**Final Project Report**

**MI-353-01-22SPTD Statistical Analysis & Pred Mo**

Group 4

Donald Bookman, Christopher Gray, Ryan Johnson, Alex Rodriguez

5/5/22

**Executive Summary**

For the final project, we chose a second hand dataset found on Kaggle.com. The dataset is from the U.S. Small Business Administration (SBA), an organization founded in 1953 to aid small businesses in their ventures in the U.S. The data collected displays small businesses and their loan applications. As a group, we decided to run four different analyses to better understand the information from our dataset. The four analyses that we ran were a neural network, a Ward cluster, a decision tree, and a logistic regression test. We implemented a neural network model to handle a wider variety of nonlinear relationships from our data. We performed the ward cluster to determine clusters from our data. In this case the clusters were formed based on the squared error. We utilized a decision tree to determine every possibility based on the data. Lastly, we did a logistic regression test to determine the outcome of individual loans. More specifically what variable resulted in loans being paid off or not. The results of these tests are intended to help us understand the results of a loan being paid off and a loan that is charged off. After the test was finished we retrieved numerous results. The neural network model told us that a majority of the SBA’s paid off their loans fully. The cluster analysis told us that the higher the loan amount is, the more likely it is an established business. This then increases the probability of them paying off the loan. The logistic regression helped explain the variation in our target variable. Lastly, the decision tree showed us that the length of the term is significant. It told us that term length is important in determining if a loan will be paid off fully or if it will be charged off. This all in all helps us solidify our assumptions about the potential variables that determine the results of a loan being processed and ultimately being paid off or charged off.

**Project Motivation/Background**

The project was based on data from the US. Small Business Administration (SBA). This is an organization that assists in the growth of small business start ups. Small businesses have been an important contributor of jobs for the world. This allows growth of the company as well as help decrease unemployment. The SBA has seen startup companies flourish from the loan support that the SBA has given them such as Fedex. They have also seen companies fall through with their guaranteed loans. To prevent the SBA from losing money they developed a system to help determine whether a company is fit to receive a loan or not. So the content of the data shows the process that the SBA does to either approve or deny a loan to a small business. It also shows the results of a return whether it was paid off fully or if it was charged off. We as a group were very motivated in regards to the effort and results that came out of this dataset. As a group we utilized this dataset as a way to better understand the functions of small businesses as well as better understand the process of loan applications.

**Data Description**

In our data an important part to talk about is what “charged off” is and why it’s important. A charge-off, also known as a charged-off account, is a debt that has been so overdue that it has been removed off the balance sheet by the creditor. Specifically for the bank it means The account has been deleted from the bank's accounting records and is no longer considered an asset. Until the charge-off account is paid off, it will have a negative influence on his credit record. It indicates the obligation has been unpaid for such a long time that creditors have designated it as a bad debt.Your debt however is not forgiven just because it has been charged off. Don't think that just because your creditor wiped off your debt, you don't have to pay it. You're still legally responsible to pay back the amount you owe as long as your charge-off remains unpaid.

**Data Preparation Activities**

To prepare the data we had to condense the data to make it easier to use. We cleansed the data by reducing the duration of the years column in the dataset. We changed the starting year for the data from 1997 to 2005, and only used Pennsylvania data. This reduced the row count from around 890,000 rows to about 7,600 rows. This will allow us to focus on banking data that is more recent and specific to our state. We use many different target variables such as geographical information like the city and the state where the check was approved, along with the location of the bank headquarters.We will also include the approval date of all the checks including the day, month, and year they were approved. Next, we prepare to use logistic regression to predict future outcomes. We decided to focus on MIS\_Status to determine if the loan will be paid off or charged off but the charge off looks to be a rare occurrence. We also had to filter to get rid of outliers and missing values. After doing this the MIS\_Status paid in full count to 5637, 1919 were charged off, and 2444 were missing. The percentage of variables missing was 24.44%, the percentage of loans paid was (5637/5637+1919) 65.96%, and the percent charged off (1919/5637+1919) is 34.04%.

**Models/Enterprise Miner Diagrams and Findings**

**Neural network**

As a part of our analysis, we chose to implement a neural network. These are a class of parametric models and hanle wider varieties of nonlinear relationships. Constructing a neural network model for the SBA national dataset involves two main phases. Defining and training the model is of vital importance when operating with neural networks, especially in the magnitude of the large dataset we are using. Yes, neural networks can be much more intricate to explain to the management of an SBA than analysis like a linear regression or a decision tree, but this is the far more superior and more detailed method of analysis. There is a much more significant verbal explanation of visuals than statistical representation when it comes to neural networks, which is while you will see this data presented in a much different way than the other analysis. Running a neural network is not a hard manual task. The diagram consisted of a total of 3 nodes; Import node, Data Partition node, and the Neural Network node at the end. Being that the dataset we are working with is very large and consists of many target levels, certain variables of non-significance needed to be rejected in order for the process to successfully run. After setting up the process and running the neural network, the results validated our predetermined notions regarding MIS\_STATUS. Our target variable for this analysis, along with other analysis performed in our project, is the MIS\_STATUS variable. As a group, we deemed this variable as the most significant indicator in our data because it provides us with the current status of an SBA’s loan agreement. The neural network results showed us that there is an overwhelming majority of SBA’s who successfully pay their loans in full. Because of this, we see an adverse effect in the GrApprv where those who successfully pay their loans in full tend to have larger gross loan amounts from their pre-approved bank. It is also important to note that for our neural network we received a standard squared-error of 0.005841 which is significant.

**Cluster Analysis**

We chose to run the ward cluster analysis in our analysis, which determines clusters based on the sum of squared errors. In running this analysis, we started with the minimum number of clusters at 2 and the maximum at 20, and made sure to standardize our variables, in order to reduce potential biases within our results. When running the analysis, we received only the minimum number of clusters of 2. These clusters varied widely in size, as cluster 1 had a frequency of 806, while cluster 2 had a frequency of 5613. Next, we looked at the variable importance that contributed to the formation of the 2 clusters, and found 6 variables with an importance over 0.9. These variables are CreateJob (1), DisbursementGross (0.983), SBA\_Appv (0.978), GrAppv(0.969), ChgOffPrinGr(0.922), and Term(0.901). 4 of these variables deal with financial data, while CreateJob displays the number of jobs created, and Term is the length in months of the loan. All of these variables are significantly higher in the first cluster than the second cluster, seemingly distinguishing larger investments in the first cluster. The images for this section are provided in the appendix below.

Next, we increased the minimum number of clusters to 5, to see if there were any changes in the variables that had a significant importance within the formation of the clusters. In this analysis, we were also returned the minimum number of clusters, this time being 5, with one cluster having 4258 data points, and the rest being as follows: 972, 402, 439, and 348. This new number of clusters gave us a very different set of variable importance measures, with only 2 variables being above 90%, LoanNr\_ChkDgt at 1 and DisbursementGross at 0.989, as compared to the 6 variables in the 2 clusters of the previous analysis. DisbursementGross, which is the total amount of the loan disbursed, was the only variable to receive a variable importance of 0.9 or above in both analyses. This makes sense, as the amount disbursed is key to whether a business will be able to pay off the loan, and is an important factor when banks decide whether or not to approve the loans. The mean DisbursementGross has a wide range between the clusters, with cluster 1 having the lowest at $57,588.97, while cluster 5 had the highest mean DisbursementGross at $475,835.40. The potential assumption here is that the higher the bank loan amount, the more likely it was going to a more established business, and the more likely it will be paid off.

**Decision Tree**

When running the decision tree, we used average square error as the model assessment statistic, and we made the maximum branch split to be 2. We also used MIS\_Status as the target variable, and connected the data partition node to the decision tree node in order to better train our data. We next had to reject the ChgOffPrinGr and ChgOffDate variables, as these are not important variables because they both deal with charge offs, but we are looking for what determines if a loan is paid in full or charged off, not the description of an already charged off loan. In this decision tree, we obtained 22 leaf nodes, and the variable used in the first split was Term, with the split being less than 59.5 months, or greater than 59.5 months. The competing variables for the first split and their variable importance measures are as follows: BankState at 0.6260, Bank at 0.2114, SBA\_Appv at 0.1840, LoanNr\_ChkDgt at 0.1505, and NoEmp at 0.1292. Term had a variable importance of 1, so it contributed the most to our decision tree. However, an interesting decision we found within the tree was that the first split that was based on BankState split the variable into California, and then the rest of the states. Overall, the decision tree shows us that the length of the term is important in determining whether a loan will be paid in full or charged off.

**Logistic Regression**

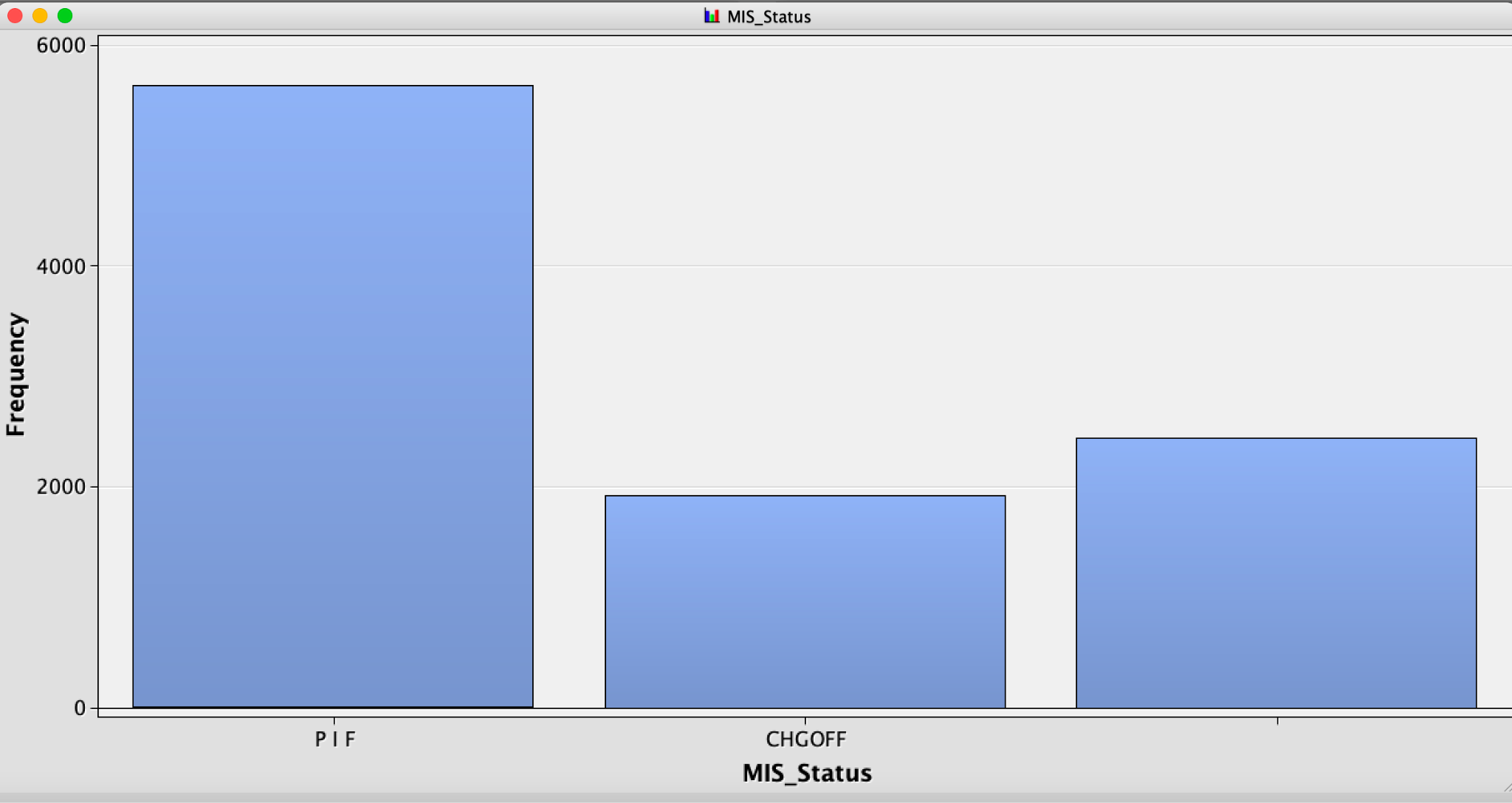
For this analysis, we chose to use the target variable MIS\_Status, in order to determine what variables may lead to a loan being paid off or charged off. Since this a nominal variable, we had to use logistic regression instead of linear regression, due to the potential issues of heteroskedasticity and non-conforming probabilites when using linear regression on nominal target variables. When running this analysis, we first added a data partition node to help the logistic regression model be trained. The misclassification rates for the training and validation set were very encouraging, as they were similar and low, with the training misclassification rate being about 0.012, and the validation misclassification rate being about 0.030. This tells us that our model is well trained and classifying relatively correctly, as our test misclassification rate is also low at 0.032, however, this may be a Next, we wanted to determine if this model was statistically significant. For this, we looked at the Likelihood Ratio Test for the Global Null Hypothesis (which is shown in the appendix below). The p-value is less than 0.05 which tells us that this model is significant, and that this model can help explain the variation in our target variable of MIS\_Status, essentially explaining whether a loan will be paid off or charged off. Next, we look at the Type 3 Analysis Effects section, pictured in the appendix below, to see which variables are statistically significant (having a p-value of less than 0.05). The statistically significant variables in this model are Bank, ChgOffPrinGr, GrAppv, and Term (the degrees of freedom for these variables are 106, 1, 1, and 1, respectively). These variables, within our model, are statistically significant in explaining the variation in our target variable, MIS\_Status.

**Managerial Implications/Conclusion**

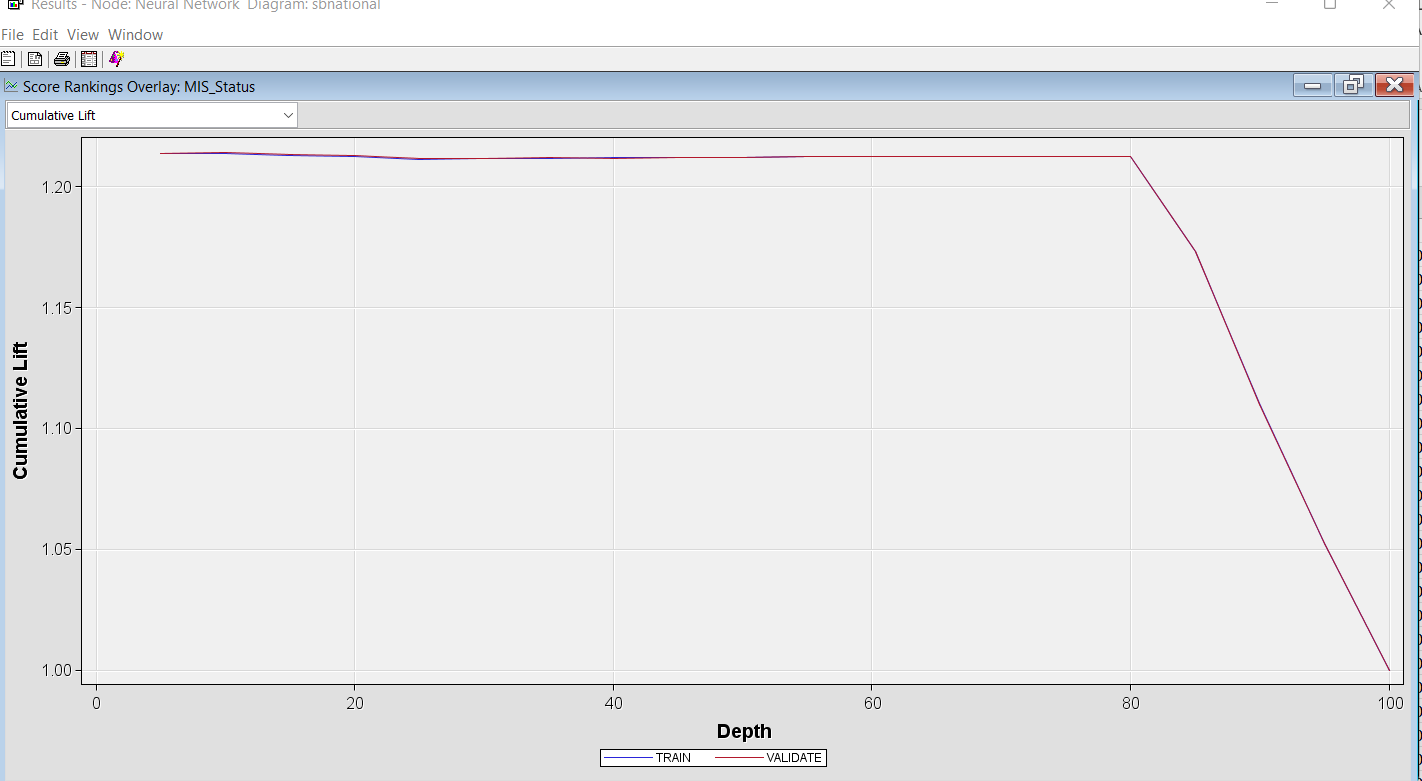
Performing multiple analysis with our U.S. Small Business Administration data not only provides more insight on the data itself, but it also equips us with alternate conclusions and predictions that are able to be drawn through the different analysis we ran. It is important to note when reviewing our specific dataset that for small businesses administrations it is always better to pay off debt in full. Like almost every administration in the country, some kind of loan is needed to fully allow operations and services to have a good chance at success in their specific sector. Paying in full and charging off accounts have their own distinctions. On the credit report of small businesses, the charge-off will be marked as “settled” whereas paying a loan in full is marked as “paid in full”. These marks send different messages to future lenders and as described above, the MIS\_STATUS in our data shows whether the administration paid their debt in full or charged it off. This is the main focal point and target variable for our analysis is the MIS\_STATUS variable. From a managerial viewpoint, this variable determines how deep an SBA’s wallet can go. All SBA’s have pre-approved loan agreements with different banks deterring the specific needs and ambitions of the particular administration. This original loan agreement is a test loan in the eyes of the bank. They want to see how exactly your business manages this money, uses the money, and pays back the money. This is why MIS\_STATUS is so important. In order to have a full picture of a business’ financial situation you have to understand how finances flow within the administration. By paying loan accounts off in full, more funds are more willingly allocated to the specific administration because the banks have tangible proof that they can maintain, pay back, and generate more revenue from the funds loaned to them. Charging off loans as small businesses can have serious and threatening effects on businesses overall ability to seek financial assistance. A charged-off account will be reported to the major credit rating bureaus and remain on that administration's credit history for seven years, making it extremely hard for that same company to get new loans for a long time. As managers of SBA’s, paying loans in full has to be the number one priority of a small business. The underlying problem with SBAs is that they are relatively small in scale, so growing organic revenue is a very hard feat to do without some sort of assistance. Not only are loans vital to the growth and future of the administration, but it also provides a safety net for the business in times of crisis and low revenue; this is something you want/need to have as an SBA.

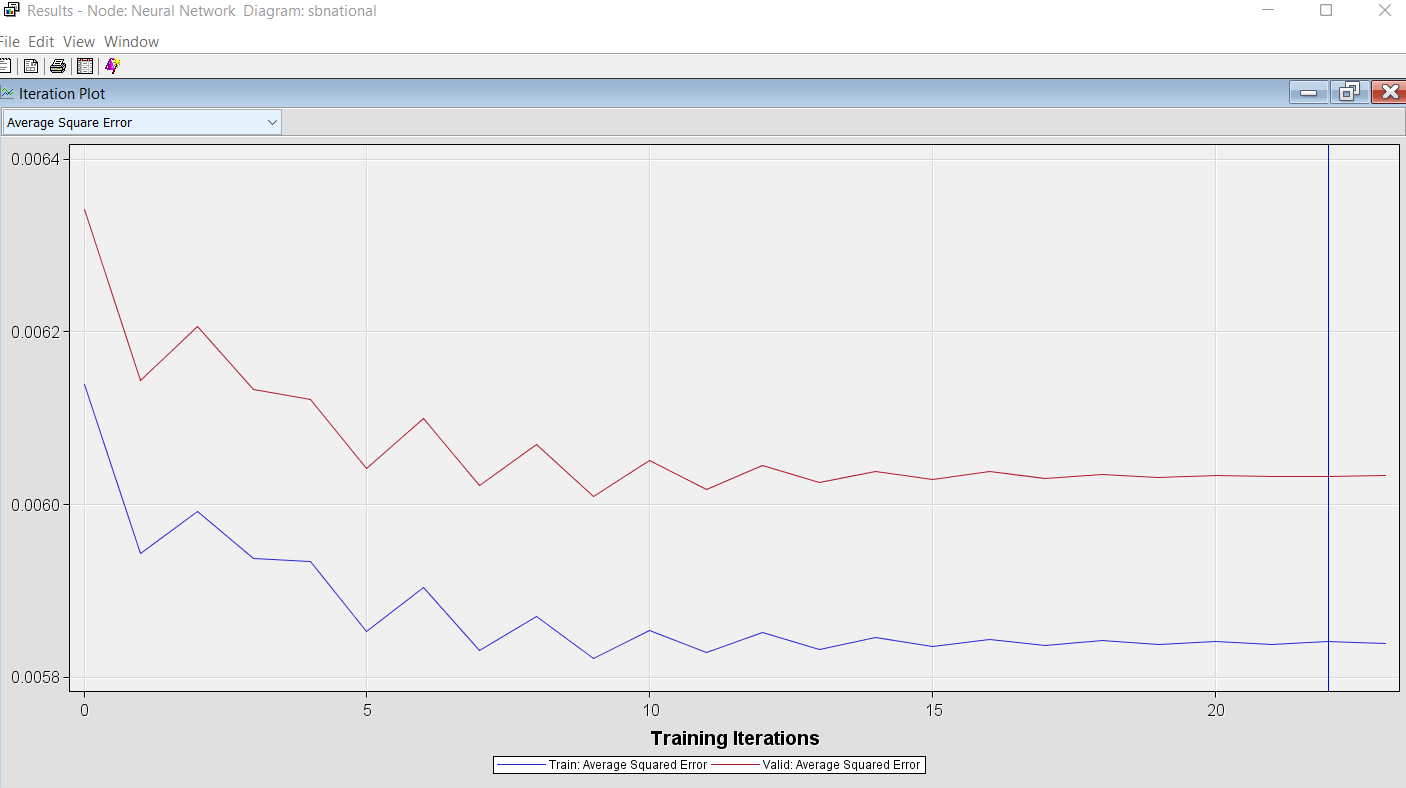
**Models/Diagrams**

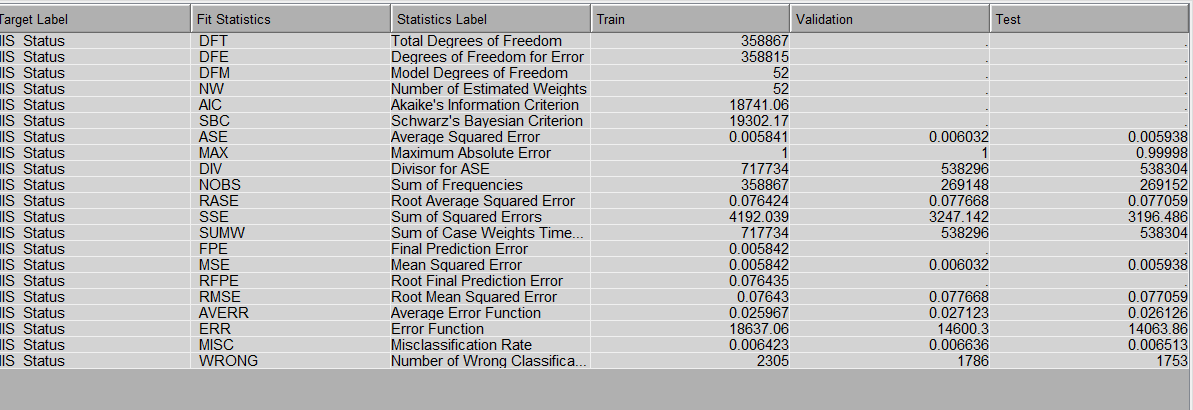
**Exploration of Target Variable MIS\_Status**

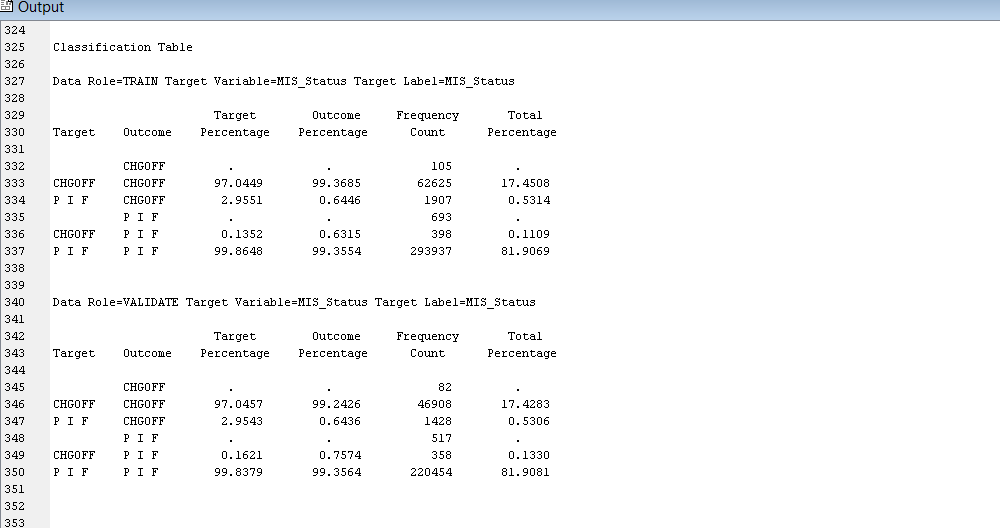
****

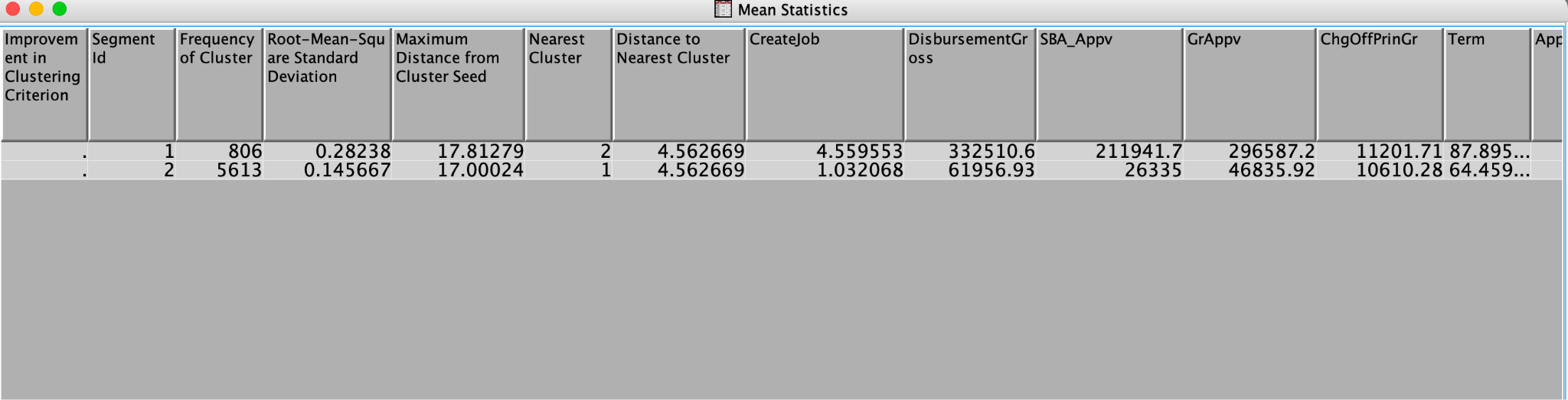
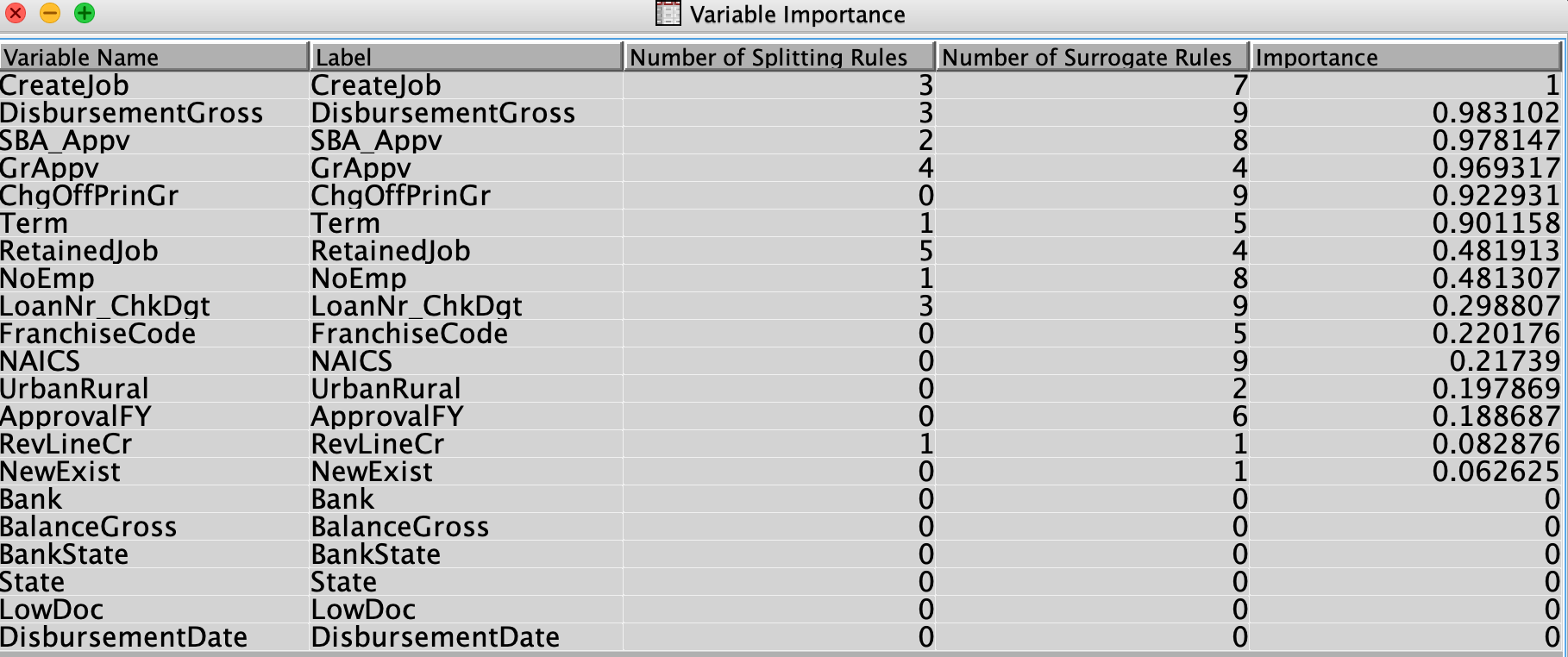
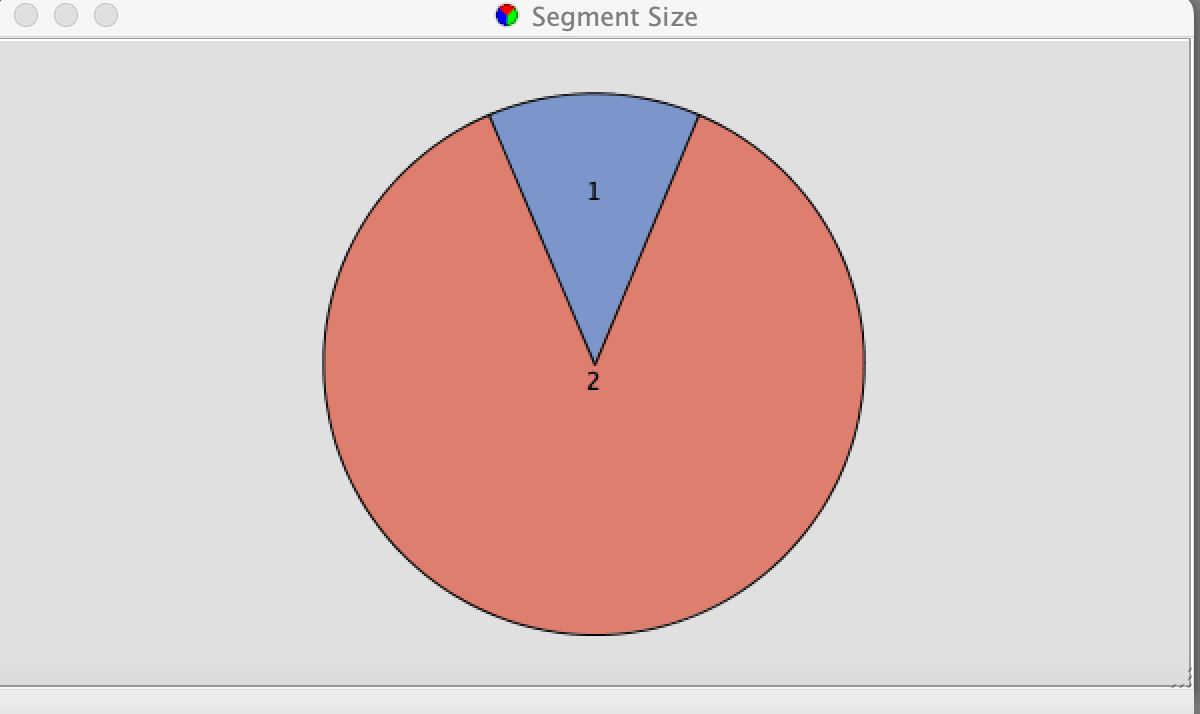
**Neural Network**

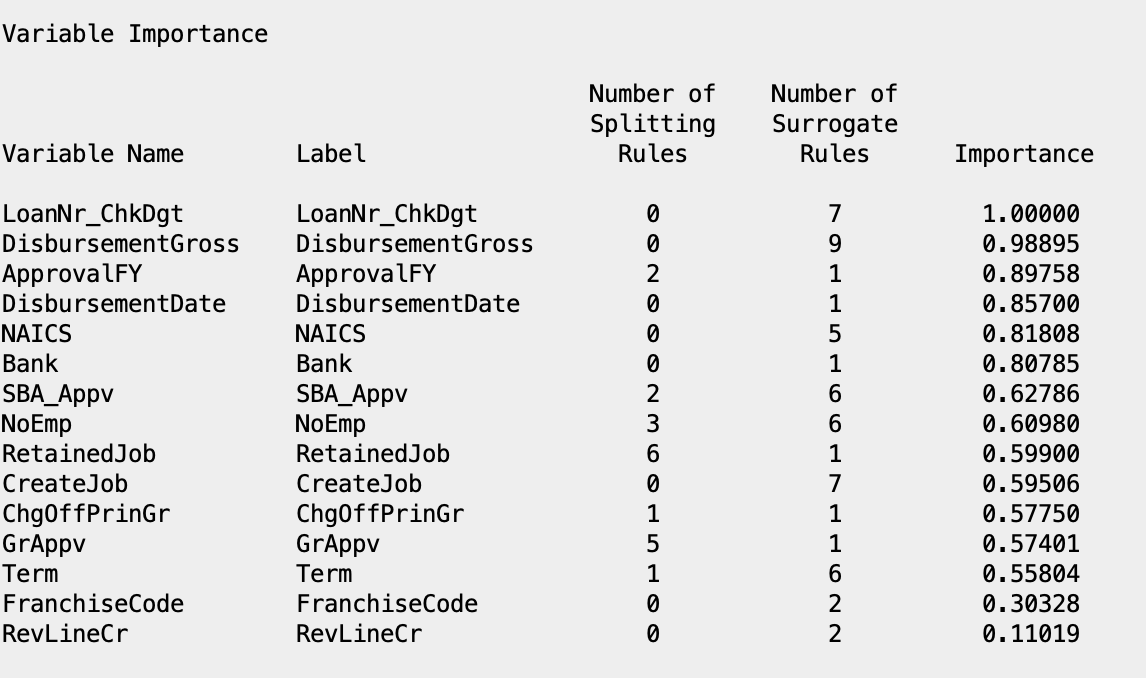
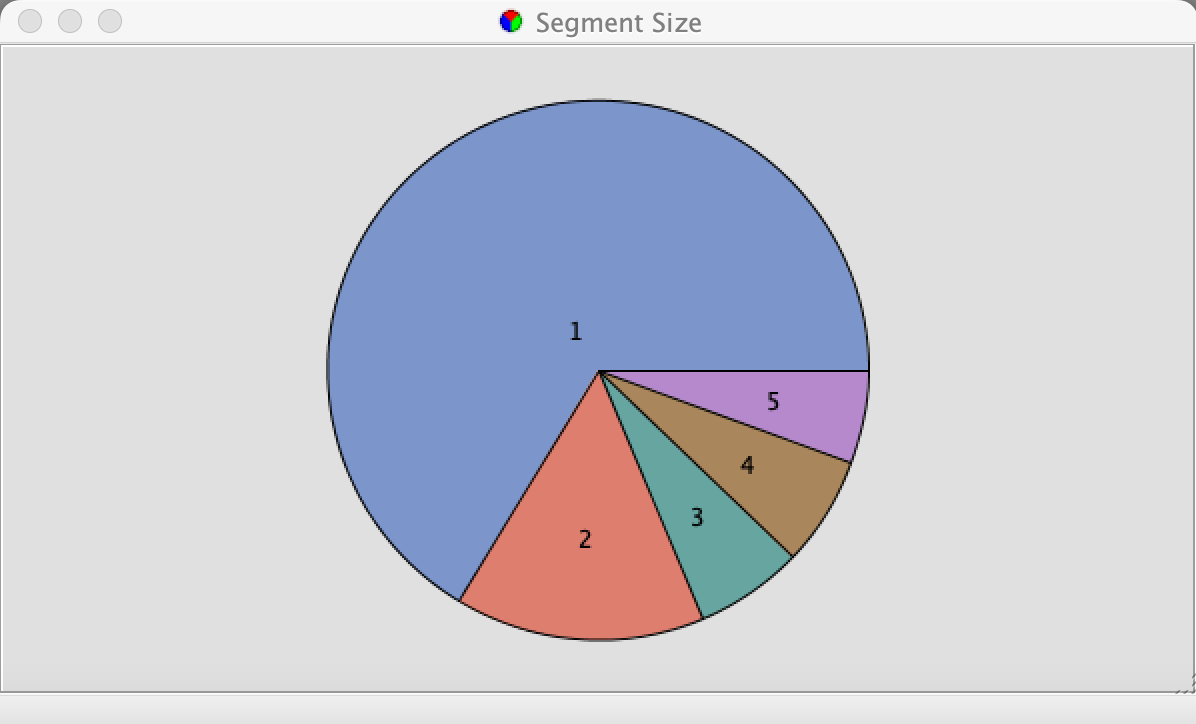
****

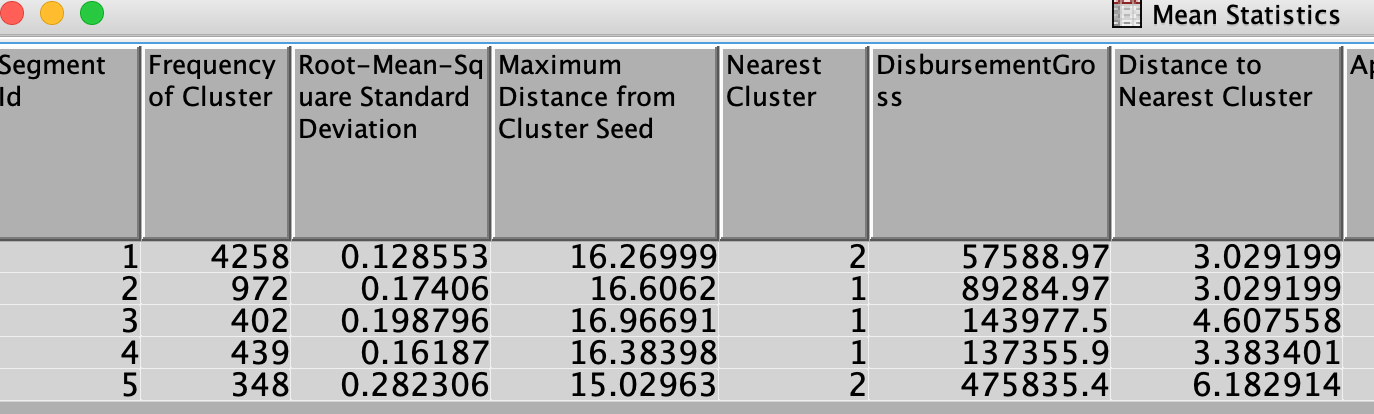
****

****

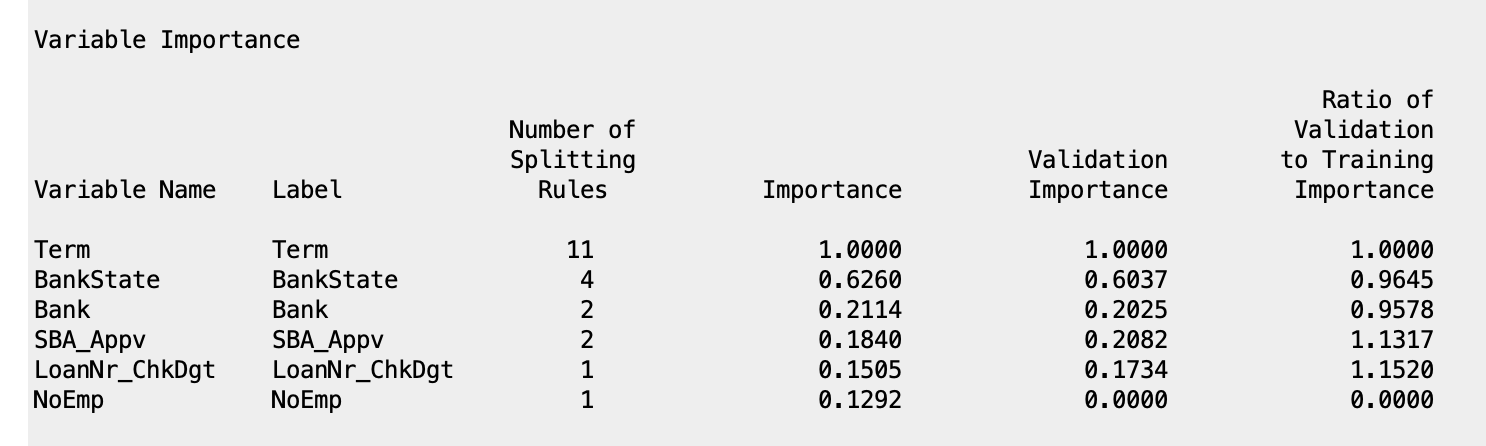
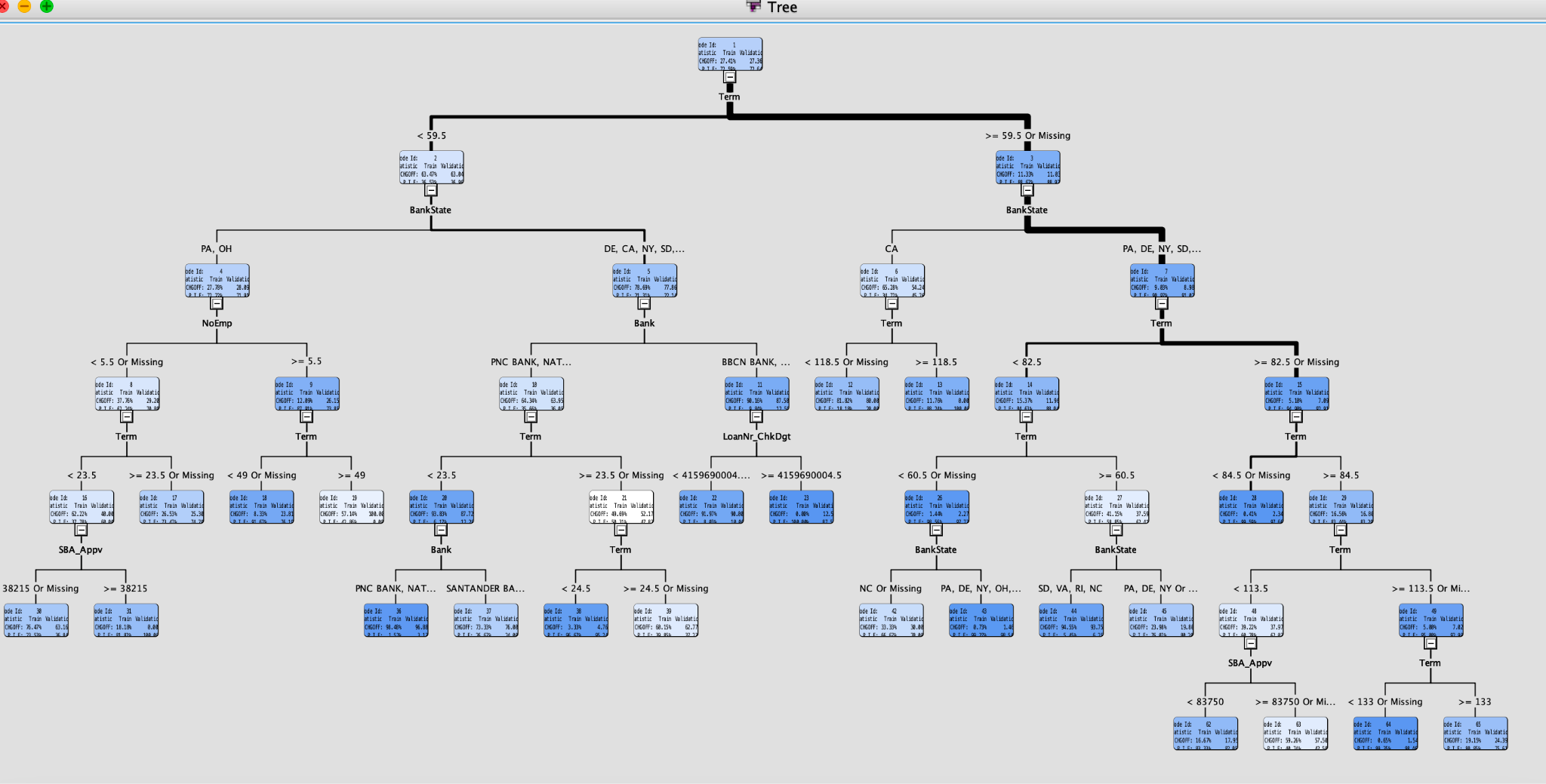
****

**Cluster Analysis 1(Ward)**

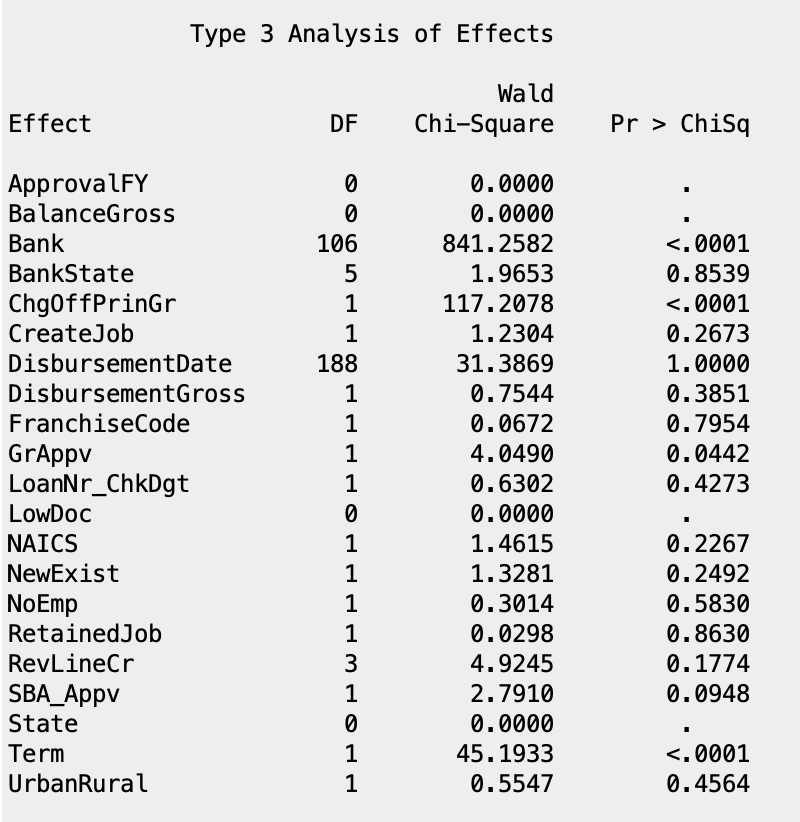
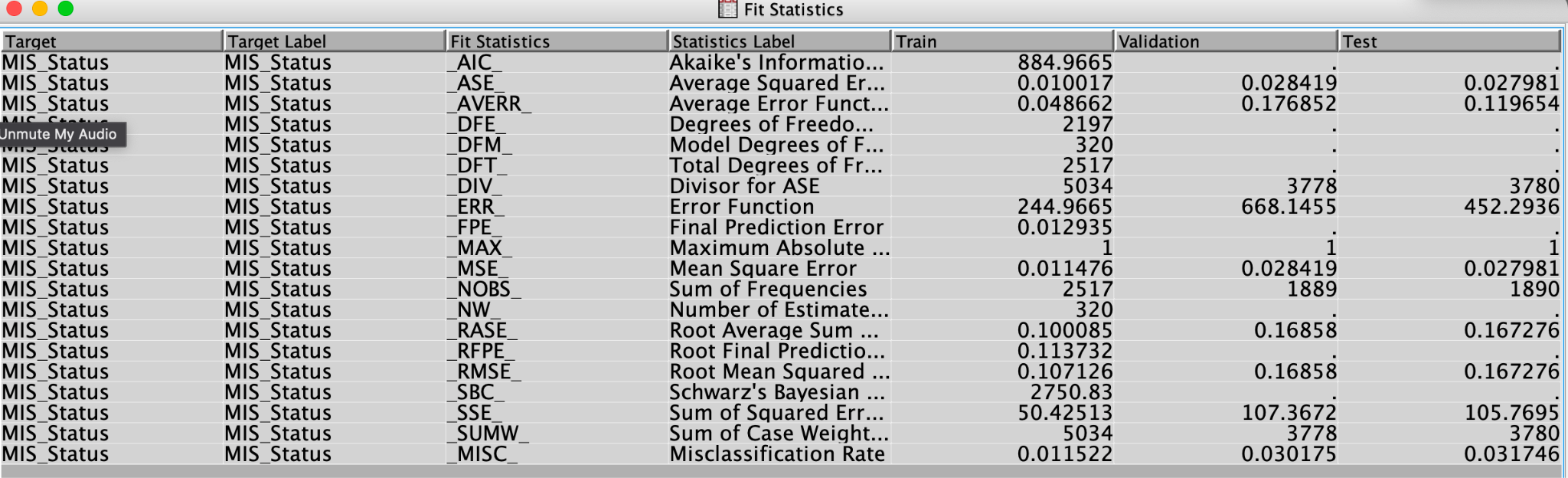
**Cluster Analysis 2(Ward)**



**Decision Tree**

****

**Logistic Regression**

****

**References**

A., & Li, M. (n.d.). "should this loan be approved or denied?": A large dataset with

class assignment guidelines. Taylor & Francis. Retrieved May 4, 2022

*Can You Get Charge-Offs Erased From Your Credit Reports?* (2021, June 4). Investopedia.

Getting Started with SAS(R) Enterprise Miner(TM) 7.1. (2011, November 3). SAS Enterprise

Miner Forum. https://support.sas.com/documentation/cdl/en/emgsj/64144/HTML/default/viewer.htm#n1docbb4hkr3nwn1ukqzqmme48yn.htm

Min Li, Amy Mickel & Stanley Taylor (2018) “Should This Loan be Approved or Denied?”: A Large Dataset with Class Assignment Guidelines, Journal of Statistics Education, 26:1, 55-66, DOI: 10.1080/10691898.2018.1434342

**Appendix**

* *Executive Summary (pg.2)*
* *Data Description(pg.3)*
* *Data Preparation Activities(pg.3)*
* *Project Motivation/Background(pg.2)*

Pg.4-5 : Neural Networks

Pg.5-6: Cluster Analysis

Pg.7 : Decision Tree

Pg.7-8: Logistic Regression

* *Managerial Implications/Conclusion(pg.8)*
* *Models/Diagrams(pg.9)*

Pg.9: Target Variable diagram

Pg.10-12: Neural Networks

Pg.13-14: Cluster Analysis

Pg.15 : Decision Tree

Pg.16-17: Logistic Regression

* *References(pg.18)*