# **MAT 303 Module Six Problem Set Report**

Decision Trees

Justin Farquhar

Justin.farquhar@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

We are exploring two separate data sets within the problem set. One is again acting as a risk analyst working for a credit card company and exploring the different variables that could impact someone defaulting on their credit. For the second data set we are an analyst working for the government and are studying how different variables effect the wage growth of an individual. Each of these results could be used specifically in their own ways, the first mostly by the company to determine which users are at risk of defaulting, who does not qualify for credit, or who would qualify for credit increases. We will be utilizing classification decision trees and regression decision tress for these data sets respectfully.

## 2. Data Preparation

Within the first data set the most important variables in my opinion would be default as that is what we are studying, credit\_utilize as that shows how much of their available credit an individual is using, missed\_payment as having missed a payment can have drastic negative effects on someone’s credit, assets which shows if someone owns a car, house, neither, or both, and education as generally having a higher education is going to lead to a higher paying job, which can help with proper credit payments. Within the second data set the most important variables in my opinon would be wage\_growth as that is what we will be studying, inflation as it can have a high correlation with the value of the dollar, unemployment as the less people that have jobs the less likely there are to be increases in wage, economy as if there is a recession wage increases are also less likely, and gdp as if gdp is high, it is likely that companies are doing well and can better afford high wage increases. There are 600 rows and 8 columns within the credit default data set and 99 rows and 6 columns within the economic data set.

## 3. Classification Decision Tree

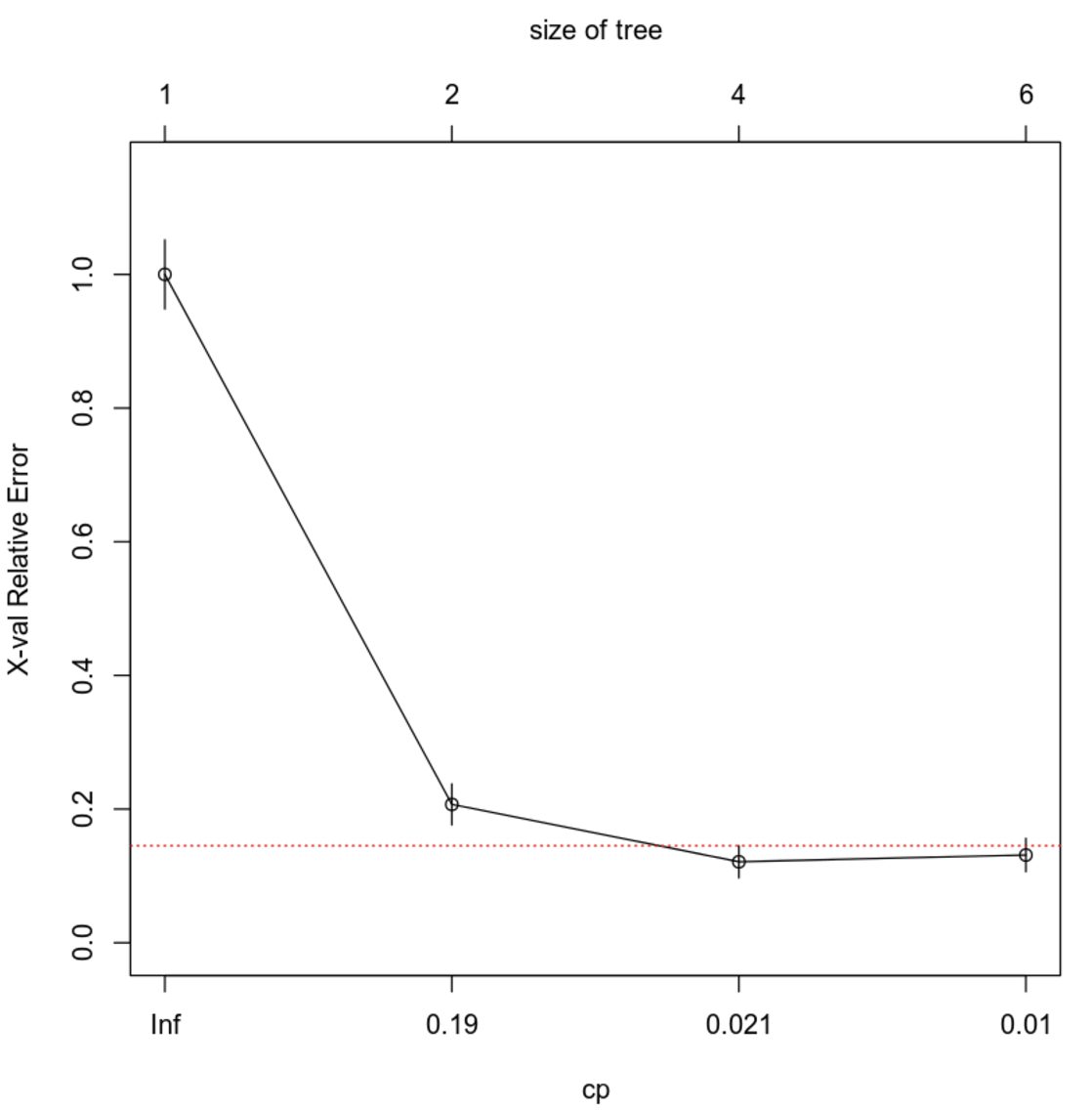
### Reporting Results

Within the original data set there were 600 rows, the training set contains 420 rows, and the validation set contains 180 rows.

A black text with black letters

Description automatically generated with medium confidence

Above is the cost-complexity table for this set seed for the default variable using missed payment, credit utilization and assets as predictors.



The above image is the validation error plotted against the cost complexity parameter. In a general sense, the lower the cp value the better, but the lower the value, the easier it is for overfitting to occur, which could capture noise of patterns rather than the actual trends we are looking for. Also in a general sense, using a value giving the lesser xerror plus roughly one standard deviation to account for avoiding overfitting is the best choice, which in this case lead me to the 0.021 cp value.

A diagram of a credit

Description automatically generated

This is the resulting pruned decision tree for our data set. I had originally chosen to go with a cp value of 0.045455 but this resulted in only the top portion of this decision tree which was only based on the credit utilization. While it is important, I feel as if there are more circumstances that go into if someone is going to default and these additional variables we included should also be present.

### Evaluating Utility of Model

I then proceeded to create a confusion matrix for the decision tree. This resulted in a true negative of 74, a true positive of 100, a false positive of 4, and a false negative of 2. These values then allow us to calculate the accuracy, precision, and the recall, which came to an accuracy value of 0.9666667, precision value of 0.9615385, and a recall value of 0.9803922.

### Making Predictions Using Model

We can then move onto making predictions with our classification model. For an individual who has not missed payments, owns a car and a house and has a 30% utilization the model resulted in them not defaulting. Then for an individual who has missed payments, does not have any assets, and has a 30% credit utilization, the model resulted in them defaulting.

## 4. Regression Decision Tree

### Reporting Results

There were 99 rows in the original set, there are 79 rows in the training set, and 20 rows in the validation set.

A black and white text with numbers

Description automatically generated

The image above represents the cost complexity table for our economy data set for wage growht using economy, unemployment, and gdp as predictors.

A graph with a line

Description automatically generated

The image above represents the plot of the validation error against the cost complexity parameter. Looking through the cp table and the plot we can better access what an appropriate cp value would be for pruning the tree. In this case, I selected 0.015015 as that is right around where the red line is crossing the plots as well as giving one of the lowest xerror values.

A diagram of a graph

Description automatically generated

The image above is the resulting pruned decision tree. We can clearly see that there is ample branches, allowing for the variables to show how each effects wage growth as each are very important in determining change.

### Evaluating Utility of Model

For this regression decision tree we got a root mean squared error value of 1.0268. This represents the average difference between the predicted and true values. This means that our decision tree has a fairly good fit to be only off by roughly 1.0268.

### Making Predictions Using Model

We can now make predictions with our regression model. Predicted wage growth if the economy is not in recession, unemployment is at 3.4%, and the GDP growth rate is 3.5% comes out to 7.0814%. Then predicted wage growth if the economy is in recession, unemployment is at 7.4%, and the GDP growth rate is at 1.4% comes out to be 4.4025.

## 5. Conclusion

These results show that our models can predict these outcomes to very high reliability. There is bound to be some amount of error, especially in the credit situation as there are many situations that can arise, but our models help account for these with the decision trees. For the wage growth data set, the room mean squared error came out to just over 1, indicating that the model is averaging to be off by as little as 1, which is incredible and extremely viable for future predictions and viewing scenarios to see how they might play out. The first data set allows the company to determine who qualifies for credit, as well as who qualifies for credit limit increases. This data can also be utilized to help provide resources or warnings to users who are starting to get dangerously close to enough predictors to have a higher potential to default, either providing resources or other help to try and prevent them from doing so. The second data set lets us see how each predictor effects overall wage growth, allowing us to plug in situations and see how it is likely to play out if these things were to happen. This can allow for us to better prepare for situations where there is lesser growth.