USING MORTGAGE ORIGINATION DATA TO PREDICT FUTURE DEFAULTS

MIS 6324.501 - Group 12

Fall 2018

Krishna Sai Chaitanya Matta

Syed Raza

Ryan Vandergriff

Charles Wilson

Alice Zheng

**Table of Contents**

[**Executive Summary** 2](#_Toc529803426)

[**Background and Objective** 3](#_Toc529803427)

[**Data Source and Metrics** 3](#_Toc529803428)

[**Training Model** 5](#_Toc529803429)

[**Decision Tree and Results** 5](#_Toc529803430)

[**Logistic Regression** 7](#_Toc529803431)

[**Neutral Networks** 8](#_Toc529803432)

[**Neutral Networks vs Decision Tree vs Logistic Regression** 10](#_Toc529803433)

[**Applying the Model** 11](#_Toc529803434)

[**Managerial Implications and Use** 13](#_Toc529803435)

[**References** 14](#_Toc529803436)

# 

# **Executive Summary**

This report summarizes the statistical modeling and analysis and results associated with residential mortgages. The purpose of this report is to document both the model and all the corresponding data analysis and inference techniques used during the subsequent statistical analysis. The primary aim of the task was to design a model so that we can predict which loans are at most risk to default from the data we acquired from Freddie Mac.

We took multiple factors (LTV ratio, credit score, DTI ratio, interest rate, and loan term) in consideration. We used three different techniques (regression analysis, decision tree, and neural network) on the 4% of our dataset to train the data in SAS Enterprise Miner and to examine the results from all three techniques to choose the best option on the basis of misclassification rate and average squared error.

After applying all the models to the training data, fit statistics were examined in order to choose the best the model for our scoring data. Regression analysis scored least on mean squared error, and hence will be applied to the rest of our dataset. The average squared error comes out for all the three techniques (neural network, regression, and decision tree) as following 0.008, 0.007, and 0.02. After applying the regression analysis to the data we were able to figure out that 2.45% of our scoring data are most at risk for defaulting in the future.

The project will be fruitful to the company as it gives them areas of opportunity in terms of probability, cost savings, and allocation of resources. This project also allows them to section out particular loans as high risk and will be able to underwrite them accordingly.

# **Background and Objective**

The residential mortgage market is vital to the American economy. Without mortgages, most Americans could never afford to buy their own home. The size of the mortgage market and its effect on the overall economy is staggering.

At the end of the second quarter of 2018, there were $10.72 trillion in outstanding residential mortgage loans. During this same time, 1.05% of residential mortgage loans were in some stage of the foreclosure process. While this is a historically low percentage of loans in foreclosure, as we have seen in past years that can quickly increase during a recession.

For many, the economic pain of the 2007-2009 recession is still fresh in their minds. During 2009 alone, 1.05 million homes were lost to foreclosure. With an average mortgage balance of $217,000 this had a $228 Billion direct impact on the economy. This figure does not even begin to account for the effect it would have on dependent jobs, local communities, and industries such as mortgage brokers, bankers, real estate agents, home improvement stores, and property taxes.

The objective of our project is to use origination data from 2017 for single family mortgages to predict which loans are most at risk for defaulting in the future. By successfully creating a useable model in SAS we would ideally be able to reduce the number of future foreclosures in the event of an economic downturn.

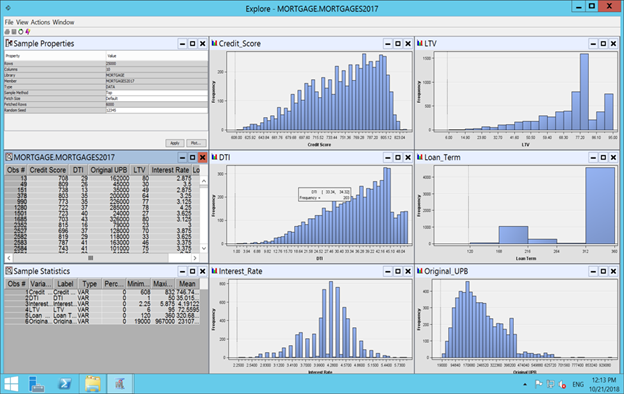
# **Data Source and Metrics**

The data we chose to use for our project was firsthand data directly from Freddie Mac. We were able to download a random sample of 25,000 single family mortgages that were originated in 2017. In total, the file contained 25 columns which we truncated down to 10 columns. The 15 columns that were eliminated were mostly descriptive or redundant data that added no value to our project. Examples of eliminated data include origination channel, servicer name, and metropolitan statistical area (MSA).

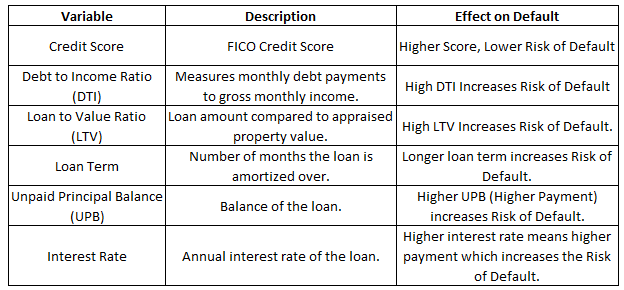
Of the ten fields that were left we ended up examining six of them for our project. A quick glance at the summary statistics in Exhibit A below shows some expected findings. The number of loans originated increased with credit scores. Most of the loans in this sample are prime loans with credit scores over 700 with our highest count being closer to 800. Our loan to value (LTV) ratio peaks around 80% which is expected since most lenders require mortgage insurance (MI) with anything less than 20% of the purchase price down payment. From the data set, we can see that most people came up with the minimum 20% requirement to avoid the monthly MI charges typically levied against a loan.

A strong majority of loan terms are either 15 year or 30 year mortgages which are typical mortgage lengths. The Debt to Income (DTI) ratio followed a very similar pattern as the credit scores where they increased at a linear rate until it peaks around 44%. The interest rates followed a rough bell curve showing that most of the homeowners in this pool were getting similar interest rates on their mortgages. Lastly, the original unpaid principal balances of the loans was skewed towards the left showing most homeowners at first glance appear to be borrowing responsibly.

**Exhibit A: Summary Statistics**



**Exhibit B: Variables Used**



Before we ran our datasets we checked for correlation between the six variables. If we had an extremely high correlation that was above 0.75, we would have examined both of those variables and likely removed one of the two to prevent our results from being skewed. Exhibit C shows the correlation between the six variables we looked at.

**Exhibit C: Correlation Coefficients**



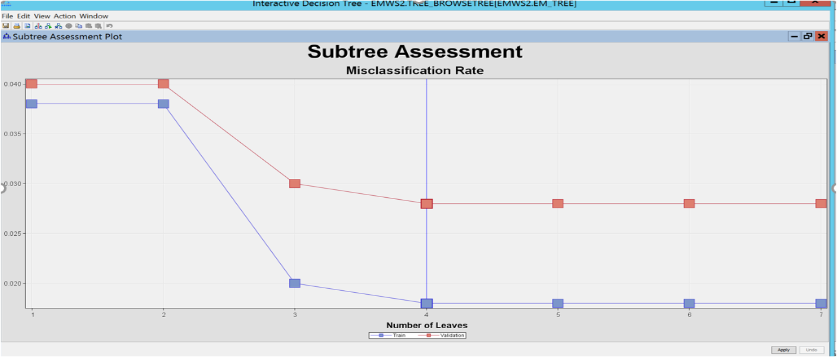
The only variables that showed a slight correlation were the loan term and interest rates. We expected some correlation between these two variables since historically the longer the loan term, the higher the rate of interest on the loan.

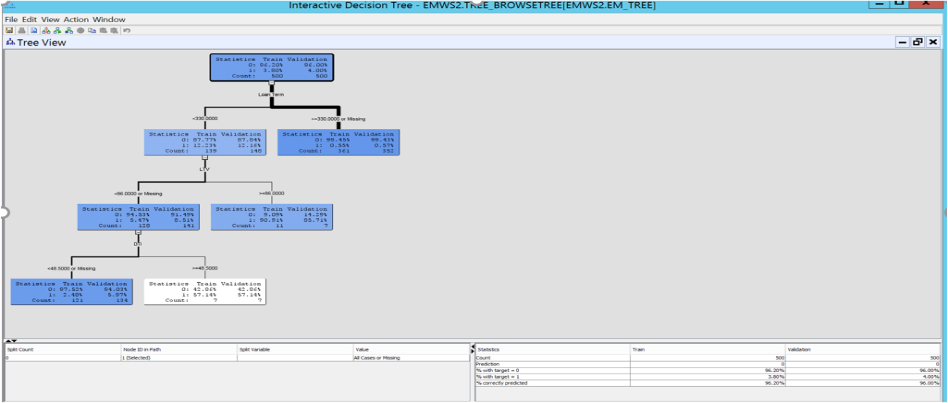
# **Training Model**

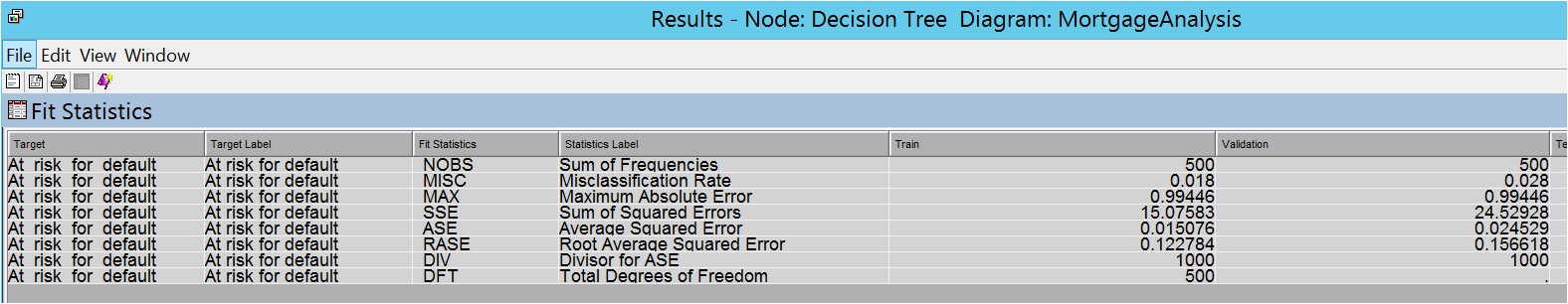
In order for our group to train the model, we had to begin by providing SAS with a number of training cases. Our training cases used the following predictors; Credit Score, DTI, LTV, Loan Term, Interest Rate, and UPB. These input variables were chosen due to their importance in relation to the likelihood of the approval of a loan. Our target or outcome from these variables was a 0 or 1. A 1 represents a loan that our model predicts has a chance to default with 0 representing a loan that our model predicts having little or low chance of defaulting. Our result was a decision prediction which fell in line with our objective.  
  
Our training model consisted of 1000 records which was complete with all the variables previously mentioned. 500 of these were randomly selected records for training the model and 500 of these were randomly selected records for validating the model. From the data set we used, we eliminated a number of inputs in order to eradicate any redundancies as well as to focus on the inputs which fell within our correlation limits.

# **Decision Tree and Results**

The first predictive modeling tool that we utilized was a decision tree. During the construction of our tree, we chose the subtree method of Assessment. The Assessment measure allowed us to see if the system has generated the smallest subtree with best assessment value or misclassification rate.

**Exhibit D**  
  
  
According to our subtree assessment, we created a predictive model which plateaus at a misclassification level equal to that of leaf 4. To create a tree that was more compact and ideally more accurate, the system selected a model with 4 leaves. The number of splits keeps our model more accurate and takes advantage of our knowledge of the misclassification rate.

**Exhibit E**  
  
  
The 3 levels in addition to our root node that our model predicted and utilized in crafting our decision tree were based on the 3 variables: Loan Term, LTV, and DTI. These splits yielded a 98.2% accuracy for our training data and 97.2% accuracy for our validation data. The fit statistics for our decision tree are as follows:

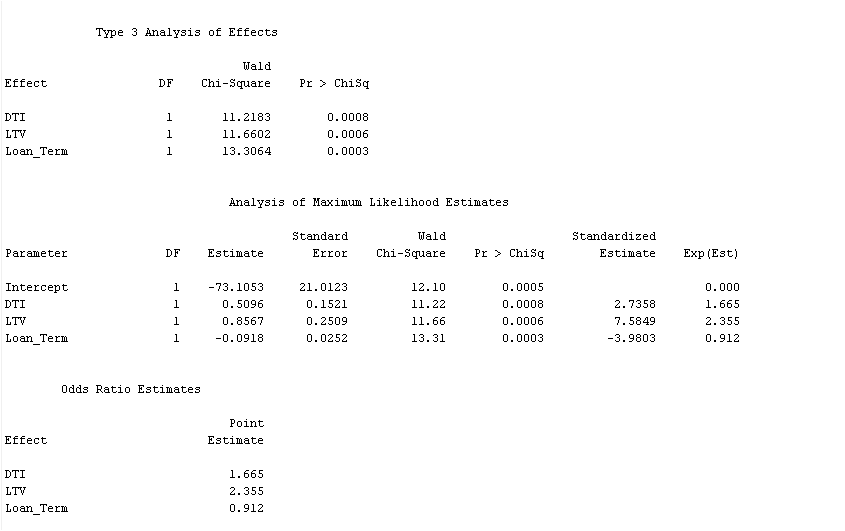
**Exhibit F**

The fit statistics of all our models will be discussed later in this report.

# **Logistic Regression**

Feeding training data set to the regression node has yielded the following values for intercept and estimates. Since the target is binary, we have chosen ‘logit’ function as the link function. Also, we have set the model selection method as ‘Stepwise’ and cut-off value for p as 0.05 for Entry and Stay significance levels.

**Exhibit G**



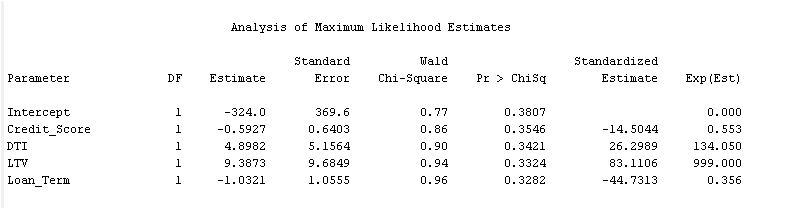
Interpreting the values of intercept and estimates gives the following equation,

At risk for default (0 or 1) = -73.1058+ 0.5096\*DTI + 0.8567\*LTV - 0.0918\*Loan\_Term.

Unit value increase in LTV has significant effect in increasing the odds ratio of loan being defaulted.

From the output, we could observe that entry of credit score has a negative effect in determining the intercept and estimate values. After the entry of credit score in the model, none of the estimated values satisfy the condition for ‘p-value’ (p<0.05)

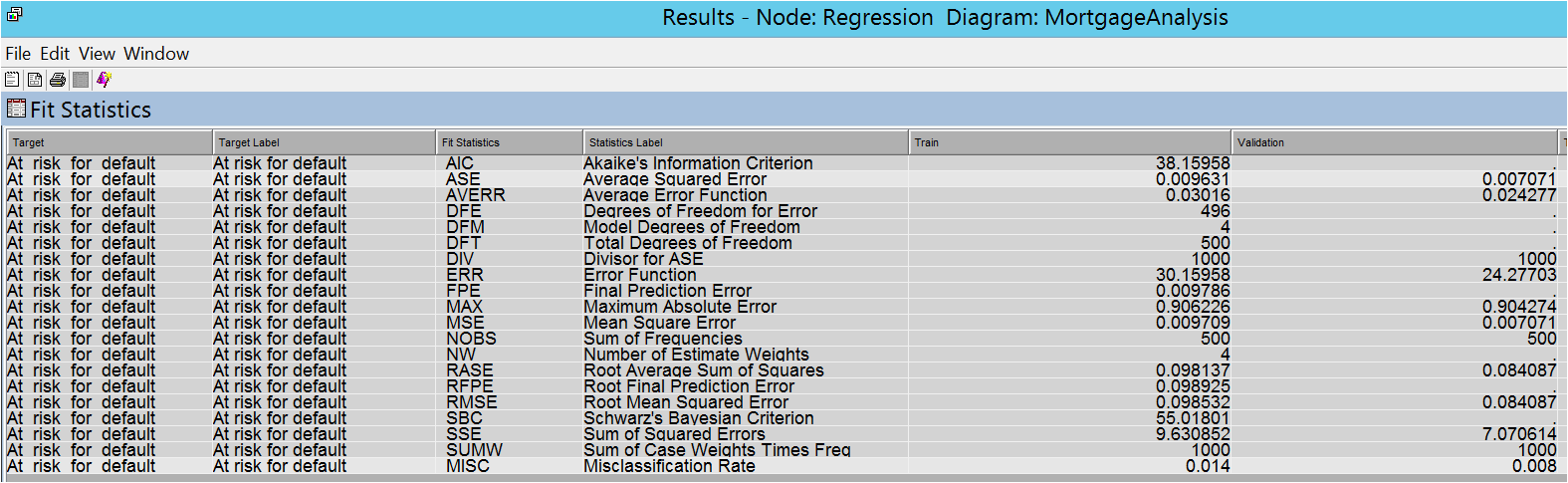
**Exhibit H**



Hence, credit score has been removedfrom the final model.

Fit statistics of our regression model are as follows:

**Exhibit I**

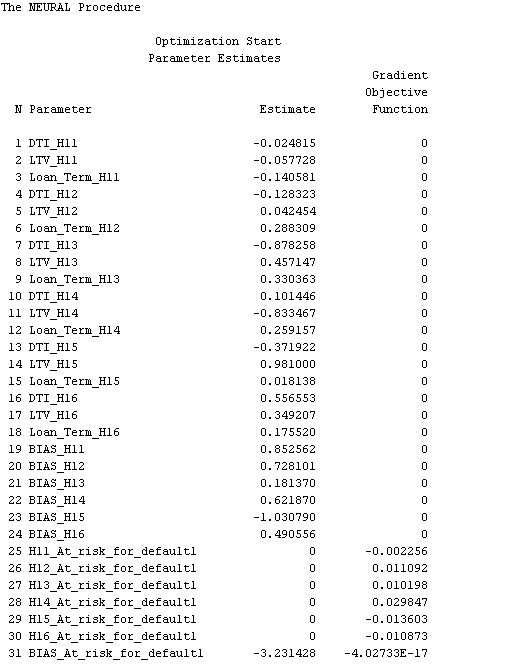


# **Neutral Networks**

Output of regression model is fed as an input to the neural networks. Neural networks do not have the ability to select and weigh only the significant inputs. Hence, the output of regression node is connected to input of neural networks node.

The model characteristics are set to build a ‘multilayered perceptron’ with 6 hidden units. The bias and average weights of input variables are as follows:

**Exhibit J**



Interpreting the bias and average weights gives us the following equations.

At risk for default = -3.231428 -0.002256\*H11+0.011092\*H12+0.010198\*H13+0.029847\*H14- 0.013603\*H15- 0.010873\*H16

H11= 0.852562-0.024815\*DTI-0.057728\*LTV-0.140581\*Loan\_term

H12= 0.728101-0.128323\*DTI+0.042454\*LTV+0.288309\*Loan\_term

H13= 0.181370-0.878258\*DTI+0.457147\*LTV+0.330363\*Loan\_term

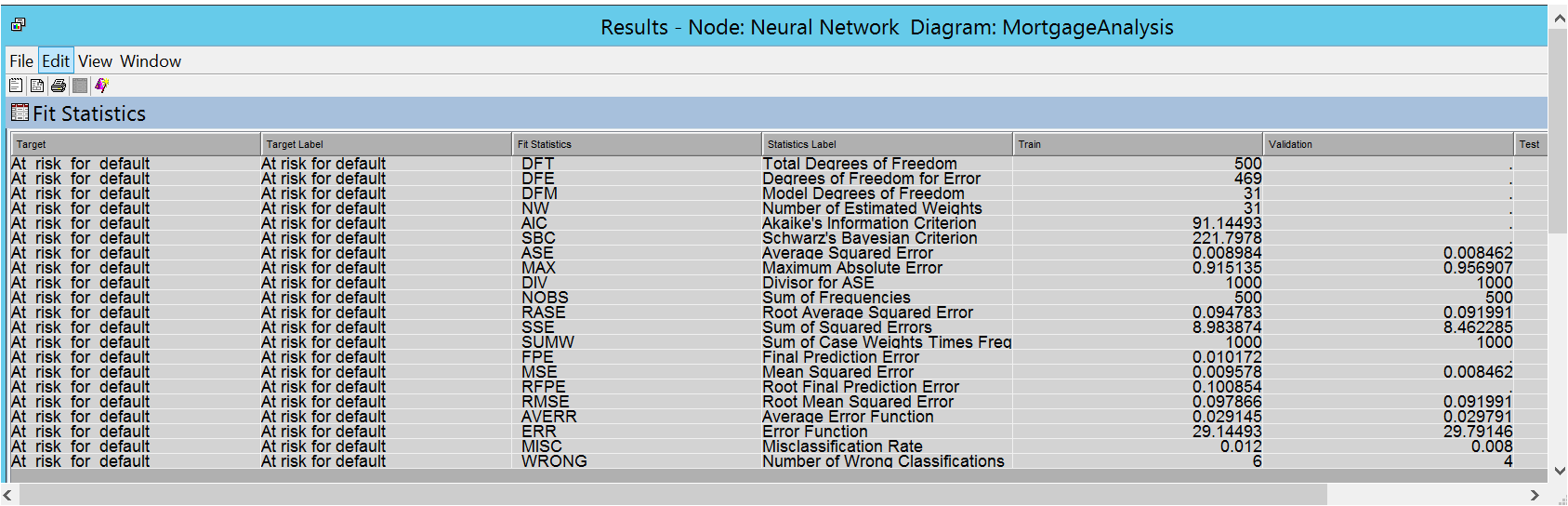
H14= 0.621870+0.101446\*DTI-0.833467\*LTV+0.259157\*Loan\_term

H15= -1.030790-0.371922\*DTI+0.981000\*LTV+0.018138\*Loan\_term

H16= 0.490556+0.556553\*DTI+0.349207\*LTV+0.175520\*Loan\_term

The fit statistics of neural network model are as follows:

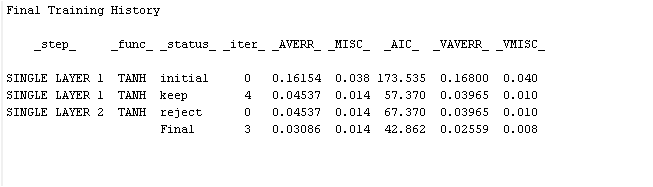
**Exhibit K**



We have also used a ‘Auto Neural network’ with 1 hidden unit and maximum number of 8 iterations. Only hyperbolic tangent function ‘TANH’ is used as an activation function.

Model in 4th iteration of 1st hidden unit has better fit statistics compared to all other models

**Exhibit L**

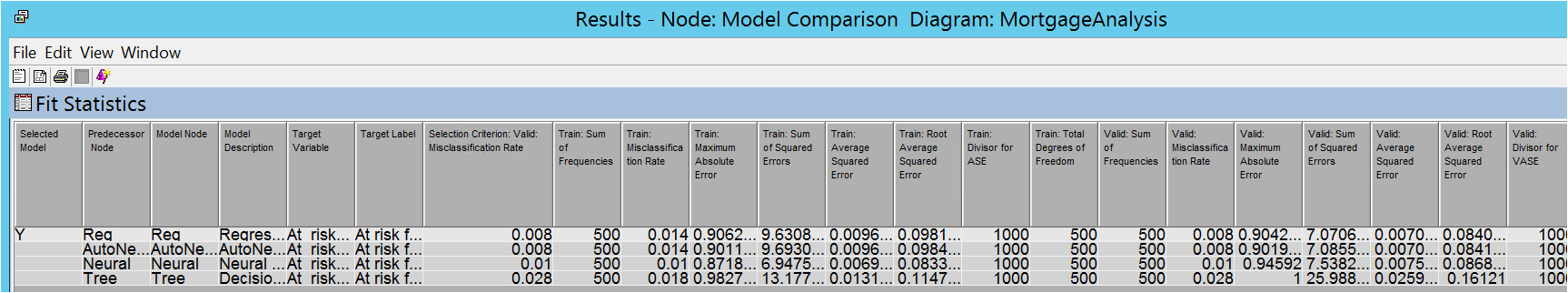


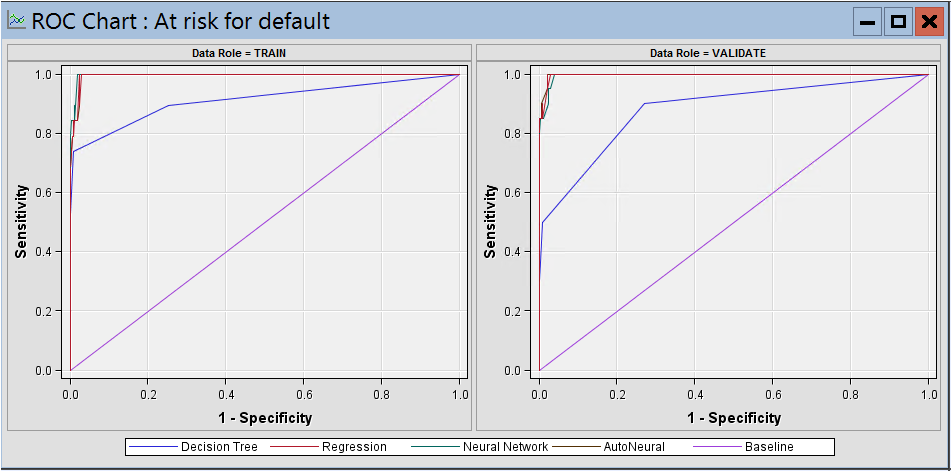
# **Neutral Networks vs Decision Tree vs Logistic Regression**

The ‘Model Comparison’ node has been used to identify the champion model out of 3 models. The node assesses the best model based on fit statistics and applies it to the scoring dataset.

In our case, the regression model is identified as the best model with the least values of ‘Validation misclassification rate’ and ‘Validation Average square error’. Hence, regression model is applied on our scoring dataset.

**Exhibit M: Results**



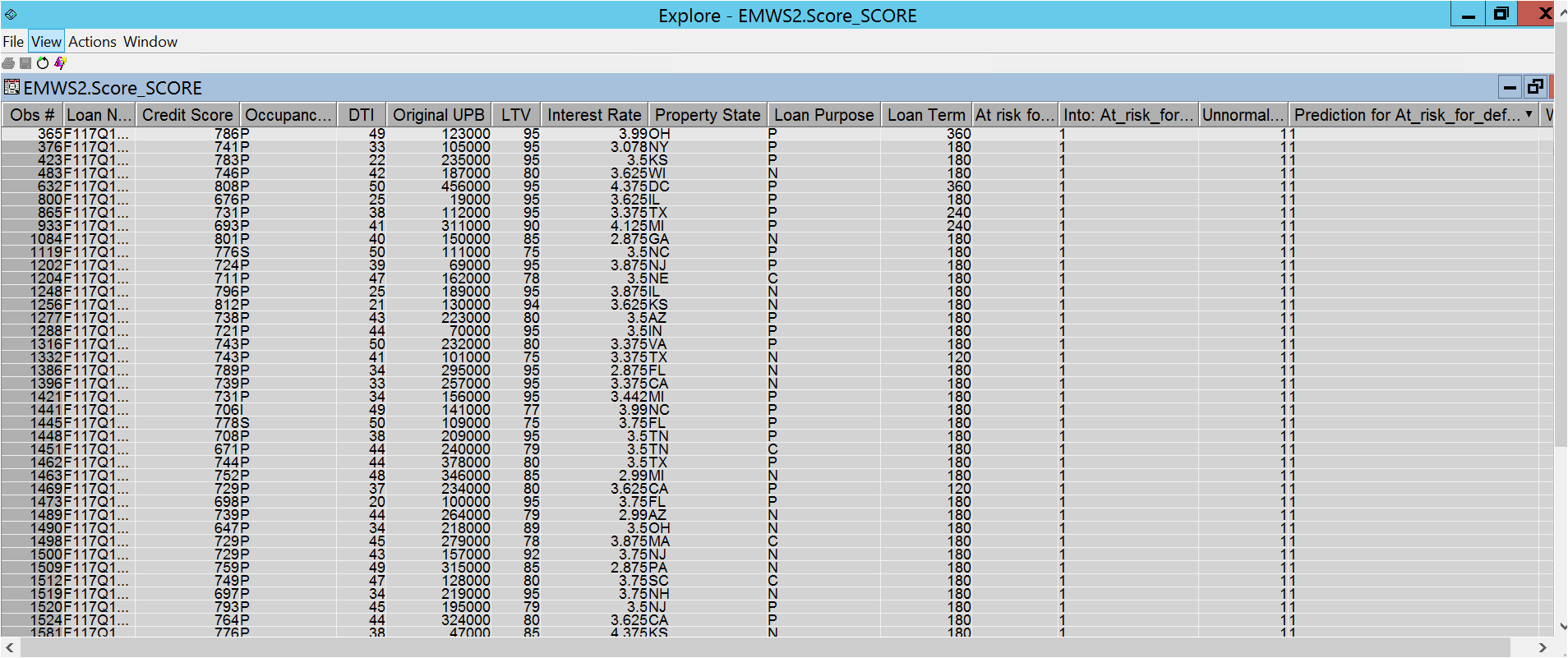
**Exhibit N**

We can observe that maximum ‘true positive’ rate has been achieved for minimal ‘false positive’ rate in case of regression model, which is the characteristic of an ideal model.

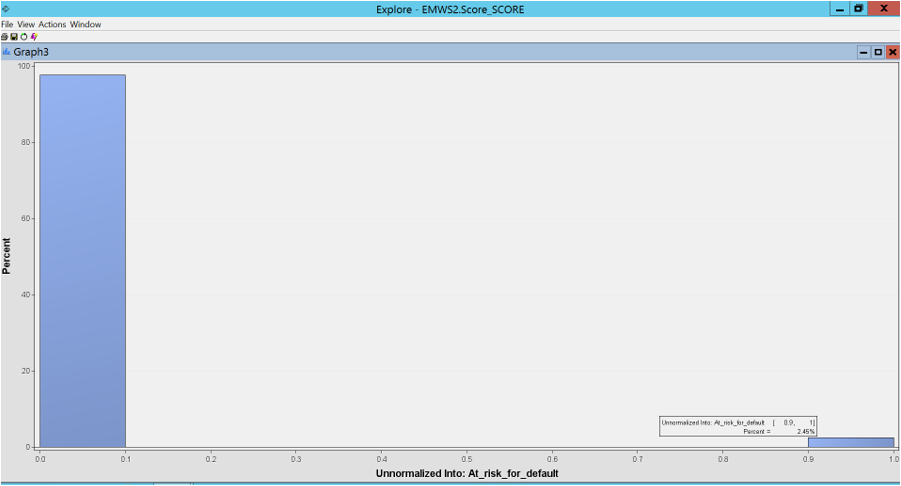
# **Applying the Model**

‘Score’ node from ‘Assess’ tab has been dragged and dropped onto the diagram. ‘ScoringMortgages’ data source and output of ‘Model comparison’ node are connected to the input of ‘score’ node. The champion model (regression) is applied on scoring dataset to predict the target variable (‘At risk for default’).

**Exhibit O: Results**

****

**Exhibit P**

****

Model predicts that 2.45% of total 24,000 loans are most likely to default in future.

# **Managerial Implications and Use**

Being able to flag so-called “high risk” loans predicted by our model will allow banks and mortgage servicers to carefully monitor these loans and immediately intervene if there are any signs of distress or early delinquency. Resolving these loans earlier will lead to lower rates of foreclosure, saved costs, fees to both the homeowner and the investors, and a better public image for the bank/servicer.

Also, if the model is successful in predicting default rates, the information taken from the model could be used in the future to modify underwriting practices. For instance, if the model is flagging loans with a high LTV, high UPB, and low credit score, the bank may require a larger down payment up front or they will not underwrite the loan. Successfully mitigating risk is extremely important for banks in the event of another recession.

# **References**

(n.d.). Retrieved from Federal Housing Finance Agency: https://www.fhfa.gov/Media/PublicAffairs/Pages/FHFA-Reports-Mortgage-Interest-Rates-April-2009.aspx

(n.d.). Retrieved from FreddieMac: http://www.freddiemac.com/research/datasets/sf\_loanlevel\_dataset.html

(2018, September 21). Retrieved from https://www.federalreserve.gov/data/mortoutstand/current.htm

Ahmad, A. (2018, August 16). *Mortgage Bankers Association*. Retrieved from https://www.mba.org/2018-press-releases/august/mortgage-delinquencies-down-in-2nd-quarter-of-2018

Andrews, J. (2018, June 13). Retrieved from Curbed: https://www.curbed.com/2018/6/13/17451278/foreclosure-rates-lowest-levels-housing-collapse

Bernstein, S. (2017, January 11). Retrieved from Reuters: https://www.reuters.com/article/us-usa-foreclosures/u-s-property-foreclosures-at-10-year-low-in-2016-idUSKBN14W0HH

*The World Bank*. (n.d.). Retrieved from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US