# 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 lower for the first few rows for 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*. 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 classify 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 |

**Evaluating Utility of the Model**

This first model will use two predictor variables as the modelling input to default probability. This model will be of the form:

Where y is ‘1’ for defaulting on credit and ‘0’ for not defaulting.

The final model is calculated as:

With *credit\_utilize* as X1, *edu2* as X2, and *edu3* as X3.

This may be linearized into the log-odds form as:

In the above equation π represents the probability (*p*) and the odds being equal to .

The model suggests that the odds of defaulting will increase 41% for each unit of credit utilization, i.e., , if education is held constant.

Based on this model and the assumption that if the probability of defaulting is over 50% then the following confusion matrix, Table 2, may be created:

**Table 2: Confusion Matrix for Model using *credit\_utilize* and *education***

|  |  |  |
| --- | --- | --- |
|  | **Prediction: default=0** | **Prediction: default=1** |
| **Actual: default=0** | 254 | 22 |
| **Actual: default=1** | 21 | 303 |

Which gives rise to the following common measures, Table 3, to help evaluate the model:

**Table 3: Common Measures of Logistic Models**

|  |  |
| --- | --- |
| **Accuracy** | 0.9283 |
| **Precision** | 0.9323 |
| **Recall** | 0.9352 |

To further determine if the model was relevant a Hosmer-Lemeshow Goodness of Fit (GOF) Test is conducted. A GOF is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: the model fits the data*

*Ha: the does not fit the data*

The null hypothesis states that there is a correlation between *credit\_utilize, education,* and *default*. The alternative states there is no correlation between *credit\_utilize, education,* and *default*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 4 shows the test statistic and its associated P-value:

**Table 4: Results for the Hosmer-Lemeshow Goodness of Fit (GOF) Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 31.582 |
| P-value | 0.9676 |

The P-value confirms that there not enough evidence to reject the null hypothesis, 0.9676 >> 0.05; thus, the model does fit the data. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the GOF test does not reveal is which of the predictor variables are relevant. To determine which are relevant a Wald Confidence Interval is conducted on each variable. The confidence intervals (95% confidence) can be used to determine statistical relevance, see Table 5.

**Table 5: Wald 95% Confidence Intervals**

|  |  |  |
| --- | --- | --- |
| ***Variable*** | **Lower Limit** | **Upper Limit** |
| ***(Intercept)*** | -11.1711 | -6.5265 |
| ***credit\_utilize*** | 26.4832 | 42.2906 |
| ***edu2*** | -2.4125 | -0.5826 |
| ***edu3*** | -5.4227 | -3.0854 |

Any variables’ interval that contains zero can be deemed statistically irrelevant. Since none of the intervals in Table 5 include zero, all the model’s predictor variables are relevant.

So far, the model has shown to be performing well based on the GOF test and all terms being relevant from the Wald confidence intervals. One more test is to evaluate the model’s discrimination sensitivity at various thresholds. The Receiver Operating Characteristic (ROC) Curve is a plot of the true-positive rate (sensitivity) against the false-positive rate (1 - specificity) – a large area under the curve (AUC) is desired with 1 being the “perfect classifier”.

Chart

Description automatically generated

**Figure 2: ROC Plot for *credit\_utilize* and *education***

The ROC curve shown in Figure 2 has an area under the curve of 0.9859. This large value (~1) again confirms the value and relevance of the generated model.

### Making Predictions Using the Model

With the new model created and confirmed as relevant 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** | **Education Level** |
| Person A | 35% | high school |
| Person B | 35% | postgraduate |

The odds of Person A defaulting are 96.0%:

The odds of Person B defaulting are 25.59%:

From these two predictions the model determines that, even with equally low credit utilization, the chances of defaulting are dramatically reduced in persons who have completed more education than those with less.

## Second Logistic Regression Model with *credit\_utilize, assets, missed\_payment*

The second will try and predict the probability of defaulting from *credit\_utilize, assets*, and *missed\_payment*.

### Reporting Results

The first step in building the model is to relevel the qualitative predictors *assets* and *missed\_payment*. In the data there are four possible values of *assets* and two for *missed\_payment*. Table 7 and 8 show the dummy variables – as before a 1 indicates the presence of a quality:

**Table 7: Dummy Variables for *assets***

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***asset1*** | ***asset2*** | ***asset3*** |
| **none** | 0 | 0 | 0 |
| **car only** | 1 | 0 | 0 |
| **house only** | 0 | 1 | 0 |
| **car and house** | 0 | 0 | 1 |

**Table 8: Dummy Variables for *missed\_payment***

|  |  |
| --- | --- |
|  | **missed** |
| Payments missed in last 3 months | 1 |
| No payments missed in last 3 months | 0 |

This second model will use three predictor variables, one quantitative and two qualitative, as the modelling input to default probability. This model will be of the form:

Where y is ‘1’ for defaulting on credit and ‘0’ for not defaulting.

The final model is calculated as:

With *credit\_utilize* as X1, *asset1* as X2, *asset2* as X3, *asset3* as X4, and *missed* as X5.

This may be linearized into the log-odds form as:

Based on this model and the assumption that if the probability is over 50% a person will default then the following confusion matrix, Table 9, may be created:

**Table 9: Confusion Matrix for Model using *credit\_utilize, assets, missed\_payment***

|  |  |  |
| --- | --- | --- |
|  | **Prediction: default=0** | **Prediction: default=1** |
| **Actual: default=0** | 262 | 14 |
| **Actual: default=1** | 21 | 303 |

Which gives rise to the following common measures, Table 10, to help evaluate the model:

**Table 10: Common Measures of Logistic Models**

|  |  |
| --- | --- |
| **Accuracy** | 0.9417 |
| **Precision** | 0.9558 |
| **Recall** | 0.9352 |

To further determine if the model was relevant a Hosmer-Lemeshow Goodness of Fit (GOF) Test is conducted. A GOF is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: the model fits the data*

*Ha: the does not fit the data*

The null hypothesis states that there is a correlation between *credit\_utilize, assets, missed\_payment,* and *default*. The alternative states there is no correlation between the variablesand *default*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 11 shows the test statistic and its associated P-value:

**Table 11: Hypothesis Test for the Hosmer-Lemeshow Goodness of Fit (GOF) Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 26.733 |
| P-value | 0.9945 |

The P-value confirms that there not enough evidence to reject the null hypothesis, 0.9945 >> 0.05; thus, the model does fit the data. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the GOF test does not reveal is which of the predictor variables are relevant. To determine which are relevant a Wald Confidence Interval is conducted on each variable. The confidence intervals (95% confidence) can be used to determine statistical relevance, see Table 12.

**Table 12: Wald 95% Confidence Intervals**

|  |  |  |
| --- | --- | --- |
| ***Variable*** | **Lower Limit** | **Upper Limit** |
| ***(Intercept)*** | -11.6518 | -6.8224 |
| ***credit\_utilize*** | 24.4513 | 40.1140 |
| ***assets1*** | -1.4624 | 0.4971 |
| ***assets2*** | -4.2167 | -1.8501 |
| ***assets3*** | -4.5947 | -2.3189 |
| ***missed*** | 0.6178 | 2.2373 |

Any variables’ interval that contains zero can be deemed statistically irrelevant. Only credit utilization (*credit\_utilize*) contains a zero and may be considered statistically irrelevant.

So far, the model has shown to be performing well based on the GOF test and only a single term being irrelevant from the Wald confidence intervals. The ROC Curve is shown in Figure 3.

Chart

Description automatically generated

**Figure 3: ROC Plot for *credit\_utilize* and *missed\_payment* and *assets***

The ROC curve shown in Figure 3 has an area under the curve of 0.9874. This large value (~1) again confirms the value and relevance of the generated model.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, two persons are considered for credit, Person A and Person B. Table 13 shows the relevant statistics for the two persons:

**Table 13: Relevant Statistics for Individuals Used in Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Person** | **Credit Utilization** | **Assets** | **History of Missed Payments** |
| Person A | 35% | only a car | yes |
| Person B | 35% | car and home | no |

The odds of Person A defaulting are 95.3%:

The odds of Person B defaulting are 19.9%:

From these two predictions the model determines that, even with equally low credit utilization, the chances of defaulting are dramatically reduced in persons who have more assets and do not have a history of missing payments.

## Conclusion

Two binary logistic models were built. Both models proved to be statistically relevant via their GOF tests, i.e., null hypotheses were not rejected, and only the second model has a predictor variable that is suspect, but which model should be used?

Table 14 shows the comparative metrics on each of the models:

**Table 14: Comparative Metrics for the Two Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **AUC** |
| **Model 1** | 0.9283 | 0.9323 | 0.9352 | 0.9859 |
| **Model 2** | 0.9417 | 0.9558 | 0.9352 | 0.9874 |

As shown in the preceding table Model Two has higher accuracy (ratio of correctness), precision (ability to predict a true positive), and AUC (closer to 1 is better performance). Therefore, even though Model Two had one predictor variable not statistically relevant, it appears to be the superior model.

This type of model could be used by a bank, credit company, or any other institution trying to evaluate the financial risk of anyone applying for credit. However, I do not believe the models are complete. In the available data there are other attributes that may be valuable, e.g., marriage status for joint applications or age (a young person will have very little history compared to an older person).

Before the Model Two is used adding *age* and *marriage* as predictor variables is advised but will be left for another day.

## 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 5 summary report*. [Unpublished report]. SNHU.