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

Decision Trees

Camilo Hoyos

camilo.hoyos@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data set we’re exploring is a credit evaluation for customers. It has attributes to determine if a customer is likely to default on their credit or not. The analysis that will be run on this data set is a classification decision tree.

There is another data set we will explore which is an economic data set. It will be used to predict wage growth dependent on a handful of attributes related to the economy. The analysis performed on this data set will be a regression decision tree.

## 2. Data Preparation

The important variables in the credit evaluation data set are the missed payments variable, credit utilization, and the assets of the customer. Missed payments identify if they have missed a payment in the past three months and assets determine if the customer has a car, house, both, or none of the assets. There are a total of 8 columns and 600 rows.

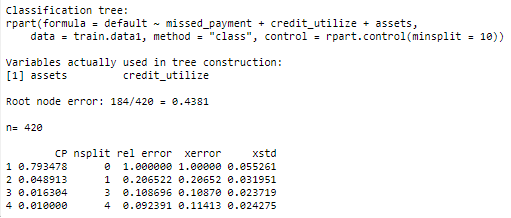
The important variables in the economic data set are economy, employment, and GDP. Economy identifies if the economy is in a recession or not, and the others are quantitative variables. There are a total of 6 columns and 99 rows.

## 3. Classification Decision Tree

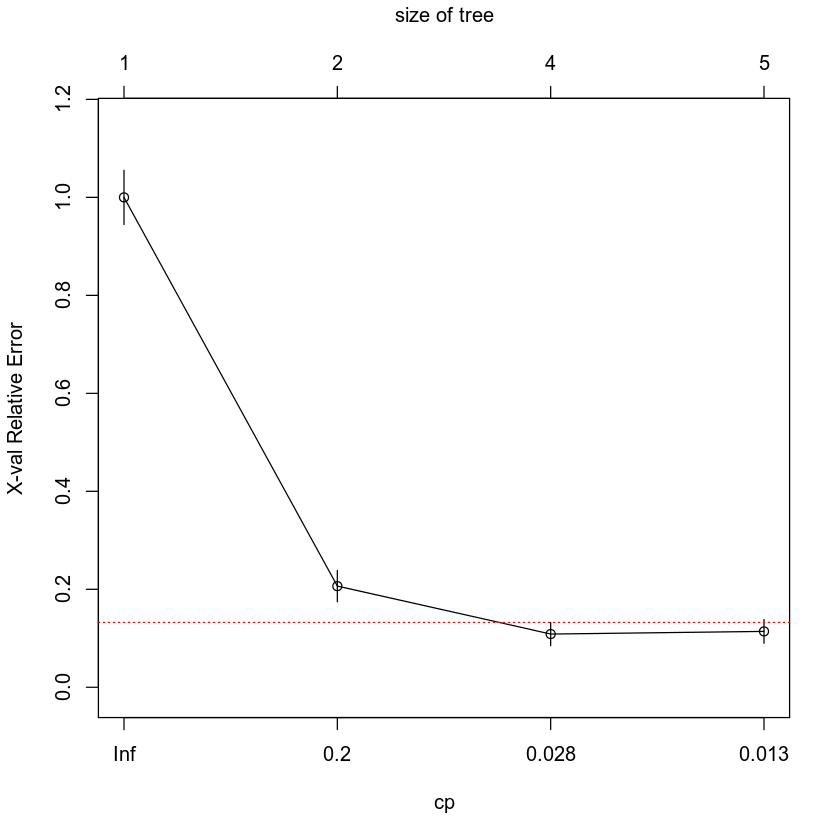
### Reporting Results

In the original data set there are 600 rows, the training set has 70% of the rows or 420, and the validation data set has the remaining 30% of the rows or 180.

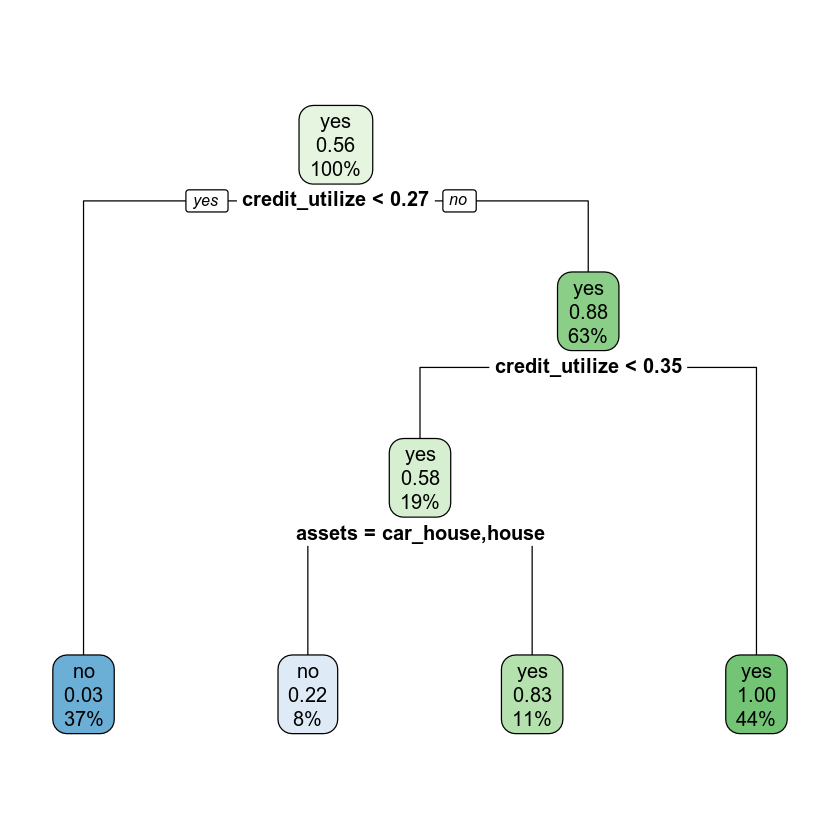
Cost-complexity parameter



CP Validation plot

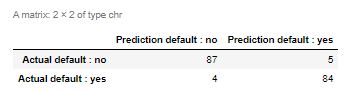


The CP value that will be used to prune the tree will be 0.028. After pruning at 0.028, the resulting decision tree will look as such:



### Evaluating Utility of Model

Confusion matrix below shows that true positive count is 84, true negative count is 87, false positive count is 5, and false negative count is 4.



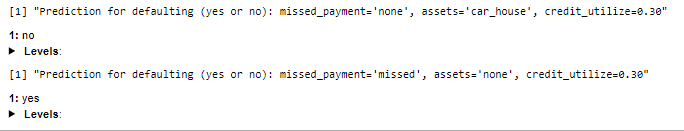
Accuracy: = = 0.95 or 95%. This provides the number of correct predictions to the total number of observations.

Precision: = = 0.9438 or 94.38%. This provides the correct positive predictions to the total predicted positives.

Recall: = = 0.9545 or 95.45%. This provides the correct positive predictions to the total positives examples.

### Making Predictions Using Model

The prediction for defaulting on credit with an individual who has not missed payments, owns a car and a house, and has 30% credit utilization is “no”; they will not default on their credit. The prediction for defaulting on credit with an individual who has missed a payment, does not have any assets, and has 30% credit utilization is “yes”; they will default on their credit.

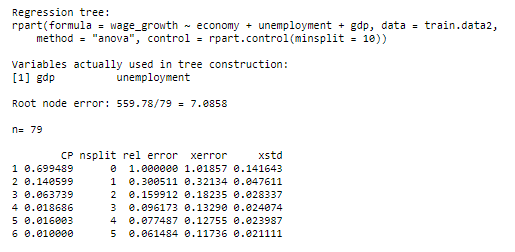


## 4. Regression Decision Tree

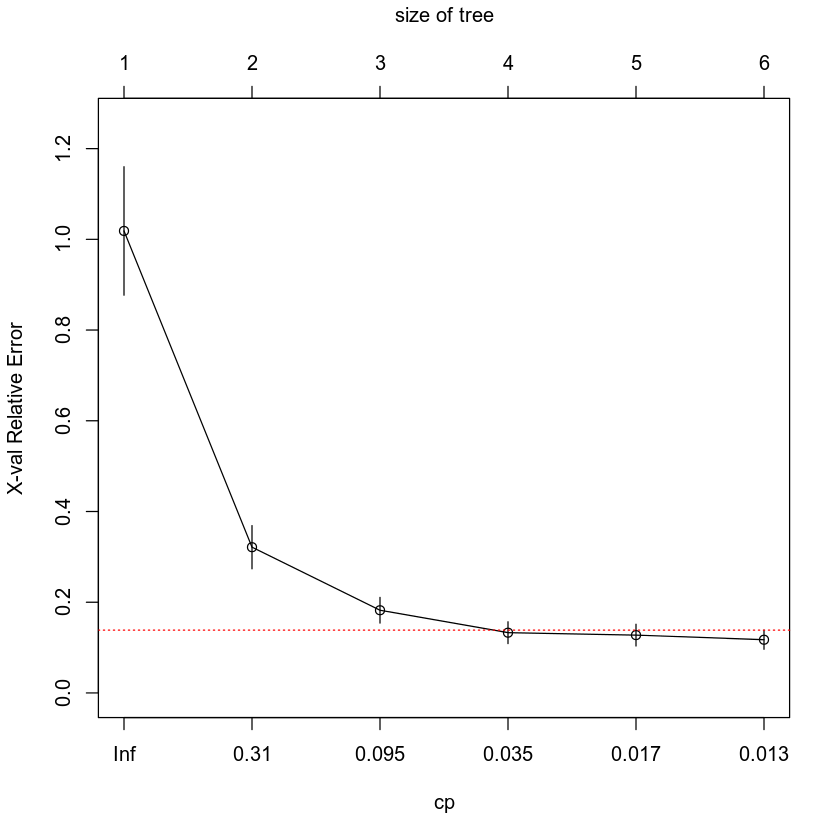
### Reporting Results

The original dataset had 99 rows. The dataset was then split to have 79 records in the training dataset and 20 in the validation data set.

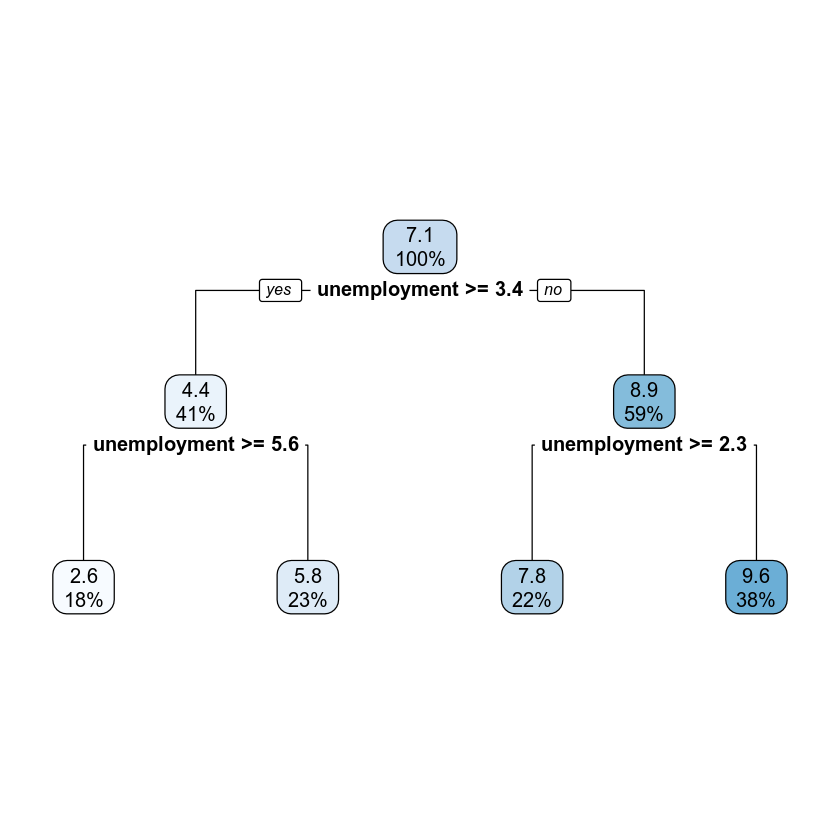
Cost-complexity parameters:



CP validation plot:



An appropriate cp value to cut off the tree is 0.035. The resulting decision regression tree will look like the following:



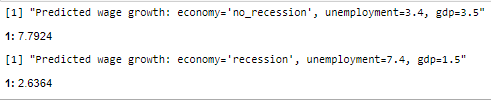
### Evaluating Utility of Model

The root mean squared error for the regression tree is 0.8386. This is the standard deviation of residuals or how much deviation there is from the regression line. The lower this value, the more accurate it is, it is simply a tool to evaluate the fit of the model. (Holmes, 2000)

### Making Predictions Using Model

The predicted wage growth if the economy is not in recession, unemployment is at 3.4% and GDP growth is 3.5% is 7.7924%.

The predicted wage growth if the economy is in recession, unemployment is at 7.4%, and the GDP growth is 1.5% is 2.6364%.

**

## 5. Conclusion

In the two analyses, classification and regression tree, we utilized quantitative and qualitative predictors to predict an outcome. The credit default project determined if a customer would default on their credit as “yes” or “no” which is a classification, not a value. The economic project determined what the wage growth would be as a value which was divided into several buckets by the predictor values. In both projects, predictor values were fed after using a training and validation data set to logically move through the classification tree until the appropriate result was identified. If the predictor values pass the criteria, they move to the left side of the node, if they fail the criteria, they move to the right side of the node until they reach the ‘bottom’ which has a result.

The practical importance of this analysis is to perform supervised learning to identify if an outcome can be predicted related to a set of variables. The difference between this and first or second order regression models utilized in the past is CART evaluations “bucket” results instead of providing directly calculated values.

## 6. Citations

Holmes, S. (2000). RMS Error. Retrieved December 03, 2020, from https://statweb.stanford.edu/~susan/courses/s60/split/node60.html