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

Decision Trees

Kayla Sacks

Kayla.sacks@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this analysis, I will be using two different data sets to create two decision tree models. The first model will be a classification decision data tree using a data set that explores the relationship between how likely an individual is to default on their credit and their characteristics. The results of this part of the analysis can be used to help credit companies learn which of their customers are more likely to default on their credit, allowing them to better manage their risk.

The second model will be a regression decision tree using data that explores wage growth patterns based on different economic factors. The results of this part of the analysis can help economists predict wage growth based on different economic situations.

I will be splitting each data set into training and testing sets to create the decision trees. I will plot the cross validation error against cost-complexity parameter (cp) to find the best cp to use to prune the decision tree to create the most appropriate model for the analysis. I will then create the decision tree and evaluate each model, using a confusion matrix for the classification tree and finding the root mean squared error for the regression tree.

## 2. Data Preparation

The first data set I will be using has a total of 8 columns and 600 rows. The rows will be divided into the training set which will be used to create our decision tree and the validation set that will be used to test the model. The variables that will be used are:

* default – If the individual defaulted on their credit. This is the response variable. 0 = did not default, 1 = default
* missed\_payment – If the individual has missed a payment in the last 3 months. Input values are ‘no’ and ‘yes’.
* credit\_utilize – How much credit has the individual utilized out of how much is allowed. Expressed as a decimal representing percentage.
* Assets – The assets owned by the individual. Input values are ‘none’, ‘car’, ‘house’, and ‘car\_house’.

The second data set I will be using has a total of 6 columns and 99 rows. The rows will again be divided into the training set which will be used to create our decision tree and the validation set that will be used to test the model. The variables that will be used are:

* wage\_growth – The wage growth rate, this will be the response variable.
* economy – If the economy is in recession or nor. Input values are ‘recession’ and ‘no\_recession’
* unemployment – The unemployment rate
* gdp – The GDP growth rate

## 3. Classification Decision Tree

### Reporting Results

For the first model, there are 600 rows, 70% (420 rows) of which will be used as the training set with the remaining 30% (180 rows) will be used as the validation set.

Once the data was split into the training and validation sets, the decision tree for default using missed payments, credit utilization, and assets was created along with the cost-complexity(cp) table which can be seen below.

A screenshot of a computer error

Description automatically generated

The plot for validation error against the cost-complexity parameter was then created to find the best cp value for pruning the tree.

A graph of a tree

Description automatically generated

From the plot we can see the best cp for our pruned model is 0.021. Using this we created the following model and decision tree:

A computer code with numbers and symbols

Description automatically generated

A diagram of a credit utility

Description automatically generated

### Evaluating Utility of Model

To evaluate the utility of this classification decision tree, a confusion matrix was created where we can see the number of true positives, true negatives, false positives, and false negatives*.*

A white rectangular box with black text

Description automatically generated

From the table we can see the count for:

* True Positives – Actual and predicted are both “yes”: 100
* True Negatives – Actual and predicted are both “no”: 74
* False Positives – Actual is “no” but predicted is “yes”: 4
* False Negatives – Actual is “yes” but predicted is “no”: 2

Accuracy was calculated to find the ratio of correct predictions to the total number of observations.

Precision was calculated to find the ratio of correct positive predictions to the total predicted positives.

Lastly, sensitivity (recall) was calculated to find the ratio of correct positive predictions to the total amount of positives.

### Making Predictions Using Model

**Prediction 1**

We will use this decision tree model to predict if an individual who has not missed payments, owns a car and a house, and has a 30% credit utilization will default on their credit.

With credit utilization is at 30% (0.30) we the answer to the first branch is “no” with the second branch being “yes”. Since the individual owns a car and house, the answer for the third branch will be “yes” so the prediction is that the individual will not default on credit.

**Prediction 2**

We will use this decision tree model to predict if an individual who has missed payments, does not have any assets, and has a 30% credit utilization will default on their credit.

With credit utilization is at 30% (0.30) we the answer to the first branch is “no” with the second branch being “yes”. Since the individual does not own a house, the answer for the third branch will be “no” so the prediction is that the individual will default on credit.

## 4. Regression Decision Tree

### Reporting Results

For this model, there are a total of 99 rows, 80% (79) of which will be in the training set with the remaining 20% (20) being in the validation set.

After splitting the data into the training and validation sets, the regression decision tree for wage growth using economy, unemployment, and gdp as predictors was created. The cost-complexity table can be seen below.

A screenshot of a computer code

Description automatically generated

The plot for the validation error against the cost-complexity parameter was created to find the best cp value to use for pruning the regression tree.

A graph with a line

Description automatically generated

From the plot, we can see that a cp of 0.014 would be an appropriate cp value to use in pruning this regression tree. The pruned model and resulting decision tree can be seen below.

A computer screen shot of a number

Description automatically generated

A diagram of a graph

Description automatically generated

### Evaluating Utility of Model

To evaluate the utility of this regression decision tree, we will find the root mean squared error for the tree. The equation for this is:

The RMSE for this decision tree is 1.0268 and is the standard deviation of the residuals.

### Making Predictions Using Model

**Prediction 1**

We will now predict wage growth if the economy is not in recession, unemployment is at 3.4%, and the GDP growth rate is 3.5% using our regression decision tree.

With unemployment at 3.4%, we will answer “no” at the first branch as it is less than the 4.2. The answer for the second branch will be “yes” as the GDP growth rate is less than 8.7 so our prediction for wage growth using the decision tree is 7.1%. This matches the R code result of 7.0814%.

**Prediction 2**

For this prediction, we will predict wage growth if the economy is in recession, unemployment is at 7.4%, and the GDP growth rate is 1.4%.

With unemployment at 7.4%, we will answer “yes” at the first branch and “no” at the second as it is greater than the 4.2 but less than the 7.6. Our prediction for wage growth using this decision tree will be 4.4% which matches the R code result of 4.4025%.

## 5. Conclusion

Based on provided test data, decision trees offer a useful way to visualize the risk between two possible outcomes. Each branch of the tree compares two outcomes for a scenario and leads to the next branch where it will again compare two variables. This process will continue though the analyses until a conclusion with the most likely outcome is found. The results of this type of model can provide predictions with high accuracy of the risks that may derive with different outcomes. The results of these analyses have practical importance as they can help companies make better decisions, which helps them to avoid risks. The first model can help credit companies to decide if they should provide or extend a credit line for an individual by determining if they are likely to default based on their credit utilization and their assets. This can help the company reduce the chance of offering a credit line to high-risk individuals.