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

Decision Trees

Yaqub Mohamud

yaqub.mohamud@snhu.edu

Southern New Hampshire University

## 1. Introduction

This problem set focuses on predicting loan default and pay growth by analyzing two data sources. The collecting of credit card default data includes individual indicators including assets, credit utilization, and missed payments, which are used to predict if a customer would default on their credit. The second type of data contains economic indicators that are used to predict wage growth, such as GDP and unemployment. As part of the analysis, regression decision trees for wage growth and categorization decision trees for credit default prediction were created.

## 2. Data Preparation

The predictor variables for the analysis of the credit default data set include credit utilization, assets, and the occurrence of a missed payment during the previous three months. The person's credit default status is the outcome variable. In the instance of the economic data set, the predictor variables are GDP, the unemployment rate, and the status of the economy (whether or not it is in a recession). There are 600 rows and 8 columns in the credit default data set and 99 rows and 6 columns in the economic data set.

## 3. Classification Decision Tree

### Reporting Results

600 rows make up the original data set, 420 rows make up the training set, and 180 rows make up the validation set when the credit card default data set is divided into training and validation sets using 70% and 30% split, respectively, using set.seed(6751342).

The cost-complexity (cp) table:

A graph with a line

AI-generated content may be incorrect.

Since 0.021 is the leftmost number for which the mean is below the red line, signifying one standard error over the minimal error, it is a suitable cp value to use when pruning the tree.

Using set.seed(6751342) and pruning the tree using the appropriate cp value gives a resulting decision tree as seen below:

A diagram of a credit utility

AI-generated content may be incorrect.

### Evaluating Utility of Model

The confusion matrix results are:

* 100 true positives
* 74 true negatives
* 4 false positives
* 2 false negatives

Accuracy = 0.97 or 97%

Precision = 0.96 or 96%

Recall = 0.98 or 98%

### Making Predictions Using Model

An individual with a 30% credit utilization, a car and a house, and no past-due payments is predicted to not default on credit. This indicates that there is no likelihood that this person will have a credit default.

If a person has missed payments, has no assets, and has a 30% credit utilization, the likelihood that they will default on their credit is yes. This indicates that a credit default is anticipated for this person.

## 4. Regression Decision Tree

### Reporting Results

The economic data set may be divided into training and validation sets using set.seed(6751342) in an 80% and 20% split, respectively. The training set consists of 79 rows, the validation set of 20 rows, and the original data set consists of 99 rows.

The cost-complexity (cp) table:

A graph of a tree

AI-generated content may be incorrect.

Since 0.014 is the leftmost number for which the mean is below the red line, signifying one standard error over the minimal error, it is a suitable cp value to use when pruning the tree.

Using set.seed(6751342) and pruning the tree using the appropriate cp value gives a resulting decision tree as seen below:

A diagram of a graph

AI-generated content may be incorrect.

### Evaluating Utility of Model

The regression decision tree's Root Mean Squared Error (RSME) is 1.0268. The standard deviation of prediction errors or residuals is known as the RSME, and a low value of 1.0268 denotes a model with excellent accuracy and minimal variance.

### Making Predictions Using Model

If the GDP growth rate is 3.5%, unemployment is 3.4%, and the economy is not in a recession, the expected wage growth is 7.0814. If the economy is in a recession, unemployment is 7.4%, and the GDP growth rate is 1.4%, the expected wage growth is 4.4025.

## 5. Conclusion

It seems that decision trees are a great way to visualize the results of risk assessments between two options. According to the results and other parameters, each branch can compare two possibilities until all the data has been analyzed. After that, it can go on to a new branch and compare two variables at a time. This can give precise projections of how risk will vary between two scenarios.

The practical significance of the analyses performed is that, in the case of the model for the economic data set, businesses may use them to forecast wage increases under various scenarios. Based on information provided to the loan business about the applicant, credit firms can use the model for the credit default data set to assess the risk of offering a new loan to the applicant.