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

Decision Trees

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

## 1. Introduction

*The data sets I will be working with are the credit default data set and the economic data set. These results might be used by a credit company looking into whether they should extend a loan to an applicant. Information obtained through these models can help them assess the risk of an applicant defaulting on a loan.*

*I will be creating and analyzing decision trees. These are good at visualizing risk by looking at two scenarios for each branch of the tree.*

## 2. Data Preparation

*The variables relevant to the analysis of the credit default data set are assets, whether a payment has been missed in the last 3 months, and credit utilization as predictor variables. The response variable is whether or not the individual has defaulted on their credit.*

*The variables relevant to the analysis of the economic data set are whether the economy is in a recession or not, unemployment rate, and GDP as predictors. The response variable is wage growth.*

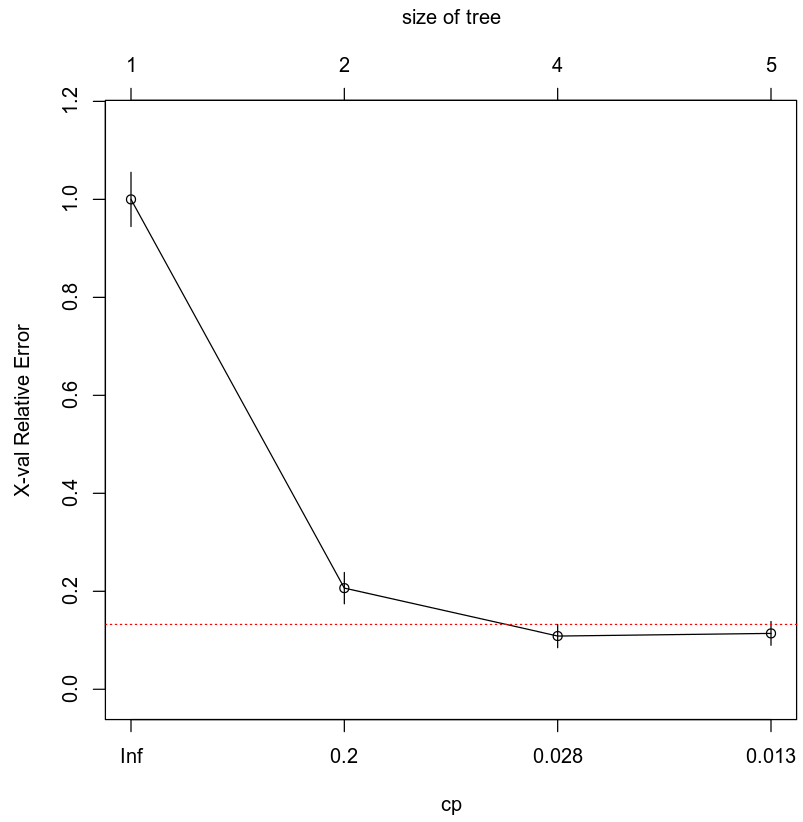
*There are 600 rows and 8 columns in the credit default data set. There are 99 rows and 6 columns in the economic data set.*

## 3. Classification Decision Tree

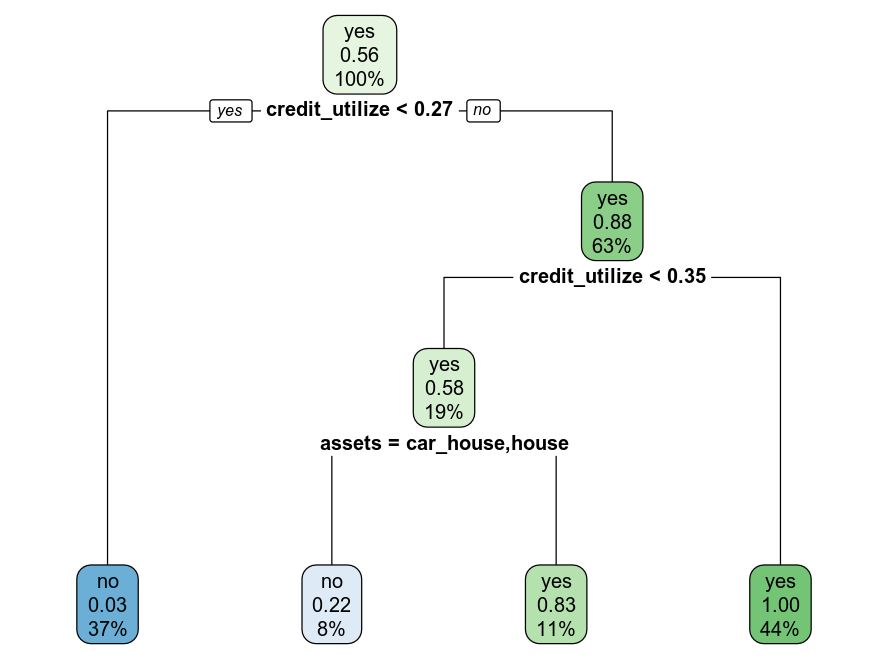
### Reporting Results

* *Using set.seed(705526) and splitting the credit card default data set into training and validation sets using 70% and 30% split, gives 600 rows in the original data set, 420 rows in the training set, and 180 rows in the validation set.*

*Cost-complexity (cp) table for this set:*

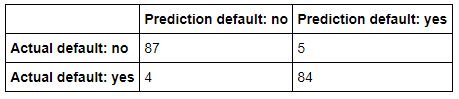
**

* *An appropriate cp value to use in pruning the tree is 0.028 as this is the leftmost value for which the mean lies below the red line denoting one standard error above the minimum error.*
* *Using set.seed(705526) and pruning the tree using the appropriate cp value gives a resulting decision tree as seen below.*

**

### Evaluating Utility of Model

*The confusion matrix for this analysis:*

**

*The confusion matrix results are:*

* *84 true positives*
* *87 true negatives*
* *5 false positives*
* *4 false negatives*
* *True Positive (TP): The actual value is yes (default = yes) and the predicted value is yes (default = yes). Hence a true positive.*
* *True Negative (TN): The actual value is no (default = no) and the predicted value is no (default = no). Hence a true negative.*
* *False Positive (FP): The actual value is no (default = no) and the predicted value is yes (default = yes). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is yes (default = yes) and the predicted value is no (default = no). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

### Making Predictions Using Model

***Prediction 1***

*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. This means this individual is not predicted to default on their credit.*

***Prediction 2***

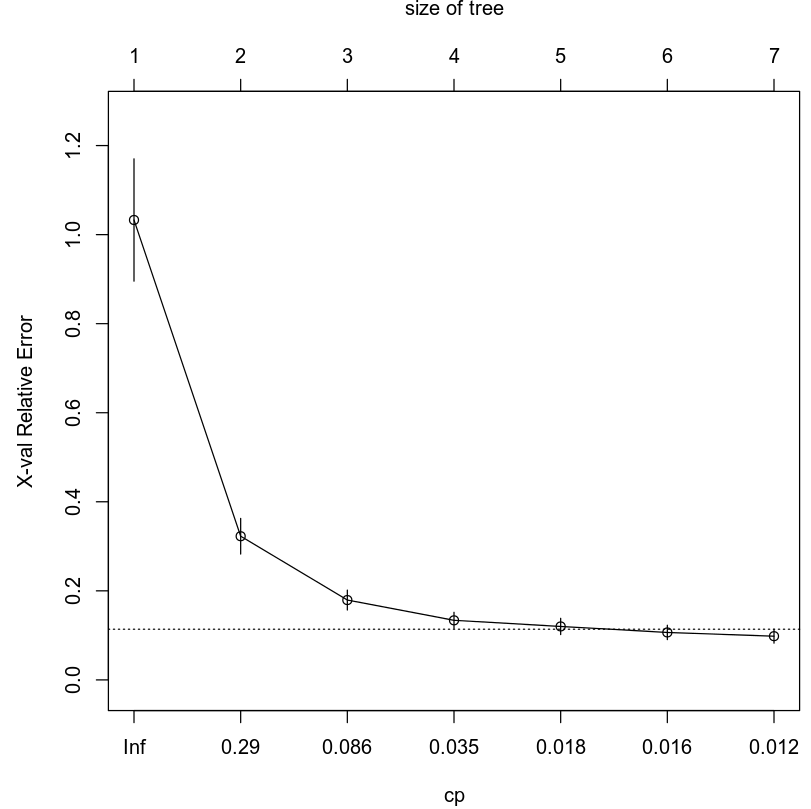
*The prediction for defaulting on credit for an individual who has missed payments, does not have any assets, and has a 30% credit utilization is yes. This means this individual is predicted to default on their credit.*

## 4. Regression Decision Tree

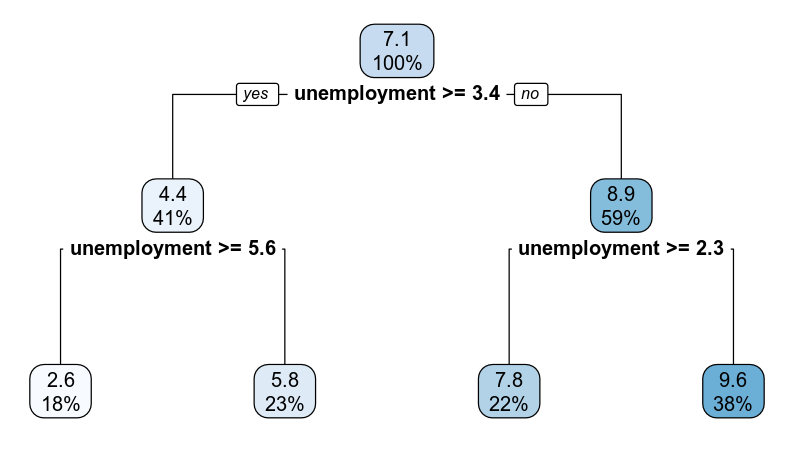
### Reporting Results

* *Using set.seed(705526) and splitting the economic data set into training and validation sets using 80% and 20% split gives 99 rows in the original data set, 79 rows in the training set and 20 rows in the validation set.*

*Cost-complexity (cp) table for this set:*

**

* *An appropriate cp value to use in pruning the tree is 0.018 as this is the leftmost value for which the mean lies below the red line denoting one standard error above the minimum error (it is right on the line and all the ones left of it are clearly above it).*
* *Using set.seed(705526) and pruning the tree using the appropriate cp value and gives the resulting decision tree seen below.*

**

### Evaluating Utility of Model

*The root mean squared error for the regression decision tree is 0.8386. This is the standard deviation of the residuals.*

### Making Predictions Using Model

***Prediction 1***

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

***Prediction 2***

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

## 5. Conclusion

*Decision trees seem to be a great way to visually see risk calculation results between two scenarios. Each branch can compare 2 scenarios and based on the results and other variables, move down to a new branch and further compare two variables at a time until all information has been analyzed. This can give good predictions of the differences in risk between two scenarios.*

*The practical importance of the analyses performed are that companies can use them to predict what will happen to wage growth under different scenarios in the case of the model for the economic data set. Credit companies can use the model for the credit default data set to evaluate the risk of extending a new loan to an individual that has applied, based on information give to the loan company about the individual.*

## 6. Citations

*Zybooks MAT 303: Applied Statistics II for Science, (2016, August).*

*Retrieved April 1, 2020, from https://learn.zybooks.com/zybook/SNHUMAT303v1*