15.095 Homework #4 Executive Summary

Hyper-parameter tuning methodology: In general, hyperparameters were first tuned for a few points over a large range. Once a general region of optimal values was found, a more granular search would be applied in that region. This was an effort to find a globally optimal value without impractical computation time.

Model 1 *(CART):*

Out-of-sample Accuracy: 0.5941, Out-of-sample AUC: 0.6115

Variables Used: The first three variables split on in the tree was the political leaning of the lower court, which district court it presided in, and the respondent type.

Model 2 *(OCT):*

Out-of-sample Accuracy: 0.6412, Out-of-sample AUC: 0.6533

Classification Trees Interpretability: Typically, CART & OCTs are viewed as some of the most interpretable models. However, given the data set is only multi-factor categorical variables, the typical interpretability suffers. No longer can you easily convey a simple tree to a decision maker.

Variables Used: The first three variables split on in the tree was the respondent type, petitioner type, and which circuit court the case was in previously.

Model 3 *(Random Forest):*

Out-of-sample Accuracy: 0.6882, Out-of-sample AUC: 0.735

Random Forest Interpretability: This method provides no easy visual to a decision-maker and thus is not very interpretable. However, one could relay the concept in layman's terms as "the average of many decision trees", with mixed success.

Variables Used: The random forest inevitably used every variable because it randomly selects a subset of variables for each tree in the forest.

Model 4 *(XGBoost):*

Out-of-sample Accuracy: 0.6412, Out-of-sample AUC: 0.6820

XGBoost Interpretability: This model is even less interpretable than random forests because there is no easy way to explain the prediction of errors to a non-data scientist.

Variables Used: Due to the large number of trees XGBoost leverages, it almost inevitably used all the variables. However, one could extract the most important features.

Model 5 *(Sparse Logistic Regression):*

Out-of-sample Accuracy: 0.6647, Out-of-sample AUC: 0.6645

Sparse Logistic Regression Interpretability: This model has slightly better interpretability compared to random forests because sparsity allows one to communicate to a decision maker

the important variables. The regression easily conveys information through coefficient's respective signs, but not interpretable values.

Variables Used: The model was tuned to only have one variable so the only one used was whether or not the lower court was liberal.

Model 6 *(OCT-H):*

Out-of-sample Accuracy: 0.6706, Out-of-sample AUC: 0.6593

Additional Model Selection: The final model selected was optimal classification trees with hyperplanes because it can achieve similar results to XGBoost with a tree of small depth, and thus increased interpretability.

OCT-H Interpretability: OCT-H achieve slightly less explainability compared to CART, which places it near the top of models in terms of interpretability.

Variables Used: OCT-H issued a combination of numerous dummy variables for levels of different categorical variables in this particularly unsparse OCT-H.

Model Selection Process: Since the most interpretable models, trees, are hardly interpretable because of the multi-class categorical variables, I decided the most important aspect in the model was out-of-sample performance. Thus, I would select the random forest model due to its accuracy and the fact that whichever model I select in this scenario, there will be substantial effort in explaining it. The biggest downside of a random forest as opposed to a difficulty interpretable CART is the lack of a visual. However, in this case I think performance trumps a confusing visual.