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

Decision Trees

Sarah Steinbaum

sarah.steinbaum@snhu.edu

Southern New Hampshire University

## 1. Introduction

There are two statistical analyses that will be performed. The first uses a set of historical data from a credit card company. The results obtained from the first analyses can be used to find correlations between customer characteristics and the likelihood the customer has of defaulting on their credit. The first analysis will use a classification decision tree which will provide a visualization to conceptualize the data and make predictions accordingly.

The second model utilizes a historical set of data from the government to assess the wage growth of the labor force based on certain factors such as the unemployment and GDP rate. The results from this analysis could be used to make predictions on wage growth based on patterns in the data and the government’s economic agenda. The second model will use a regression decision tree.

## 2. Data Preparation

There are important variables in both sets of data. The set of data that the first model utilizes is the credit card default dataset. The important variables contained in this set include defaulting on credit (*default*), whether or not an individual has missed a payment within the last three months (*missed\_payment*), how much of an individual’s credit is allowed to be used (*credit\_utilized*), and assets owned by the individual (*assets*). This variable could be a car only, house only, car and a house, or none for no assets. The credit card default dataset has 600 rows and 8 columns that each represent a variable.

The second set of data for the second model is the economic dataset. The important variables in this set include the wage growth rate (*wage\_growth*), the unemployment rate (*unemployment*), whether or not the economy is in a recession (*economy*), and the GDP growth rate (*gdp*). The economy dataset contains 99 rows and 6 columns each representing a variable.

## 3. Classification Decision Tree

### Reporting Results

When using set.seed(671342) using a 70% and 30% split, the credit card default data set has 600 rows. The training set has 420 rows and the validation set has 180. The graph below shows the validation error against the cost-complexity parameter (cp) which assists in selecting the most appropriate model by pruning the tree. This graph contains a horizontal red line that is one-standard error above the minimum error. An appropriate cost-complexity parameter (cp) value to prune the tree is 0.021 as it is the largest cp value below the red line.

Chart, line chart

Description automatically generated

Timeline

Description automatically generated

### Evaluating Utility of Model

The values for the confusion matrix are listed below:

* True positive = 100
* True negative = 74
* False positive = 4
* False negative = 2

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.9666667

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.96153846

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.98039216

### Making Predictions Using Model

According to the classification model, an individual who has not missed payments in the last three months, owns a car and house, and has a 30% credit utilization is not likely to default of their credit. However, an individual who has missed payments within the last three months, does not have assets, and has a credit utilization of 30% is likely to default on their credit.

## 4. Regression Decision Tree

### Reporting Results

When using the set.seed(6751342) using a 80% and 20% split the economic data set has 99 rows. The training set has 79 rows and the validation set has 20. The validation error against the cost-complexity parameter (cp) graph is portrayed below. An appropriate cost-complexity parameter (cp) value to prune the tree is 0.014 as this value is the closest to the left side and the largest value below the red line.

Chart, line chart

Description automatically generated

Diagram

Description automatically generated

### Evaluating Utility of Model

The root mean squared error (RMSE) evaluates the performance of the model and is residual’s standard deviation. The closer it is to 0, the more accurate it is. The RMSE for the regression decision tree is 1.0268, which indicates that the performance of this model highly likely to be accurate.

### Making Predictions Using Model

According to the regression model, the predicted wage growth if the economy is not in a recession, unemployment is at 3.4% and the GDP growth rate is 3.5% is 7.0814. The predicted wage growth if the economy is in a recession, unemployment is at 7.4%, and the GDP growth rate is 1.4% is 4.4025.

## 5. Conclusion

In conclusion, both models used decision trees to assist in making predictions. The first model assessed the likelihood of an individual defaulting on their credit based on certain characteristics. The second model assisted in predicting the wage growth given the status of the economy along with other variables.

In the first model, we created a classification decision tree using a set of data from a credit card company using variables such as defaulting on credit (default), missed payment within the last three months (*missed\_payment*), how much of an individual’s credit is allowed to be used (*credit\_utilized*), and assets owned by the individual (*assets*). We used a 70% and 30% split and separated the data into training and validation sets. We created a graph to show the validation error against the cost-complexity parameter (cp) and determined the most appropriate cp value needed to prune the tree was 0.021 as this was the largest cp value below the red line on the graph. We then created the plot for the pruned decision tree. We then obtained the confusion matrix and reported the counts for true positives, true negatives, false positives, and false negatives. We also calculated the values for accuracy which equaled 0.9666667, precision which equaled 0.96153846, and recall which equaled 0.98039216. We then made predictions to determine whether an individual was likely to default on their credit based on certain characteristics with the binary answers of ‘yes’ or ‘no’.

In the second model, we used an economic dataset to create a regression decision tree using variables such as the wage growth rate (*wage\_growth*), the unemployment rate (*unemployment*), whether or not the economy is in a recession (*economy*), and the GDP growth rate (*gdp*). We used an 80% and 20% split to separate the data into training and validation sets. We created the validation error against the cost-complexity parameter (cp) graph and determined the most appropriate cp value to prune the tree was 0.014 and used this value to create a plot for the pruned decision tree. We then calculated the root mean squared error (RMSE) to evaluate the performance of the model and found that the value was 1.0268. We made predictions to determine the wage growth rate based on the status of the economy, the unemployment rate, and the GDP growth rate.

Overall, the practical importance of the first analysis that was performed is that a credit card company can determine if it is appropriate to approve an applicant for a loan. The practical importance of the second analysis is that an economist can make predictions for the wage growth rate based on different scenarios.

## 6. Citations

Berrier, J. (2016). MAT 303: Applied Statistics 2 for Science. Zyante Inc. (zyBooks.com)