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

Decision Trees

Tammy Hartline

[tammy.hartline@snhu.edu](mailto:tammy.hartline@snhu.edu)

Week 6

Southern New Hampshire University

## 1. Introduction

*For this week’s analysis we are exploring the ‘credit\_card\_default’ dataset. It contains historical data related to credit card customers and their credit behavior. The dataset includes prediction variables, such as education level, assets owned, missed payments in the past three months, and credit utilization. The response variable being used is the default variable. (a binary value indicating whether the customer has defaulted on their credit)*

*The results of this analysis have one primary use as a risk analysis. It is used to assess the creditworthiness of individuals, helping financial institutions make informed decisions about lending money, including who they lend to, and how much the customer qualifies for, with the least chance of the customer defaulting on their line of credit. (Not paying it back) This analysis can also be used for risk assessments, developing credit scoring models, and identifying strategies to mitigate credit default risks.*

*The analysis will be performed using a classification decision tree, which will then be studied as a binary classification analysis.*

*The second dataset we will be working with is the ‘economic’ dataset. It includes economic factors, such as wage growth rate, inflation rate, economy status, education level, and GDP growth rate.*

*For this dataset we will performing a regression analysis, after creating a regression decision tree model, meant to predict wage growth based on economic features.*

*This information can be valuable for economic policy planning and analyses.*

## 2. Data Preparation

*In the credit\_card\_default dataset, there are several important variables, such as age, which represents the age of the potential borrower and can be a crucial factor when determining credit worthiness, as younger borrowers may have less financial history to consider when compared those advanced in age. Another important variable is sex, indicating gender, which can be associated with disparities and financial behaviors. Next, education level, which is the highest education attained by the person, and could be used to help determine the borrower’s financial literacy as well there earning potential. Then we have marriage: status of marriage for the borrower, assets: indicates whether the borrower owns assets of value such as a car and/or home, which can be used to reflect the person’s financial stability and ability to repay their debts, missed payment: specifies whether the individual has missed payments in the past 3 months, as past payment behavior is a strong predictor of future default probability, credit utilization: which measures how much of their total available credit is being used, and can be used to estimate financial stress in cases where the borrower has maxed out or close to maxed out usage of credit. The final important variable is the default variable, and it is our response variable that indicates whether the borrower has defaulted on their credit or not.*

*For this dataset, there is a total of 600 rows and 8 columns. Of those, we split the rows into 2 sets. The total rows were split with 60% to 40% respectively, resulting in a training set row count of 360, and a validation set row count of 240, which accounts for all 600 rows and 8 columns.*

## 3. Classification Decision Tree

### Reporting Results

*Using 6,751,342 as the set seed, with a 70%:30% split, there are now 420 rows for the training set, and 180 rows for the validation set, compared to the original 600 rows before the split, or the previous split when the set seed was 507,690, with a 60%:40% split, there were 360 rows for the training set, and 240 rows for the validation set.*

*Next, using the 6,751,342 set seed, I created a classification decision tree for the default variable, using missed payment, credit utilization and assets as my predictors. I also included the cp, (cost-complexity table, which can be seen below:*

*A computer code with numbers and symbols

Description automatically generated*

*After plotting the validation error against the cost-complexity parameter (cp), I was able to easily determine (0.021) as the appropriate cp value to be used in pruning the tree, which can be seen in the plot below:*

*A graph with a line

Description automatically generated*

*Using the appropriate value from the previous paragraph of 0.021, we get the following classification tree result:*

*A computer code with numbers and symbols

Description automatically generated*

*Pictured below, is the plot results for the updated classification tree:*

*A diagram of a credit

Description automatically generated*

### Evaluating Utility of Model

*Evaluate the utility of the classification decision tree. Address the following questions in your analysis:*

*The model correctly predicted that 100 borrowers would default for our true positives value, where they did indeed default. The true negatives for the model correctly predicted that 74 borrowers would not default, and they did not. The false positives for the model incorrectly predicted that 4 borrowers would default, where in fact, they did not. Lastly, the false negatives for the model incorrectly predicted that two borrowers would not default, where they actually did indeed default on their credit.*

*For this model the following formulas are how we determined the accuracy, precision, recall, and specificity:*

*Accuracy is measured as (TP + TN) / (TP + TN + FP + FN),*

*In this case the calculation was (100 + 74) / (100 + 74 + 4 + 2) = 174/180 = 0.9667,*

*meaning the model has a high accuracy rate of* ***96.67%***

*Precision is measured as (TP) / (TP + FP),*

*In this case the calculation was (100) / (100 + 4) = 100/104 = 0.9615,*

*indicating when our model predicts a customer will default, the prediction is correct approximately 96.15% of the time.*

*Recall is measured as (TP) / (TP + FN),*

*In this case the calculation was (100) / (100 + 2) = 0.9804,*

*meaning the model correctly identifies 98.04% of customers who actually default on their credit.*

*Specificity is measured as (TN) / (TN + FP),*

*In this case the calculation was (74) / (74 + 4) = 74/78 = 0.9487,*

*indicating that the model correctly identifies 94.87% of the customers who did not default on their credit.*

*Overall the classification decision tree model appears to be highly accurate in predicting credit card default. It has a high accuracy rate, and it performs well in terms of precision, recall, and specificity, indicating its utility in identifying borrowers at risk of defaulting. Though, it is worth noting, it is always essential with any analysis to consider other factors not included in our assessments, such as business objectives, and the cost of false positives and negatives when assessing the model’s utility for real world context.*

### Making Predictions Using Model

*In this section, we were given the task of performing two predictions. The first prediction involved looking at a borrower who had not missed any payments within the last three months, owns a car and house, and utilized 30% of their available credit. The prediction returned ‘no.’ Meaning, based on the model, this borrower is predicted* ***not*** *to default on their credit.*

*For the second prediction, we looked at a borrower who had missed at least one payment in the past three months, and who did not own a car or house, but also utilized 30% of their total available credit. This prediction returned ‘yes.’ Meaning, based on the model, this borrower* ***is*** *predicted to default on their credit.*

*Each of these predictions are for their specific variables and use cases. It is worth noting that these predictions are probabilistic, and there may still be some uncertainty involved.*

## 4. Regression Decision Tree

### Reporting Results

*In this dataset, (economic), the original row count was 99, and the column count was six. After setting the seed to 6,751,342 and splitting the economic data into training and validations sets with an 80%:20% split, respectively, there are 79 training set rows, and 20 validation set rows.*

*With the seed set to 6,751,342, I created a regression decision tree for wage-growth, using economy, unemployment, and GDP as predictor variables. The following was a result of that decision tree and includes the cost-complexity (cp) table:*

*A screenshot of a computer code

Description automatically generated*

*Next, I plotted the validation error against the cost-complexity parameter (cp), to determine the appropriate value for pruning the tree. As you can see in the image below, the appropriate value to prune the tree in this case is 0.014:*

*A graph with a line

Description automatically generated*

*The next step was to prune the tree using the appropriate cp value (0.014), which resulted in the following cp table:*

*A screenshot of a computer code

Description automatically generated*

*The final step, was to plot the pruned tree, which can be seen below:*

*A diagram of a graph

Description automatically generated*

### Evaluating Utility of Model

*To evaluate the utility of the model, I found the RMSE (root-mean-squared-error), which was 0.8065. The interpretation of this in the context of our model is the RMSE value of approximately 3.8065 suggests that, on average, the predictions made by the regression decision tree model for wage growth have an error of approximately 3.8065 units when compared to the actual wage growth values in the testing dataset. We use RMSE as a measure of the prediction error of a regression model, as it quantifies the difference between the predicted values and the actual observed values in the dataset.*

*A lower RMSE value indicates the model’s predictions are closer to the actual value, which is the desired result for regression tasks.*

*In this specific scenario, our RMSE value of 3.8065, indicates the model’s predictions for wage-growth may have some level of error, and the actual wage-growth values could deviate from the predictions by around 3.8065 units on average. As to whether this an acceptable error level, would depend on the specific context of the problem and the application’s requirements.*

### Making Predictions Using Model

*Here I was given a scenario where the wage-growth was to be predicted for an economy* ***not*** *in a recession, with an unemployment rate of 3.4%, and the GDP growth rate of 3.5%. The prediction result for this scenario was 7.0814, which can be translated as, the model predicts that with these specific circumstances, wage-growth is expected to be approximately 7.0814.*

*The next scenario given was again predicting wage-growth, but with the economy being* ***in*** *a recession, with an unemployment rate of 7.4%, and the GDP growth rate of 1.4%. The prediction result for this scenario was 4.4025, which can be translated as, the model predicts that with these specific circumstances, wage-growth is expected to be approximately 4.4025.*

*These predictions can be valuable for economic planning and analysis, helping to understand how different economic factors impact wage-growth.*

## 5. Conclusion

The results of the statistical analyses in this report provide valuable insights into two different scenarios involving decision trees: a classification decision tree for predicting credit card default and a regression decision tree for predicting wage growth based on economic factors.

Classification Decision Tree Analysis:

The classification decision tree effectively predicts credit card default with a high degree of accuracy, precision, recall, and specificity. These statistical terms are defined as follows:

Accuracy: The proportion of correct predictions out of the total predictions. In this analysis, it indicates that 96.67% of predictions were accurate, meaning that the model correctly identified whether a customer would default or not in nearly 97% of cases.

Precision: Precision measures the percentage of true positive predictions (correctly identifying customers who defaulted) among all positive predictions (the sum of true positives and false positives). In this case, it tells us that when the model predicts a customer will default, it is correct approximately 96.15% of the time.

Recall: Recall measures the percentage of true positive predictions among all actual positive cases (true positives plus false negatives). Here, the model correctly identifies 98.04% of customers who genuinely defaulted.

Specificity: Specificity measures the percentage of true negative predictions (correctly identifying customers who did not default) among all actual negative cases (true negatives plus false positives). The model correctly identifies 94.87% of customers who did not default.

The practical importance of this classification decision tree lies in its utility for financial institutions. It can help them make informed lending decisions by assessing the creditworthiness of individuals. High accuracy and precision mean that the model can identify high-risk customers, reducing the chance of default. However, it's essential to consider other real-world factors and business objectives when applying this model.

Regression Decision Tree Analysis:

The regression decision tree predicts wage growth based on economic factors, with a resulting Root Mean Squared Error (RMSE) of approximately 3.8065. RMSE is a statistical measure that quantifies the difference between predicted values and actual observed values in the dataset. A lower RMSE indicates more accurate predictions.

In this context, an RMSE of 3.8065 suggests that, on average, the predictions made by the regression decision tree model for wage growth have an error of approximately 3.8065 units when compared to the actual wage growth values in the testing dataset. This means that the model's predictions may deviate from actual wage growth values by around 3.8065 units on average.

The practical importance of this regression decision tree lies in its ability to provide economic insights. It allows policymakers and analysts to understand how different economic factors, such as recession status, unemployment rate, and GDP growth rate, impact wage growth. However, the acceptability of this level of prediction error depends on the specific context and requirements of the application.

In summary, both decision tree models offer practical utility in their respective scenarios, with the classification model aiding in credit risk assessment and the regression model supporting economic policy planning and analysis. However, users should consider the limitations and real-world factors when applying these models in practice.