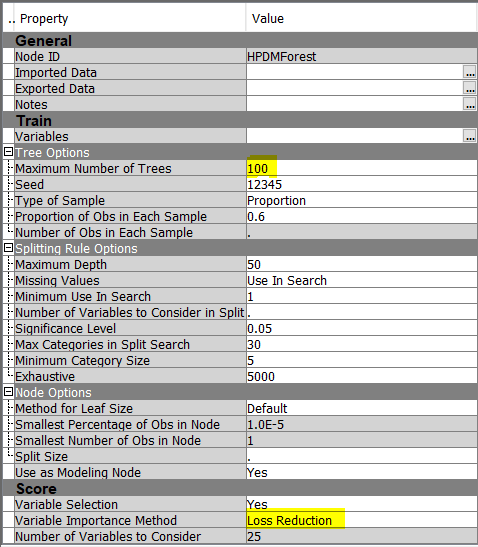
* + Based on the validation misclassification rate (0.3296), the model is pruned when the number of leaves is at 9.

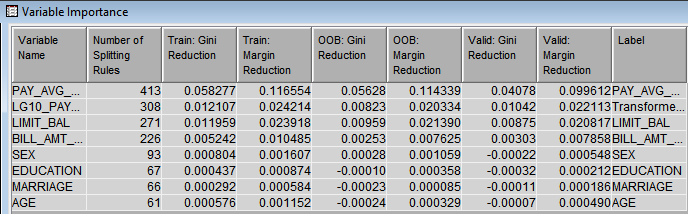


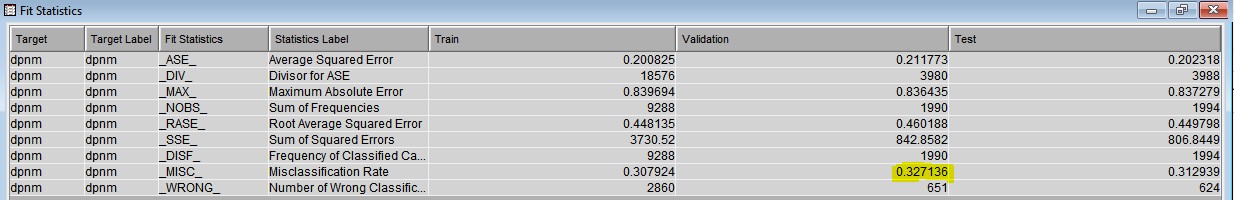
# HP FOREST

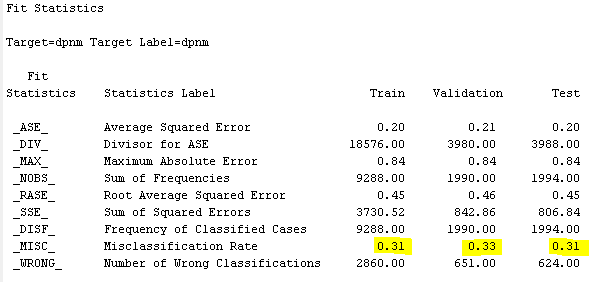
Instead of just coming to a conclusion based on the decision tree model that was built on the entire dataset, we are using random forest. Random forest is similar to decision trees, except it splits our total data set into different trees randomly and then makes a conclusion on which class that particular observation should belong to.

* + We decided to split our dataset into a total of 100 trees and selected the best nodes based on minimum loss reduction.

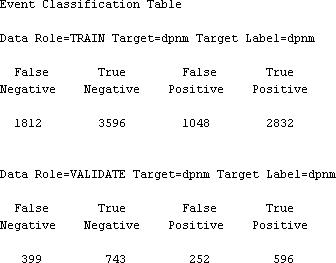








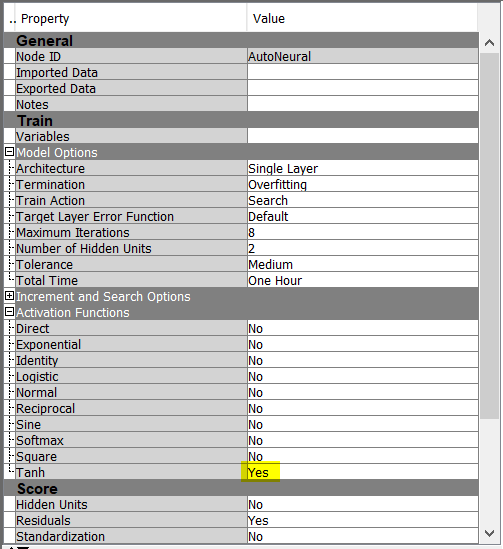
* + We can see that the validation misclassification rate for the decision tree model is 0.33.

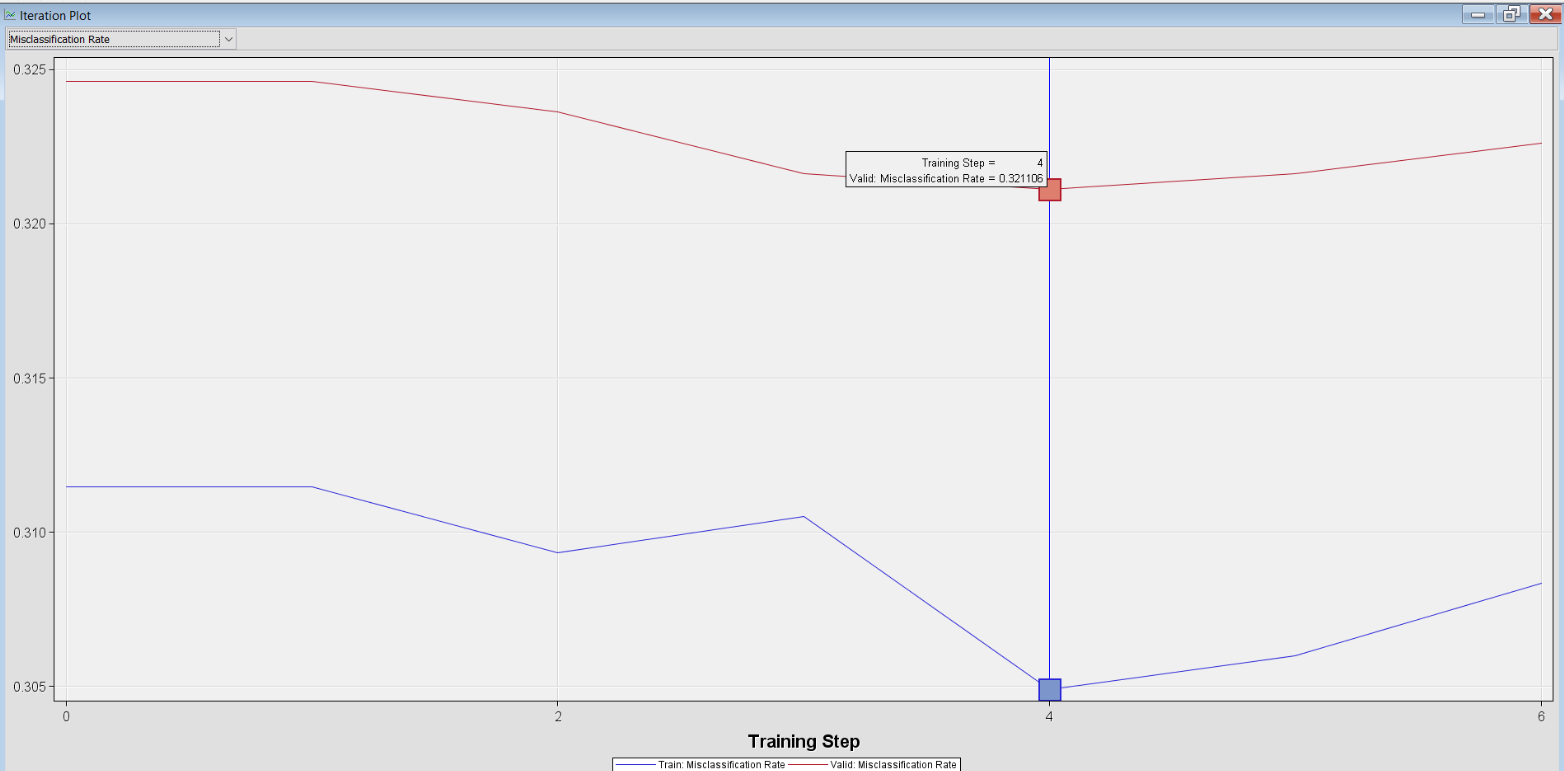


# AUTONEURAL WITH TANH ACTIVATION

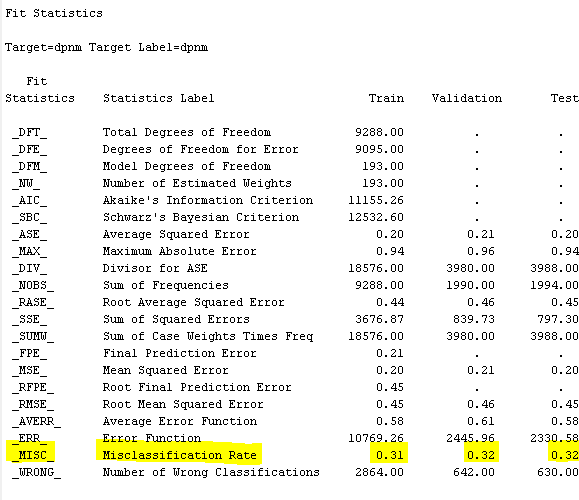
The AutoNeural node, as its name implies, automatically configures a Neural Network model. It makes use of default combination and error functions. The algorithm compares multiple activation functions and chooses the best one.

* + After connecting the Auto Neural node, in the properties panel, from the model options, the number of hidden units chosen is 2.
  + We are running an auto neural network by selecting yes for Tanh activation function as it is mainly used for sigmoid and classification purposes.



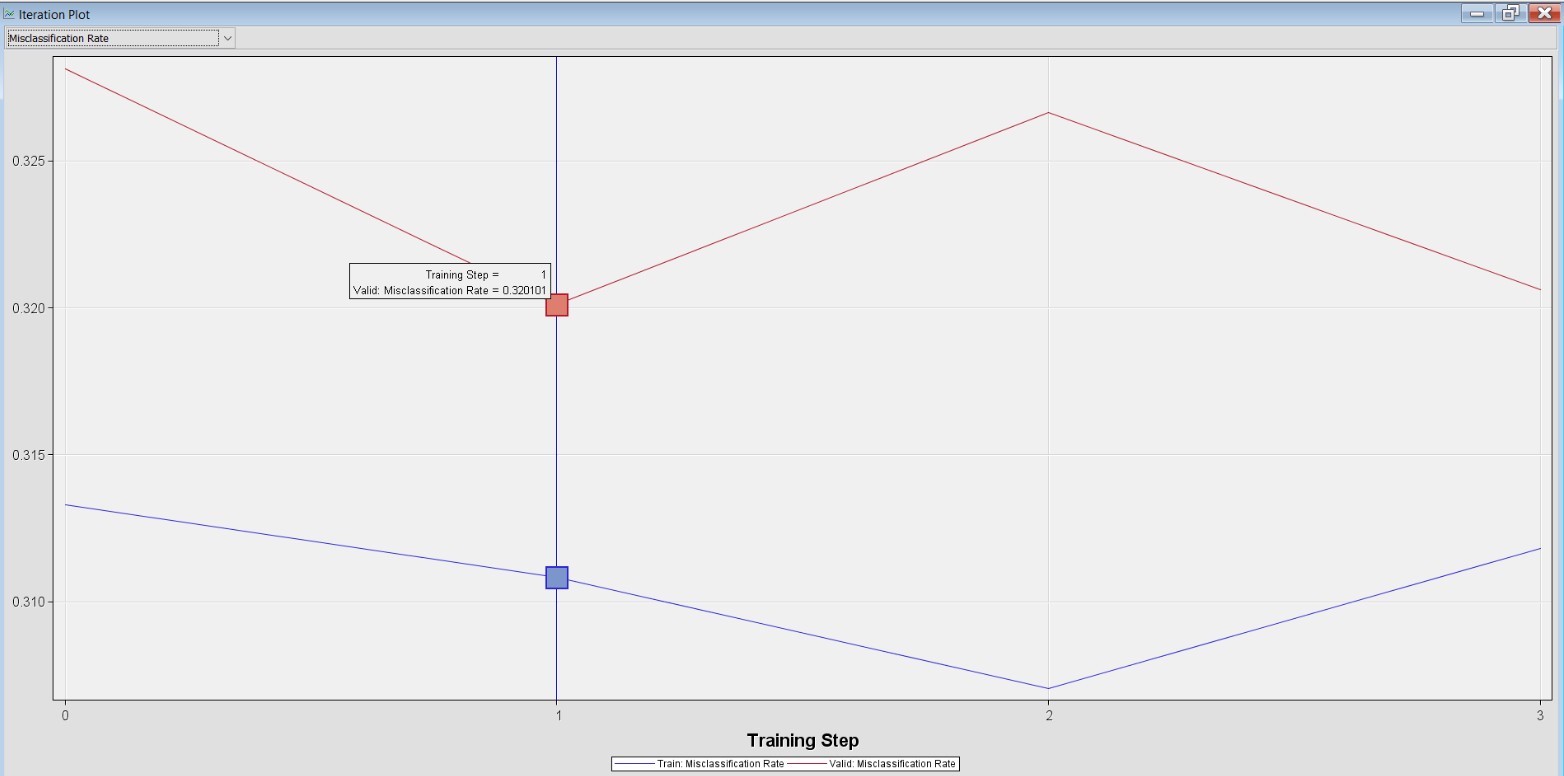
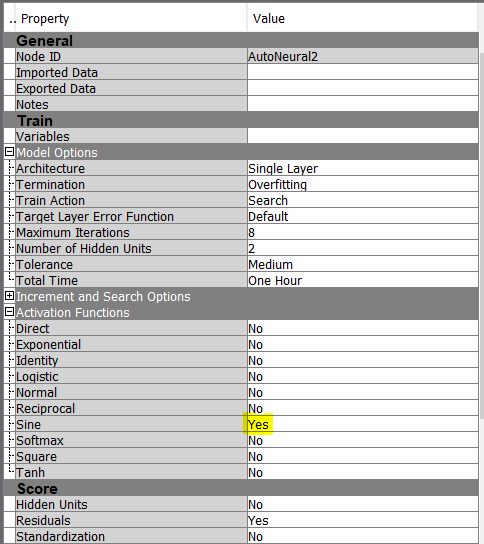
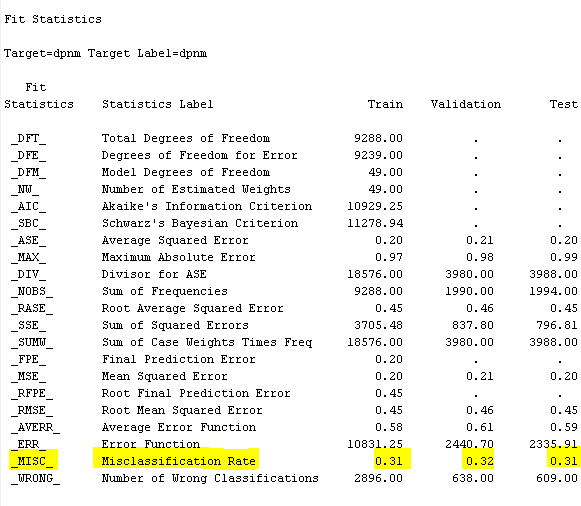


* + - This iteration plot shows us when the misclassification rate is at its minimum for the validation set.
    - So, this describes after the 4th iteration the misclassification rate of the model starts to increase.



* + We can see that the validation misclassification rate for the AutoNeural model with TanH activation function is 0.32 which is lesser than the previous models.

# AUTONEURAL WITH SINE ACTIVATION

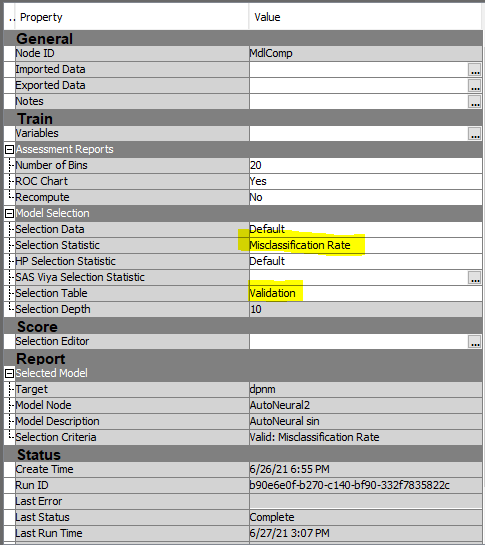
* + After connecting the Auto Neural node, in the properties panel, from the model options, the number of hidden units chosen is 2.
  + We are running an auto neural network by selecting yes for sine activation function.
    - This iteration plot shows us when the misclassification rate is at its minimum for the validation set.
    - So, this describes after the 1st iteration the misclassification rate of the model starts to increase.
  + We can see that the validation misclassification rate for the AutoNeural model with Sine activation function is 0.32 which is similar to the AutoNeural model with TanH activation function.

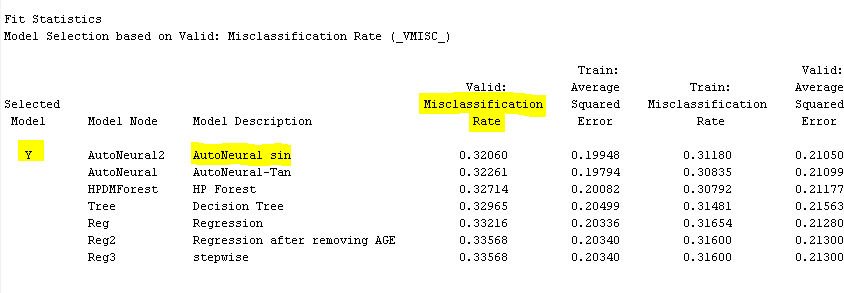
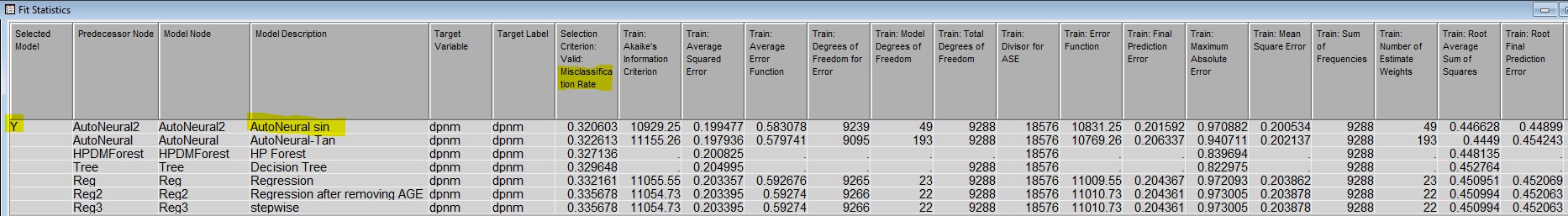
**MODEL COMPARISON**

The model comparison node from the Assess group is used to compare the performance efficiency of all of our models. The results of all the models are compared using two criteria - misclassification rates and ROC curves.

# PREDICTING THE BEST MODEL BASED ON MISCLASSIFICATION RATE

We are using the model comparison node to compare which model is the best based on the validation misclassification rate of each model.



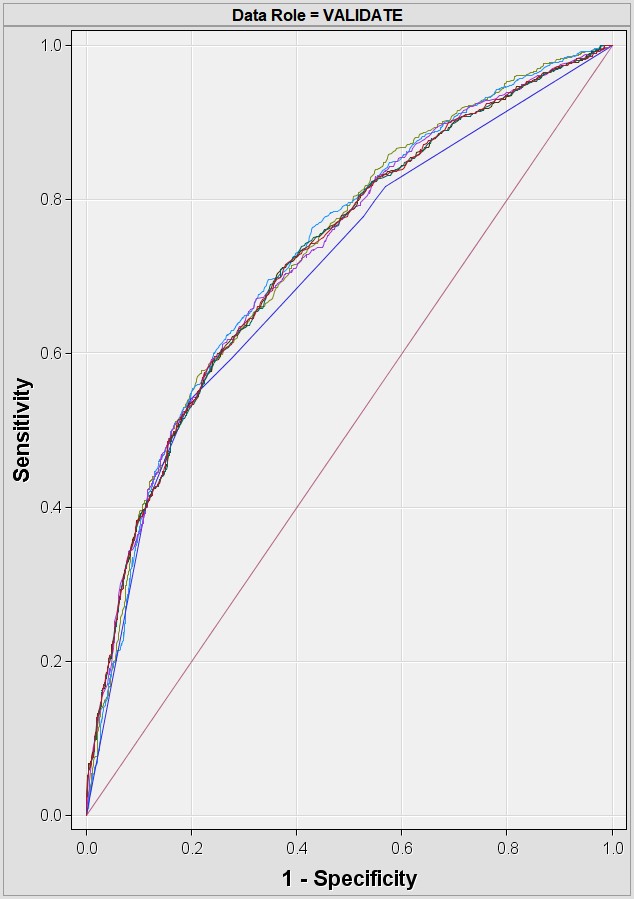


* + The misclassification rate for the model AutoNeural with Sine activation function has the least value (0.32) when compared to the validation misclassification rates of other models.
  + Therefore, the best model selected for classifying persons who will default is AutoNeural with Sine activation function.

# PREDICTING THE BEST MODEL BASED ON ROC CURVE

* + To predict the best model based on the ROC curve, we need to calculate the sensitivity and 1-specificity from each model.
  + Sensitivity is calculated by the formula TP/TP+FN and,
  + Specificity is calculated by the formula TN/TN+FP.
  + And the values are tabulated below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **True Positive** | **True Negative** | **False Positive** | **False Negative** | **Sensitivity** | **1 - Specificity** |
| **Logistic Regression (all variables)** | **517** | **812** | **183** | **478** | **0.52** | **0.82** |
| **Logistic Regression (age removed)** | **513** | **809** | **186** | **482** | **0.52** | **0.81** |
| **Stepwise Regression** | **511** | **820** | **175** | **484** | **0.56** | **0.81** |
| **Decision Tree** | **537** | **797** | **198** | **458** | **0.54** | **0.80** |
| **HP Forest** | **596** | **743** | **252** | **399** | **0.60** | **0.75** |
| **Autoneural (TanH)** | **620** | **728** | **267** | **375** | **0.62** | **0.73** |
| **Autoneural (Sine)** | **567** | **785** | **210** | **428** | **0.57** | **0.79** |



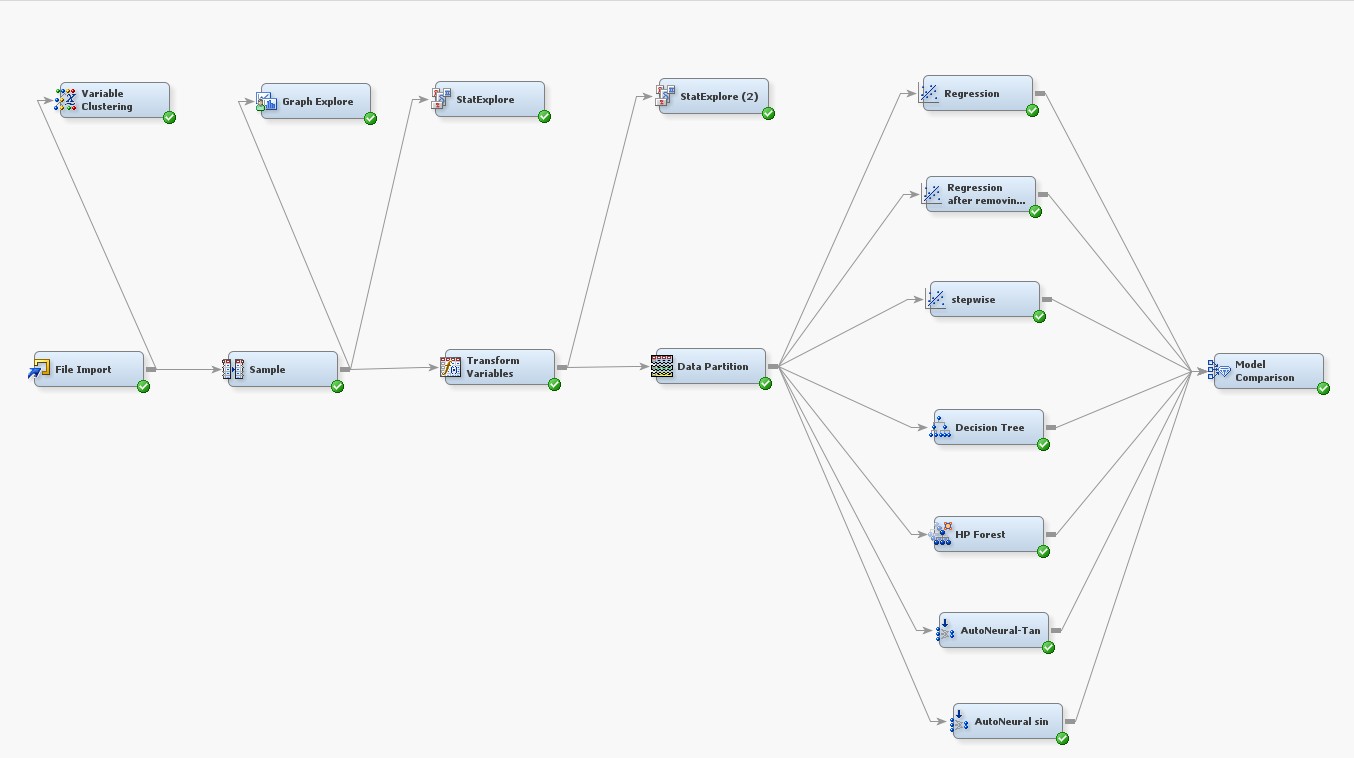


* + The ROC curve has the added benefit of being able to depict a model's performance against a variety of cutoff parameters. It represents the trade off between sensitivity and 1 - specificity.
  + If the sensitivity is closer to 1 and 1 - specificity is closer to 0 it is said to be a better model.
  + In our case, the ROC graph suggests, the auto neural model with Tanh activation function is better when compared with the other models.

# RESULTS

After checking the performance results of all the models, the AutoNeural Network model with Sine activation function was the best model to classify who will default on the credit card loan based on the lowest misclassification rate. Based on the ROC curve performance measure, we find the AutoNeural Network model with TanH activation function to be the best model.

# FINAL DIAGRAM



The Final Diagram of all the models executed are shown in the diagram above.

# CONCLUSION

The objective of this project is to train multiple supervised learning algorithms to predict customers behavior on paying off credit card balance. We first investigated the data by performing exploratory data analysis which includes cleaning missing or invalid values and exploring the relationship between the independent variables. The Explore option of File Import helps us to visualize these relationships and important features. Since the original dataset was imbalanced, we used oversampling to balance the dataset. Then we partition the dataset into training, validation and testing dataset. We started with the logistic regression algorithm with all variables, then a second logistic regression with only significant variables and removing the age variable, followed by stepwise logistic regression. In all these models, the best variables are selected, that is the variables which are significant with the p-value <0.05. We also built a decision tree model. Then we tried a random forest model which has a better performance than the former models. At the end, we tried to build two neural networks, one with Tanh activation function and the second with Sine activation function.

A model with lowest misclassification rates is obtained successfully and we can predict who is more likely to default on credit card loans. We conclude that the AutoNeural with Sine activation function is the best model when compared to the validation misclassification rates of other models. By using the ROC curve and confusion matrix to evaluate the model performance, we conclude that the AutoNeural model with Tanh activation function is the best model to classify the person who will default on the credit card loan when compared with the other models.

**REFERENCES**

1. UCI machine learning repository: Default of credit card clients data set. (n.d.). [https://archive.ics.uci.edu/ml/datasets/default+of+](https://archive.ics.uci.edu/ml/datasets/default%2Bof%2B)