Anna Nguyen
Branden Ho

Case 4 Report Team 5

1. Boston Housing data

i: Fit a regression tree (CART) on the training data. Report the model's in-sample MSE performance.

Without cp value specified	In-Sample MSE
Original Regression Tree	16.69697

ii:Test the out-of-sample performance. Using tree model built from (i) on the raining data, test with the remaining 10% testing data. Report out-of-sample model MSE.

Without cp value specified	Out of sample MSE
Original Linear Regression vs Testing Data	17.66738

With cp value specified	In sample MSE
Optimal Regression Tree vs Training Data	14.21091

With cp value specified	Out of sample MSE
Optimal Regression Tree vs Testing Data	19.49892

Description: As we compare the out of sample MSE from the optimal tree to the original tree, we can see that the optimal tree actually has a slightly higher MSE. To an extent this was within expectation because we found the best cp and now the tree is fit to find

the optimal complexity, but the cp went from what is normally 0.001 to 0.006. This means that our tree is going to be a bit more complex and therefore have a bit more error due to how it is more fitted.

iii:Conduct linear regression using all explanatory variables except "indus" and "age" on the training data. Report the model's in-sample MSE. Test the out-of-sample performance with the remaining 10% testing data.
Report out-of-sample model MSE etc?

	MSE
Linear regression In Sample	23.23503
Linear regression Out of Sample	16.09352

Description: From this table we can see that the linear regression has the opposite effect for this data. The linear regression of the best model actually has a lower MSE with the out of sample data.

Iv:What do you find comparing CART to the linear regression model fits from (iii)?

While the linear regression may show that as a model it is doing a good job of handling error with the out of sample data, the tree regression also allows us the freedom to find the best complexity parameter and then build the optimal model. In the end, we believe that each has their own strengths. The linear regression is very straightforward and can produce solid results with modeling, but the tree regression allows for an easy to follow tree graph and has the cp evaluator.

2. German Credit Score data

(i).

In-sample performance after pruning the tree to find optimal tree model:

Symmetric cost (after specifying symmetric cost function)	0.5

Asymmetric cost (after specifying asymmetric cost function)	0.62125
Misclassification cost (calculated from the matrix table)	0.3275
Asymmetric misclassification cost (calculated from the matrix table)	0.3725

After pruning the tree, we can see that 32.75% of the training observations are misclassified, or 37.25% misclassified if we take the ratio of 5 from the revised cutoff value into our calculation. Comparing this misclassification cost with the symmetric and asymmetric cost, we can see that it has a much lower error rate. Also, it is worth noting that after pruning the tree, our error rate is lower compared to the original error rate. Therefore, the pruning process has improved our classification accuracy a lot.

(ii).

Out-of-sample performance:

AUC	0.7331731
Symmetric cost (after specifying symmetric cost function)	0.5
Asymmetric cost (after specifying asymmetric cost function)	0.7175
Misclassification cost (calculated from the matrix table)	0.39
Asymmetric misclassification cost (calculated from the matrix table)	0.57

The AUC is greater than 0.7 so it has acceptable discriminatory power. Also, it can be seen that 39% of testing observations are misclassified, or 57% misclassified if we take the ratio of 5 from the revised cutoff value into our calculation. Comparing this with our symmetric and asymmetric cost, it has a lower error rate but the only difference is that our asymmetric misclassification cost has a higher error rate than the symmetric cost.