Group 18:

Daniel Rubio | Mohammad ALRUKHAIS

Research on BI Applications

Mortgage Analysis

# Executive Summary

This report was written to satisfy the University of Texas at Dallas BA with SAS project requirements named Research on Business Intelligent Applications. Mortgage defaults and pay-offs were analyzed for 41,747 borrowers and used to create a regression, decision tree, and neural network model to predict the status of a borrower given certain variables such as, GDP growth, interest rate, and unemployment rate.

We found that the neural network model was the best model created in SAS Enterprise Miner, with a misclassification rate of 21.112%. Not much can be interpreted from the model, however, because it is a black box. We recommend using the neural network model for accuracy and application. However, to better understand how the variables interact with each other a regression or decision tree model is recommended. Using the regression model and decision tree lenders can anticipate the status of borrowers in their current portfolios’. This information will help lenders re-align their portfolios to achieve the optimal portfolio depending on their risk-appetite levels.

# Project Motivation / Background

Both Mohammad and Daniel have a background in dealing with financial data and thought it would be appropriate in using such data. We found mortgage data from creditriskanalytics.net. What makes mortgages interesting to us is that they are highly associated with the last financial crisis in 2008.

In 2008 there was a significant amount of defaults occurring in the United States, primarily because of how the mortgage-backed securities were being represented to investors. Had investors investigated these portfolios, a detailed analysis would have seen the shortcomings of such securities. Our goal is to try to predict the status of a borrower, that is whether they defaulted or paid-off their mortgage. Predicting the status of borrowers will help lenders mitigate the risks they face and optimally making better decisions on their originated loans.

# Data Description

## Overview

The panel data consists of 50,000 borrowers and over 600,000 instances total with an average of 24 instances for each borrower. The time period of the data is unknown, as it was intended since when the data was collected shouldn’t be relevant, however, time intervals are used so that it would be possible to conduct a time series analysis.

It is important to note that because our analysis does not take a time series analytical approach, we decided to only analyze the final observation of each borrow that had a default or payoff. This limits the data analyzed to only 41,747 borrowers and instances, the remaining 8,523 borrows continue to pay their mortgage and are not analyzed. There is a total of 15,158 instances where a borrower defaults and 26,589 instances where a borrower paid-off their loan. It is not possible to distinguish between paying-off or refinancing of a mortgage with the data provided.

## Key-Variables

Some notable attributes the data contains are: whether a borrower defaulted, paid off, or made a monthly payment, Loan type, Interest at observation, Unemployment rate at observation, GDP at observation, House Pricing Index (HPI) at loan origination, HPI at observation, FICO score of borrower at loan origination, Loan to Value ratio at time of loan origination, and Loan-to-Value at observation.

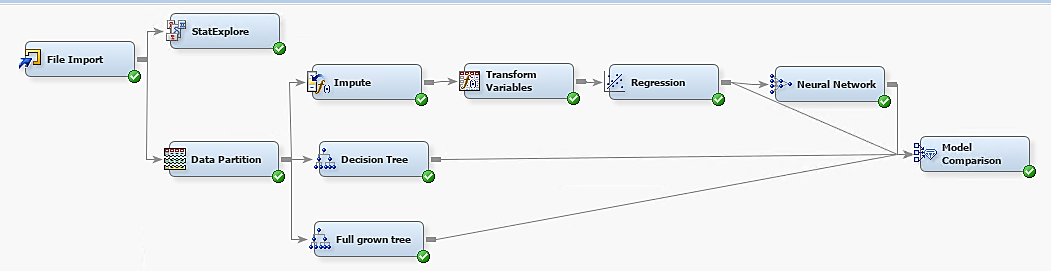
## Variable-Specific Findings

Something that was expected is that the average FICO score at loan origination was about 10 points below average for someone who defaulted, and the score was about 6 points above average for someone who paid off their loan.

The LTV ratio is a formula commonly used in the finance industry to assess the risk of a given loan by taking the mortgage amount / property value. The average LTV at observation in the data set is 80.967, and as expected, the average LTV is at 96.94 when a person defaults. This means that the property value is decreasing at a faster rate than the mortgage amount (which could be explained by an economic downturn), or the mortgage amount is increasing at a faster rate than the property value (which could be caused by interest accruing on missed payments).

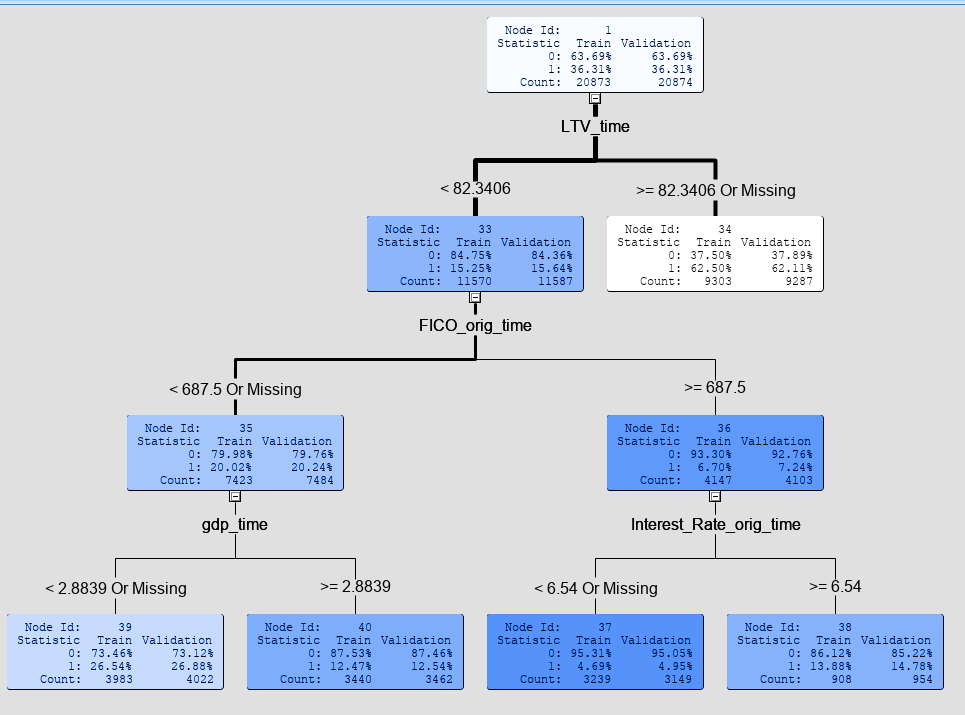
Lastly, the GDP at observation time is another expected indicator that will be important in determining a borrower’s status. The average GDP growth is 1.47%, and when a borrower pays-off their mortgage the economy is fairing better than average at 2%. And as expected, the GDP growth is below average, at 0.5%, when a borrower defaults on their mortgage.

# BI Model’s



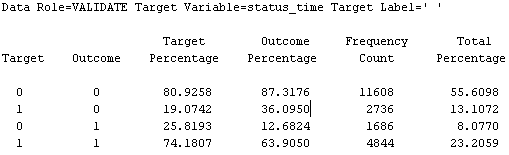
## Decision Tree

As shown in the decision tree below the first spilt occurs using the variable LTV (loan to value time of observation). When LTV is greater or equal to 82.3%, individuals who pay off are 37% while individuals who default are 62.5%. On the contrary when LTV is less than 82.3%, individuals who pay off are 84% while individuals who default are 15.25%. The second spilt occurs for individuals with an LTV than is less than 82.3, this split occurs based on FICO score. Within the sub-group of individuals who have a FICO score greater than or equal 687.5, 93.3% pay off and 7.24% default. Whereas when individuals have a FICO score below 687.5, 79% pay off and 20.24% default. Two more splits occur on variables interest rate and GDP which changes the percentages of defaults and payoffs further.



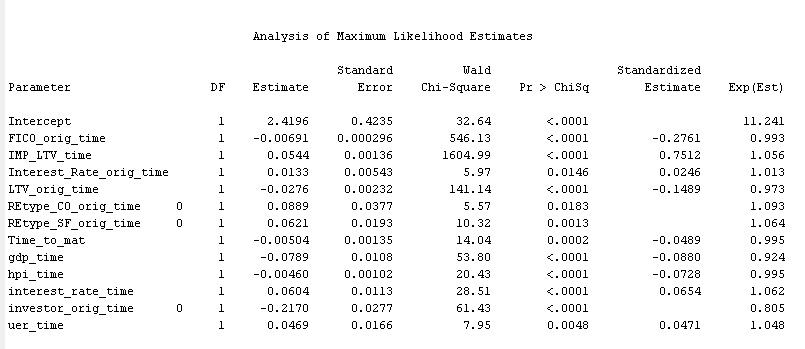
## Neural Network Model

The Neural Network (NN) model was created after a regression node was ran so that we only use the same variables in the regression, this helps optimize the neural network. Initially, we were afraid that the misclassification occurred mostly with the default target (1), however, we see that it was only 36% misclassified in this area. This gives us some confidence that the neural network can accurately predict when a borrow is going to default or payoff.



## Logistic Regression Model

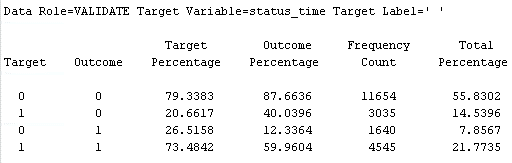
Because our target variable is binary, it is not reasonable to conduct a linear regression. Instead, a Logistic Regression node was ran after we standardized the data in the “Transform Variables” node. We had to standardize the data because of the different scales used in the variables, including FICO scores, GDP % growth, and mortgage balances. Within the node we specified that the selection model criteria would be stepwise to ensure that SAS will begin my removing variables that did not seem important in explaining the status of a borrower.



Shown above are the variables used in the regression, we were left with 12 variables. When the estimate has a negative sign, it shows that there is a positive relationship between the variable and paying off (status 0). The results of the regression seem to be consistent with what we have found using the decision tree, in terms of the relationship of a variable with the status of the borrower. For example, we have previously determined using the decision tree that a high FICO score will reduce the probability of a borrower defaulting, this is also shown by the regression analysis.

A picture containing text, receipt

Description automatically generatedWhen we take a look at the odds ratio we see some dramatic results. When the unemployment rate increases by 1% point the likeliness of defaulting increases by almost 4.8%, a non-investor borrow is 35.2% more likely to default than an investor borrower, and for every point a borrower’s FICO score increases the likelihood of defaulting decreases 0.7%.We also see that mortgages for with the underlying asset being a Condominium are 19.5% more likely default and Single-Family homes are 13% less likely to default. Lastly, we noticed that for every month the loan is close to maturity the likelihood of a default decreases by 0.5%.

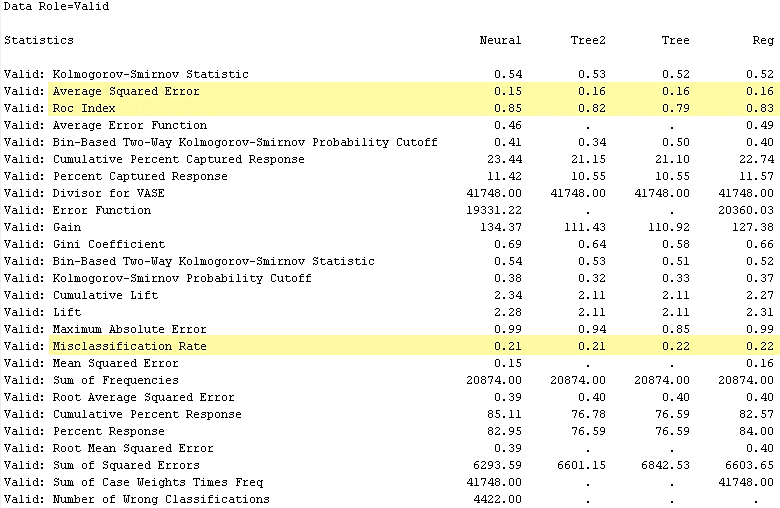


Above, we have the detailed misclassifications in the regression model. Most notable, we see that the model is not able to correctly predict the defaults on a loan 40% of the time and can accurately predict defaults almost 60% of the time. Overall, the model had a 22.3963% Misclassification rate, which we think is a respectable rate in predicting the status of a borrow.

# Conclusion

## Comparison

To ensure that our regression model is can be used to make predication, we have compared it with other models. Above are the results of comparing Neural network analysis, a full-grown tree, a pruned tree, and our regression model. Although, the best model is most scenarios would be the Neural Network Model. It has the lowest misclassification rate at 21.112%, and the highest ROC index at 0.85. The high ROC indicates that a risk adverse agent would also prefer the NN model because of the ratio of True Positives to False Negatives and True Negatives to False Positives. The model is a black box, and no predication can be made with it.



Moreover, the results of our regression model and decision tree, specifically using ASE, ROC index and Misclassification rate are very similar to the neural network model. This gives us confidence that our models can be used to make predictions on the status of borrowers.

## Uses of the models developed

Lenders face two kinds of risks, default risk and prepayment risk, both of which can be predicted using our models. Knowing that a borrower has a higher probability of default will allow lenders to increase provisions for such borrowers. Further, lenders can charge higher interest for such borrowers in the future so that they are fairly compensated for the risk these borrowers pose. The second type of risk, prepayment risk can also be medicated using our models. The model can be used to predict which borrowers will pay early. This will help lenders manage their assets and liabilities so that their durations match, a concept known as balance sheet immunization. The model also considers economic variables and how they effect the status of a borrower. This will allow lenders to hedge against their portfolio if they expect any changes to occur in the economy overall.

# References

*B. Baesens, D. Roesch, H. Scheule, Credit Risk Analytics: Measurement Techniques, Applications and Examples in SAS, Wiley, 2016*