# MAT 303 Module Six Problem Set Report

Decision Trees

Brandon Hobbs

Brandon.Hobbs@snhu.edu

Southern New Hampshire University

## Introduction

Two CART analyses will be created in this report. The first tree will be a classification tree using the credit data from Week 5. The second data set consists of the economic data set from Week 3.

The credit data set being analyzed consists of 600 rows and 8 columns. Each row contains data about a particular credit seeking individual, e.g., *default*, *sex*, *education*, etc. See Figure 1 upper for the first few rows for data.

The economic data set being analyzed consists of 99 rows and 6 columns. Each row contains data about a particular economic period, e.g., *wage\_growth*, *inflation*, *gdp*, etc. See Figure 1 for the first few rows of data.

The data sets will be used to build trees model with the purpose of predicting an individual’s likelihood to default on credit (*default*) or the growth of wages (*wage\_growth*), respectively, from the other available data.  
  
First, the data in the csv-file will be ingested into a data frame so the R-language may be used for the stated purpose. Next, it will be plotted to provide a sense of the data and then the regression models, and their appropriateness, will be calculated. Finally, the models will be used to make predictions.

Table

Description automatically generated

Table

Description automatically generated

**Figure 1: First 5 Rows of Data from Credit Data (upper); First 5 Rows of Data from Economic Data (lower)**

## Data Preparation

To begin the analysis each data set was imported into their own data frame for consumption in the R-language. Of particular interest from the economic dataset are the *default*, *credit\_utilize, assets*, and *missed\_payment* parameters. From the economic dataset the parameters of interest are *wage\_growth*, *economy*, *unemployment*, and *gdp*. From this data two trees will be created.

The first tree will be a classification tree and will try to classify an individual’s likelihood of defaulting from *credit\_utilize* and *assets,* and *missed\_payments*.

The second tree will be a regression tree and will try to forecast an individual economic period’s wage growth from *economy*, *unemployment*, and *gdp*.

## Classification Decision Tree

The first tree will be a classification tree and will try to classify an individual’s likelihood of defaulting from *credit\_utilize* and *assets,* and *missed\_payments*.

### Reporting Results

The first step in building the model is to split the data into training and validation sets. The original data set consisted of 600 samples – this was split into 70% training and 30% validation. Table 1 shows the number of rows in each new set.

**Table 1: Training and Validation Samples used in Classification Tree**

|  |  |
| --- | --- |
|  | **Samples** |
| **Training Set** | 420 |
| **Validation Set** | 180 |

Once the data was segregated between training and validation sets the tree was generated using a seed value of 6751342. The unpruned tree cost-complexity parameter (CP) table is shown below in Table 2.

**Table 2: CP Table for the Unpruned Classification Tree**

|  |  |  |  |
| --- | --- | --- | --- |
| **CP** | **Number of Splits** | **Relative Error** | **Cross-validation Error** |
| 0.7930 | 0 | 1 | 1 |
| 0.0455 | 1 | 0.2071 | 0.2071 |
| 0.0101 | 3 | 0.1162 | 0.1212 |
| 0.01 | 5 | 0.0960 | 0.1313 |

To reduce the cross-validation error to a minimum, 5 splits were generated. This increase in number of splits also increases the complexity of the model which is not always advantageous as this can overfit the data. To help determine the “optimum” tree the CP value is plotted against the cross-validation error, see Figure 2.

Chart, line chart

Description automatically generated

**Figure 2: Cross-validation Error versus CP**

The first CP value to be at or below the red line, minimizing error and complexity, becomes a new control input and the model is generated again, with pruning at this minimized CP value. Figure 2 suggests an optimum CP value of 0.021.

The now pruned tree’s cost-complexity parameter (CP) table shown below in Table 3.

**Table 3: CP Table for the Pruned Tree**

|  |  |  |  |
| --- | --- | --- | --- |
| **CP** | **Number of Splits** | **Relative Error** | **Cross-validation Error** |
| 0.7930 | 0 | 1 | 1 |
| 0.0455 | 1 | 0.2071 | 0.2071 |
| 0.021 | 3 | 0.1162 | 0.1212 |

The final pruned classification tree is shown in Figure 3.

Timeline

Description automatically generated

**Figure 3: Pruned Classification Tree to Determine Default Probability**

**Evaluating Utility of the Model**

The pruned tree is then tested using the segregated validation data. The confusion matrix is shown in Table 4.

**Table 4: Confusion Matrix for Pruned Classification Tree**

|  |  |  |
| --- | --- | --- |
|  | **Prediction: default=no** | **Prediction: default=yes** |
| **Actual: default=no** | 74 | 4 |
| **Actual: default=yes** | 2 | 100 |

Which gives rise to the following common measures, Table 5, to help evaluate the model:

**Table 5: Common Measures of Classification Models**

|  |  |
| --- | --- |
| **Accuracy** | 0.9667 |
| **Precision** | 0.9615 |
| **Recall** | 0.9804 |

### Making Predictions Using the Model

With the new model created and optimized it is useable for predictions. As an example, two persons are considered for credit, Person A and Person B. Table 6 shows the relevant statistics for the two persons:

**Table 6: Relevant Statistics for Individuals Used in Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Person** | **Credit Utilization** | **History of Missed Payments** | **Assets** |
| A | 30% | No | Car and House |
| B | 30% | Yes | None |

Person A is not expected to default while Person B is expected to default. Figure 4 shows the path through the classification tree for each person.

Timeline

Description automatically generatedTimeline

Description automatically generated

**Figure 4: Classification Path for Person A (left); Classification Path for Person B (right)**

From these two predictions the model determines that, even with equally low credit utilization, the chances of defaulting are increased with fewer assets.

## Regression Decision Tree

The second tree will be a regression tree and will try to forecast an individual economic period’s wage growth from *economy*, *unemployment*, and *gdp*.

### Reporting Results

The first step in building the model is to split the data into training and validation sets. The original data set consisted of 99 samples – this was split into 80% training and 20% validation. Table 7 shows the number of rows in each new set.

**Table 7: Training and Validation Samples used in Regression Tree**

|  |  |
| --- | --- |
|  | **Samples** |
| **Training Set** | 79 |
| **Validation Set** | 20 |

Once the data was segregated between training and validation sets the tree was generated using a seed value of 6751342. The unpruned tree cost-complexity parameter (CP) table shown below in Table 8.

**Table 8: CP Table for the Unpruned Regression Tree**

|  |  |  |  |
| --- | --- | --- | --- |
| **CP** | **Number of splits** | **Relative error** | **Cross-validation error** |
| 0.7087 | 0 | 1.0000 | 1.0223 |
| 0.1290 | 1 | 0.2913 | 0.3582 |
| 0.0592 | 2 | 0.1623 | 0.2365 |
| 0.0238 | 3 | 0.1031 | 0.1799 |
| 0.0150 | 4 | 0.0794 | 0.1711 |
| 0.0123 | 5 | 0.0644 | 0.1386 |
| 0.0100 | 6 | 0.0520 | 0.1326 |

To reduce the cross-validation error to a minimum 6 splits were generated. This increase in number of splits increases the complexity of the model which is not always advantageous as this can overfit the data. To help determine the “optimum” tree the CP value is plotted against the cross-validation error, see Figure 5.

Chart, line chart

Description automatically generated

**Figure 5: Cross-validation Error versus CP**

As before, the first CP value to be at or below the red line becomes a new control input and the model is generated again. Figure 5 suggests an optimum CP value of 0.014.

The now pruned tree’s cost-complexity parameter (CP) table shown below in Table 9.

**Table 9: CP Table for the Pruned Tree**

|  |  |  |  |
| --- | --- | --- | --- |
| **CP** | **Number of splits** | **Relative error** | **Cross-validation error** |
| 0.7087 | 0 | 1.0000 | 1.0223 |
| 0.1290 | 1 | 0.2913 | 0.3582 |
| 0.0592 | 2 | 0.1623 | 0.2365 |
| 0.0238 | 3 | 0.1031 | 0.2303 |
| 0.0140 | 4 | 0.0794 | 0.2028 |

The final pruned classification tree is shown in Figure 6.

Diagram

Description automatically generated

**Figure 6: Pruned Regression Tree to Determine Wage Growth**

**Evaluating Utility of the Model**

The pruned tree is then tested using the segregated validation data. Instead of a confusion matrix and its calculable metrics to determine how correct the tree is a regression tress uses root mean square error (RSME). The lowest RSME possible, assuming a perfect match on the predicted and observed, would be 0%.

When the tree shown in Figure 6 is tested using the segregated validation dataset the RSME is 1.027%. This is low value, i.e., close to 0, and should be considered a good result – especially since the validation set is rather small (20 samples).

### Making Predictions Using the Model

With the new model created and optimized it is useable for predictions. As an example, two economic periods are considered, Period A and Period B. Table 10 shows the relevant statistics for the two periods:

**Table 10: Relevant Statistics for Individuals Periods used in Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Period** | **Within Recession** | **Unemployment Rate** | **GDP** |
| A | No | 3.4% | 3.5% |
| B | Yes | 7.4% | 1.4% |

Period A is expected to have a wage growth of 7.1% (7.0814) and Period B 4.4% (4.4025). Figure 7 shows the path through the regression tree for each period.

Timeline

Description automatically generated Timeline

Description automatically generated

**Figure 7: Regression Path for Period A (left); Regression Path for Period B (right)**

## Conclusion

Two CART analyses were performed over two different data sets. The classification tree attempted to predict how likely a person would default on credit and the regression tree attempted to predict the wage growth of a given economic period.

The classification tree’s performance is evaluated with a confusion matrix. The confusion matrix provides the results of the model by processing the validation data and measuring how many correct and incorrect predictions there are. The classification tree had an overall accuracy of 96.7% (ratio of correctness), a precision of 96.2% (ability to correctly predict a positive outcome), and a recall of 98.0%. As these are all above 95% this tree can be considered as performing well.

Moreover, the tree was pruned using the cost-complexity parameter (CP) so as not to overfit and be more generalized for any future data sets.

The regression tree’s performance is evaluated using RSME. The closer to zero the better the tree can predict an observed outcome. The computed tree has an RSME-value of 1.03%. Again, this value being rather low, i.e., less than 5%, the tree can be considered as performing well.

This tree was also pruned using CP to reduce the chance of overfitting the data.

Both models could be used, as is, by a person to decide wage growth or the chances of default. However, trees such as these are more valuable in machine learning techniques to characterize data sets.

## Citations

Hobbs, B. (2022). *MAT 303 module one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module two summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module three summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 project one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module five summary report*. [Unpublished report]. SNHU.