Prediction Challenge 2

Jeevanandan Ramasamy

Initial Steps

- I first created a default decision tree with the full training data without any adjustments
- Since I received a high error value, I tried adjusting the parameters to the rpart function such as minsplit and cp
- I was able to get the error to 12% at best

Cleaning the Data

- I found many outliers that interfered with the prediction accuracy of the decision tree
- So I removed outliers that seemed too far away from the range of values for each category
- I decided to keep some outliers that do not fit in the ranges of other categories

Refining the Decision Tree

- Since I had a very high accuracy for the first Prediction Challenge, I used the results I got as a baseline for this challenge
- After cleaning the data, I applied rpart again and got an error of about 4-5% with the training dataset
- I also had an error of 4% when compared to my first prediction dataset

Refining the Decision Tree Cont.

- Afterward, I modified the parameters to rpart to match the new data
- This decreased the overall error to around 3%
- Cross validation of this decision tree yielded around 88% average accuracy for the subset and 0.05% average variance for the subset

Final Prediction Model

- I decided to split the data into the same categories I did for the first prediction challenge and created a decision tree for each one
- This resulted in a total of 8 decision trees
- After adjusting minbucket and cp again, I was able to get the error to less than 11%
- Though I am not sure if this prediction model surpasses my previous one, I have high hopes for this challenge