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

Decision Trees

Ashley De Venuto

ashley.devenuto@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this analysis, the data sets being used are the credit card default data set and the economic data set. When it comes to the credit card default data set, creditors working with the credit card company can use the results of the analysis to help determine if the client/customer is a good fit for a loan or line of credit or if they will default on payments and not be granted a loan or line of credit. For this part of the analysis, a classification decision tree will be used. For the analysis of the economic data set, the results of the analysis will show if certain factors aid in improving wage growth. The type of analyses used here will be a regression decision tree.

## 2. Data Preparation

The important variables that are being used in the first part of the analysis using the credit card default data set are default as the response variable and missed payments, credit utilization, and assets as the prediction variables. The important variables used in the second part of the analysis using the economic data set are wage growth as response variable and economy, unemployment, and gdp as the prediction variables. There are 8 columns that represent the variables and 600 rows representing the records within the credit card default data set. There are 6 columns the represent the variables and 99 rows the represent the records within the economic data set.

## 3. Classification Decision Tree

### Reporting Results

When using set.seed(6751342) and split the credit card default data set using a 70%(training) and 30%(validation) split, there are 600 rows in the original data set, 420 rows in the training set, and 180 in the validation set. After creating the model for the classification decision tree, a table was made that is cp against x-val relative error. The cp value that is appropriate to use is 0.021. This number is used during the tree pruning because it is the leftmost value where the mean sits below the red line. This indicates that it is one standard error above the minimal error.

Here is the visualization of the CP for the data set:

A graph with a line

Description automatically generated

When using set.seed(6751342) and use the appropriate cp value, the resulting decision tree looks like this:

A diagram of a credit utility

Description automatically generated

### Evaluating Utility of Model

The confusion matrix looks like this:

A screenshot of a computer

Description automatically generated

The true positives are 100 where the actual and predicted values are 1. The true negatives are 74 where both the actual and predicted values are 0. The false positives are 4 where the actual value is 0 and the predicted value is 1. The false negatives is 2 where the actual value is 1 and the predicted value is 0.

Accuracy =

Accuracy =

Accuracy = 97%

Precision =

Precision =

Precision = 96%

Recall =

Recall =

Recall = 98%

### Making Predictions Using Model

The prediction for defaulting on credit for an individual who has not missed payments, owns a car and a house, and has a 30% credit utilization is no or low chances of defaulting. This means that this individual has a low risk of this person not paying back their loans. The prediction for defaulting on credit for an individual who has missed payments, does not have any assets, and has a 30% credit utilization is yes or a high chance of defaulting on payments.

## 4. Regression Decision Tree

### Reporting Results

Using set.seed(6751342) and split the economic data set 80%(training) and 20%(validation) split, the original data set has 99 rows, the training set has 79 rows, and the validation set has 20 rows. After creating the model for the classification decision tree, a table was made that is cp against x-val relative error. The cp value that is appropriate to use is 0.014. This number is used during the tree pruning because it is the leftmost value where the mean sits below the red line. This indicates that it is one standard error above the minimal error.

Here is the visualization of the CP for the data set:

A graph with a line

Description automatically generated

When using set.seed(6751342) and use the appropriate cp value, the resulting decision tree looks like this:

A diagram of a graph

Description automatically generated

### Evaluating Utility of Model

The root mean squared error for the regression decision tree is 1.0268. This number is also considered the residuals standard deviation.

### Making Predictions Using Model

The predicted wage growth if the economy is not in a recession, unemployment is at 3.4% and the GDP growth rate is 3.5% is 7.0814. The predicted wage growth if the economy is in a recession, unemployment is at 7.4% and the GDP growth rate is 1.4% is 4.4025.

## 5. Conclusion

It seems that both the classification and regression decision trees are useful visualizations to use when determining the risks of an individual defaulting on payments or what variables help improve wage growth. Each branch of the decision trees compares two possibilities and as an individual moves down the branches, two variables are compared. This process is repeated until all variables involved have been analyzed. This specific type of model can accurately determine the risks associated with each outcome. In the real world, the results of the found using the economic data set can tell a business what future wage growth may look like. When it comes to using this model in the real world using the credit card default data set, lending institutes can use this model to access the risks associated with giving an applicant a new loan or line of credit.