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

Decision Trees

Zachary Tankersley

[Zachary.tankersley@snhu.edu](mailto:Zachary.tankersley@snhu.edu)

Southern New Hampshire University

## 1. Introduction

The data to be analyzed is comprised of two historical datasets, one containing credit card default information, and the other containing economic data. Two models will be constructed and may be used to accurately predict how likely an individual may be to default on their credit line and predict wage growth. Both analyses will be based off of either a decision or regression tree model.

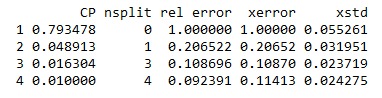
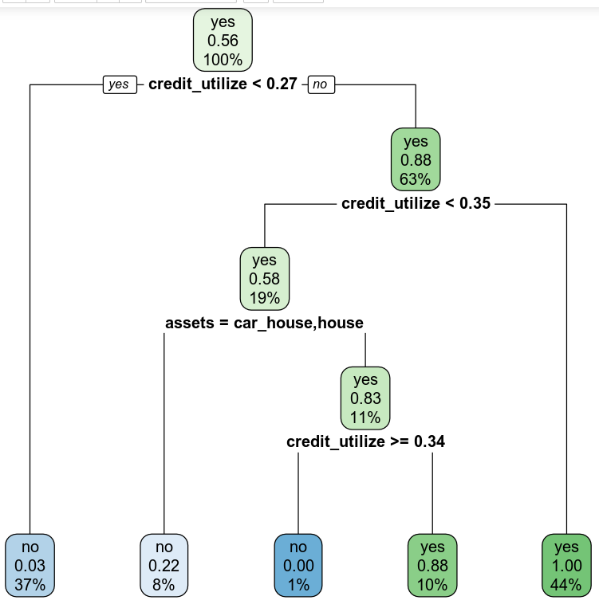
## 2. Data Preparation

In the credit card default dataset, the variables of interest will be default, missed\_payment, credit\_utilize, and assets. There are 600 rows and 8 columns of data in this dataset. In the economic dataset there are 99 rows and 6 columns of data, and the variables of interest will be: wage\_growth, economy, unemployment, and GDP.

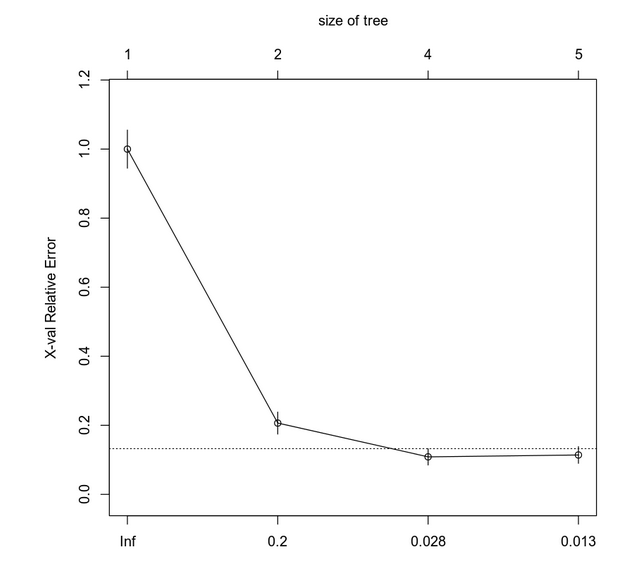
## 3. Classification Decision Tree

### Reporting Results

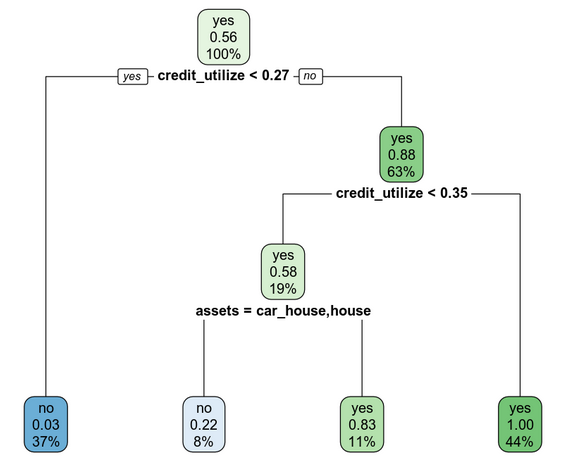
When splitting the credit default data set into training and validation sets at a 70% ratio, the training set ends up containing 420 rows of data, while the validation data contains 180 rows. Using this data to train a classification decision tree model, the unpruned tree and its complexity parameter table are created and appear below:



When plotting the validation error against the cost-complexity parameter the following plot is shown in R:



This plot gives an idea of the appropriate values with which to prune the model, and I chose 0.028 as it is the leftmost value where the mean lies below the line which symbolizes a standard error above the minimum error value. With this value in mind, another tree can be plotted with the pruned model, and it is pictured below:



### Evaluating Utility of Model

The confusion matrix for this model is shown below, which will allow the calculation of the accuracy, precision, and recall:

|  |  |  |
| --- | --- | --- |
|  | Prediction: no | Prediction: yes |
| Actual: no | 87 (tn) | 5 (fp) |
| Actual: yes | 4 (fn) | 84 (tp) |

* Accuracy: 95% or 0.95 (this shows the ratio of true positive and negative predictions to all predictions)
* Precision: 94% or ~0.943 (this shows the ratio of true positive to all positive predictions)
* Recall: 95% or ~0.954 (this shows the ratio of true positive to true positive *and* false negative predictions)

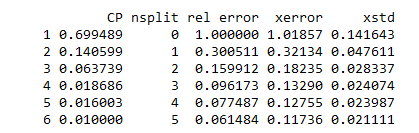
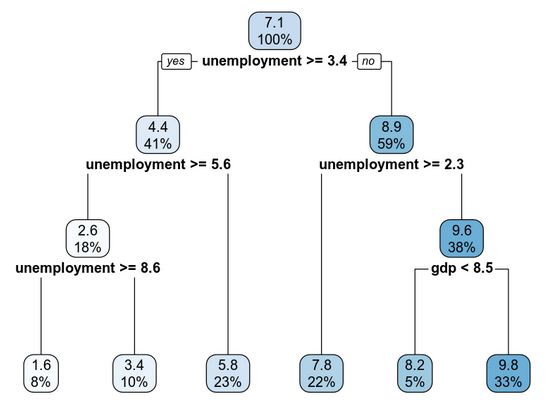
### Making Predictions Using Model

Using the pruned model and an individual who has not missed any payments, owns both a car and a house, and has a 30% credit utilization rate, the model predicts that this individual will *not* default on their credit. For a similar individual with missed payments, no assets, and also with a 30% credit utilization rate, the model predicts that the individual *will* default on credit.

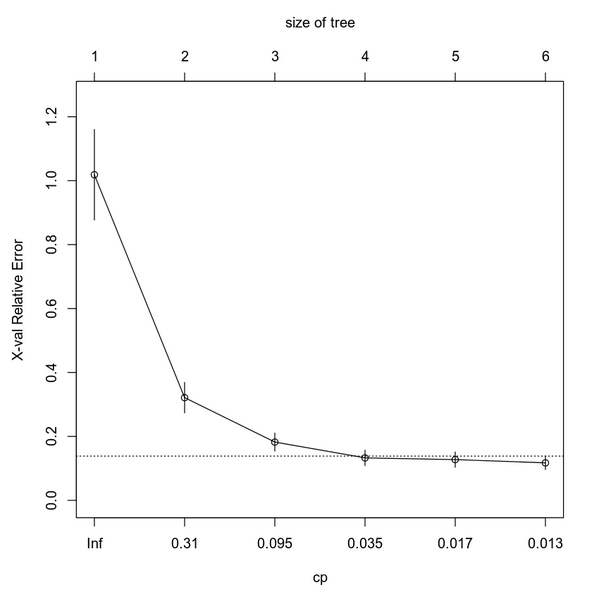
## 4. Regression Decision Tree

### Reporting Results

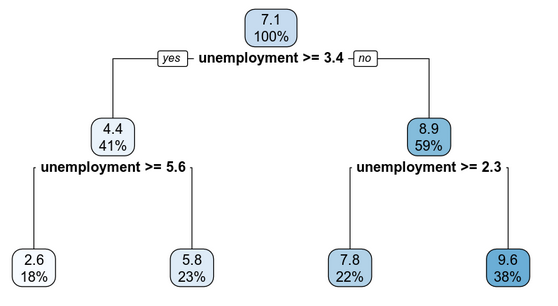
When splitting the economic data set into training and validation sets, an 80% ratio is used instead, and the training set ends up containing 79 rows of data, while the validation set contains 20 rows. Using the data to train a regression decision tree model, the unpruned tree and its complexity parameter table are created and are shown below:



When plotting the validation error against the cost-complexity parameter, the following plot is created in R:



Again, this plot just gives an idea of the appropriate values to prune the model with, and this time the leftmost value under the line is 0.035 which is the cp parameter I chose to prune this model with, and the resulting model tree is shown below:



### Evaluating Utility of Model

The measure the difference between predicted values from this model and the observed values, the RMSE is calculated at around 0.8339, which is fairly close to the wage growth mark we would hope to see, however, a lower value would be better.

### Making Predictions Using Model

Using the model to predict wage growth for an economy that is not in a recession with 3.4% unemployment and a GDP of 3.5%, the wage growth percentage predicted by the model is 7.792%. A much more dire growth percentage of 2.636% is predicted when the economy is in a recession, and unemployment and GDP are at 7.4% and 1.5% respectively.

## 5. Conclusion

The advantage of using a decision or regression tree model is the visual aspect, wherein you can take an observation and fit it to a class or node or branch on a tree and know, generally speaking, where the outcome will fall without having to create a model specifically for that observation. This can be used to establish a baseline with which to judge criteria and can make decisions easier and quicker on average. They are also very powerful at determining outcomes when several different combinations of variables need to be used. These types of models (especially in the case of the credit information data above) can be very useful in determining risk. In a more general sense, decision trees can be used for a wide variety of applications as we have seen here, ranging from financial sector to government predictive applications. Most recently however, machine learning algorithms are being based on decision trees, so the future is bright for this type of model.