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

Decision Trees

Catherine Taft

Catherine.taft@snhu.edu

Southern New Hampshire University

## 1. Introduction

Miraculously I’ve managed to hold onto this job as a risk assessor at the credit card company (obviously they didn’t check my credit history) which is great because I suspect my family was planning an intervention. I must live my truth but I must also feed my cats premium cat food so here we are.

In the meantime, we’re again looking at whether or not we should extend loans to people based on a default risk calculated from certain financial and demographic information. This time however we’ll be using decision trees of two types (classification and regression).

## 2. Data Preparation

The variables we’ll be using are mostly the same as before: assets (house, car, both), credit utilization, missed payments within the last three months, education, and age. Then just for the sake of nostalgia we’ll run a decision tree on the economic dataset from my old job, where we’ll use economy (in recession or not), GDP and inflation rate to determine wage growth.

There are 600 rows and 8 columns in the credit card dataset, and 99 rows and 6 columns in the economic dataset.

## 3. Classification Decision Tree

### Reporting Results

Having split the data 70/30, now we can see that there are 600 rows in the original data set, 420 rows in the training set and 180 rows in the validation set.

Here’s the cost complexity table and the plot of the cp values against the relative error:

Text

Description automatically generated

Chart, line chart

Description automatically generated

Generally when selecting a cp value we want the largest one that’s still below the red line. This graph threw me because the x-axis is decreasing not increasing (?!) but based on that we’d choose 0.028 as our value. We will use this to prune our decision tree (I love the tree vocabulary so much):

Timeline

Description automatically generated

### Evaluating Utility of Model

Here’s the confusion matrix for that model:

Table

Description automatically generated with medium confidence

This gives us 87 true negatives, 4 false negatives, 84 true positives and 5 false positives.

**Accuracy**

The accuracy is:

Which evaluates to 0.95 or 95%.

**Precision**

The precision is:

This evaluates to 0.944 or 94.4%.

**Recall**

The recall is:

That evaluates to 0.9545 or 95.45%.

All of that indicates that this is a pretty good model.

### Making Predictions Using Model

Next we have a couple of predictions. The first one is whether or not a person is likely to default if they have not missed payments, own a car and a house, and have 30% credit utilization. The result was no, they are not likely to.

The second prediction is if someone will default if they have missed payments, do not have any assets at all, and have a 30% credit utilization, and the result is that yes they’re likely to.

This makes a lot more sense than the predictions in the previous problem set where it made it seem like a master’s degree alone would keep you from defaulting on your credit card bills.

## 4. Regression Decision Tree

### Reporting Results

Now we’ll violate my security clearance and go back and use a data set from my previous government job analyzing labor statistics to predict wage growth, just for fun.

First we’ll split it 80/20. The original data set has 99 rows. After splitting it, there are 79 rows in the training set and 20 rows in the validation set.

Then we’ll run the model and create the cp table and then the plot of the cp values against relative error:

Text

Description automatically generated

Chart, line chart

Description automatically generated

Based on this we’d choose the cp value of 0.035 (it’s actually very slightly below the line) to use for pruning the tree.

Here is the resulting decision tree:

Timeline

Description automatically generated

### Evaluating Utility of Model

The root mean squared error is 0.8386. The RMSE is the standard deviation of the residuals. In general a lower RMSE is better and this is pretty close to 1 so that’s maybe not great and could indicate a higher level of error than may be ideal.

### Making Predictions Using Model

Next we’ll do two predictions. The first is what the wage growth will be if the economy is not in recession, unemployment is at 3.4% and the GDP growth is at 3.5%. The model predicts it’ll be 7.7924%.

The second is what the wage growth will be if the economy IS in recession, unemployment is at 7.4% and GDP growth rate is 1.5% (that sounds horrible). For this it predicts it’ll be at 2.6364%. Yikes.

## 5. Conclusion

Both these types of decision trees have interesting applications. The classification tree basically gave us a yes/no prediction for a binary variable but also probability of default (if I’m understanding correctly) based on our parameters, which are both useful in the context of risk assessment. If the yes/no doesn’t satisfy you, you could make a more educated decision based on the provided likelihoods.

The regression tree provided a prediction for wage growth based on walking through a provided scenario which is also useful though honestly I’m not sure why a decision tree specifically is the best in terms of a specific prediction like in the two scenarios, since the previous models we used seemed to behave similarly. However the printed tree plot gave a lot of information about what wage growth would be based on specific parameter values which is very useful.