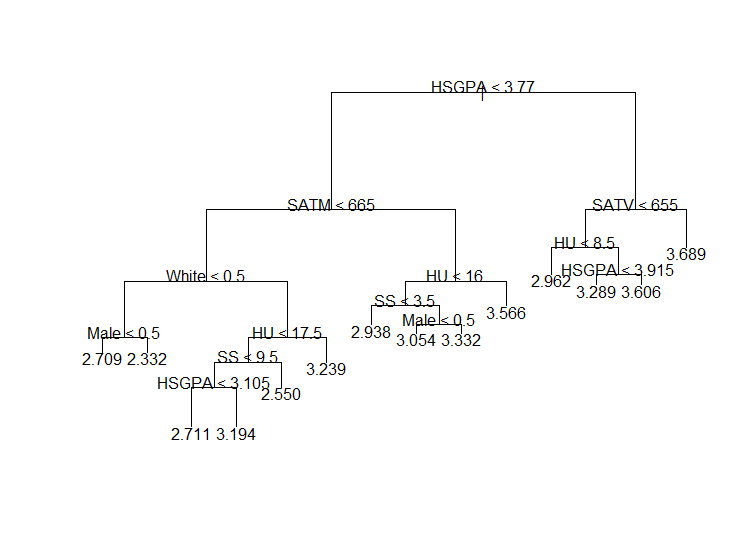
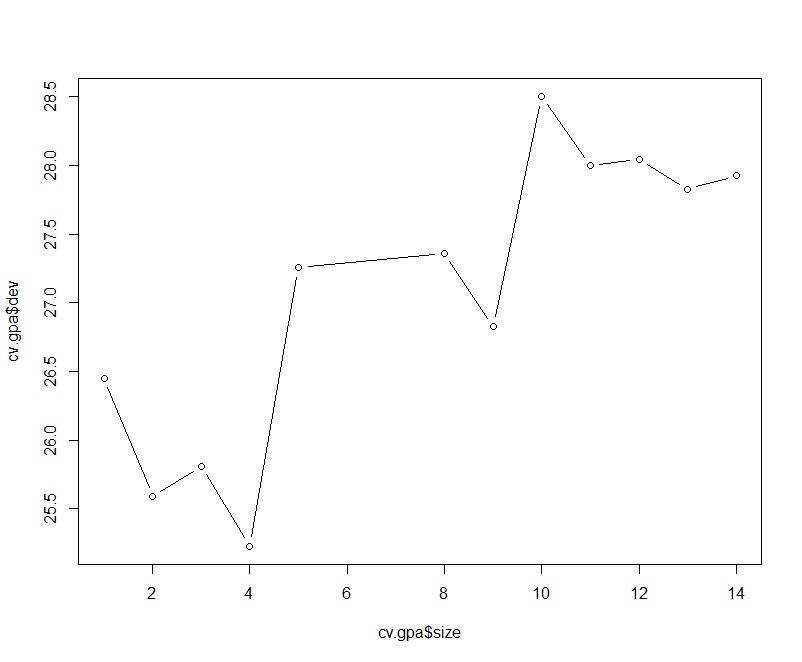
Decision Trees and Random Forests with Boosting and Bagging

In this project we are working with titled “FirstYearGPA”. This is a dataset we have used before. It has one response variable named “GPA” which is a quantitative variable that represents a student’s GPA after their first year in college. In addition to the response variable, the dataset also has 9 predictor variables named “HSGPA”, “SATM”, “SATV”, “White”, “HU”, “Male”, “FirstGen”, “CollegeBound”, and “SS”. The goal of this assignment was to predict the value of our response variable (GPA) using decision trees, random forests, boosting, and bagging. Once we have created several different models that can make predictions, we will compare our findings with what we found the last time we worked with this dataset.

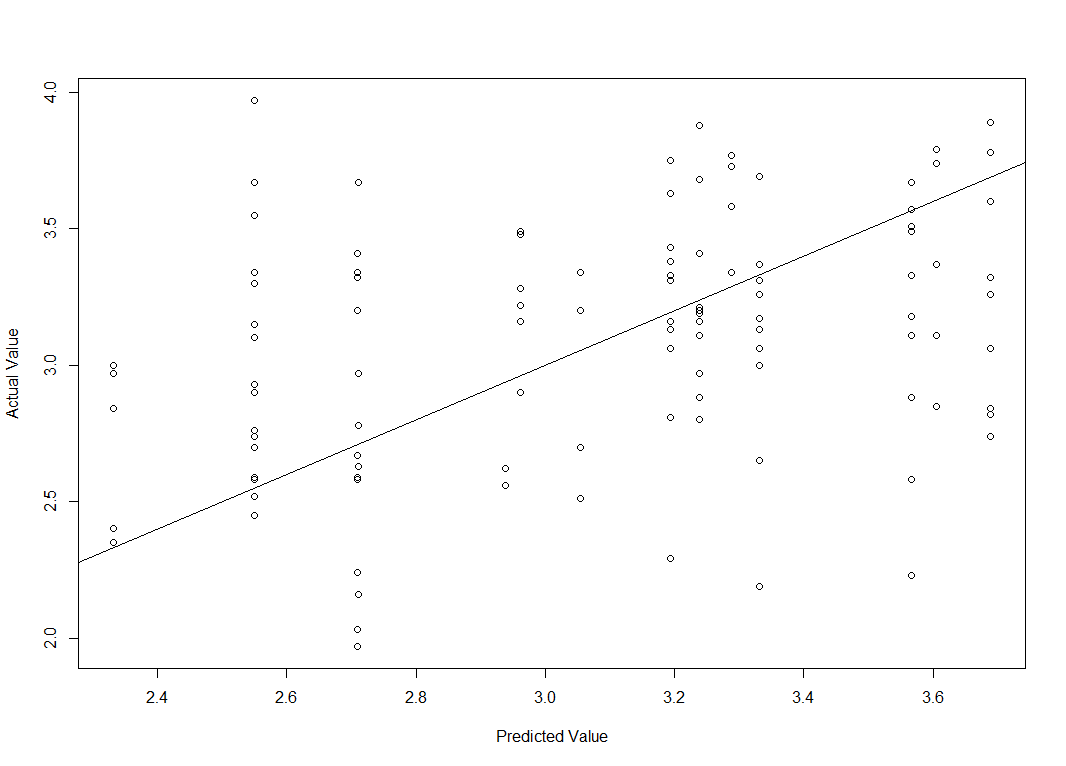
To begin we first split our dataset into test and training groups. We then fit a “large” tree to the training group we created. The variables that were included in this tree were “HSGPA”, “SATM”, “White”, “Male”, “HU”, “SS”, and “SATV”. The plot of this tree has been included below.



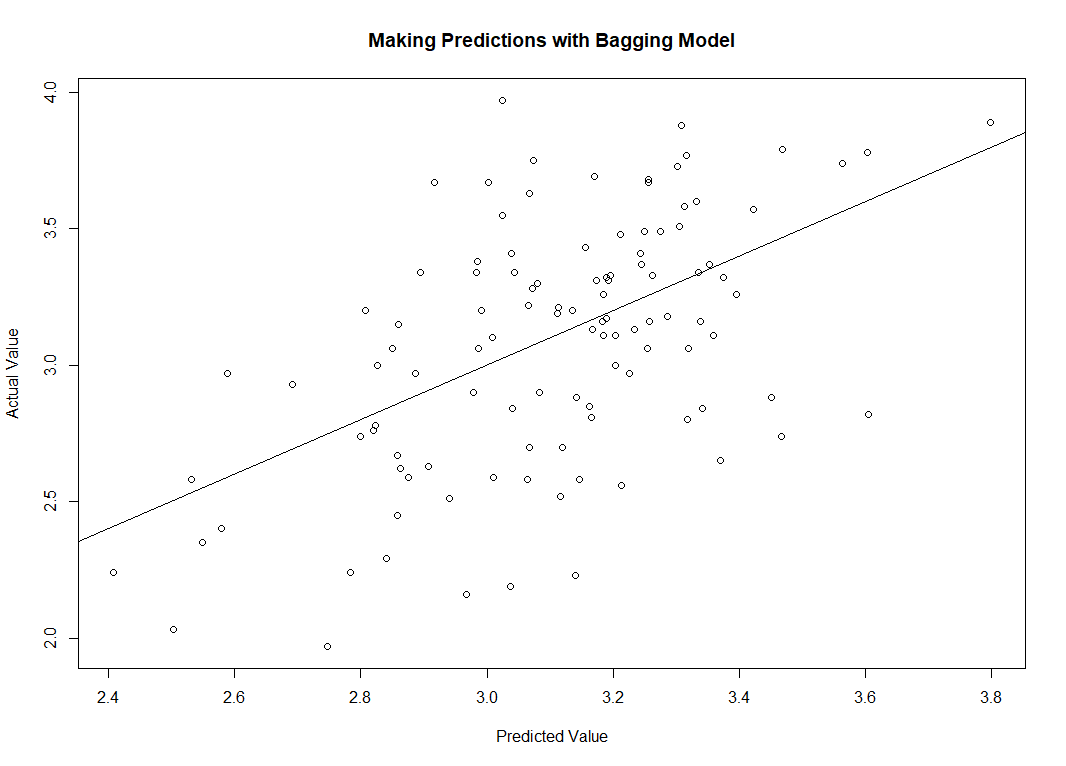
After we created the largest possible tree with our training data the next step was to prune it. To figure out how many terminal nodes our pruned tree should have, we used 5-fold cross validation and plotted the deviance versus the size. This plot has been included below.



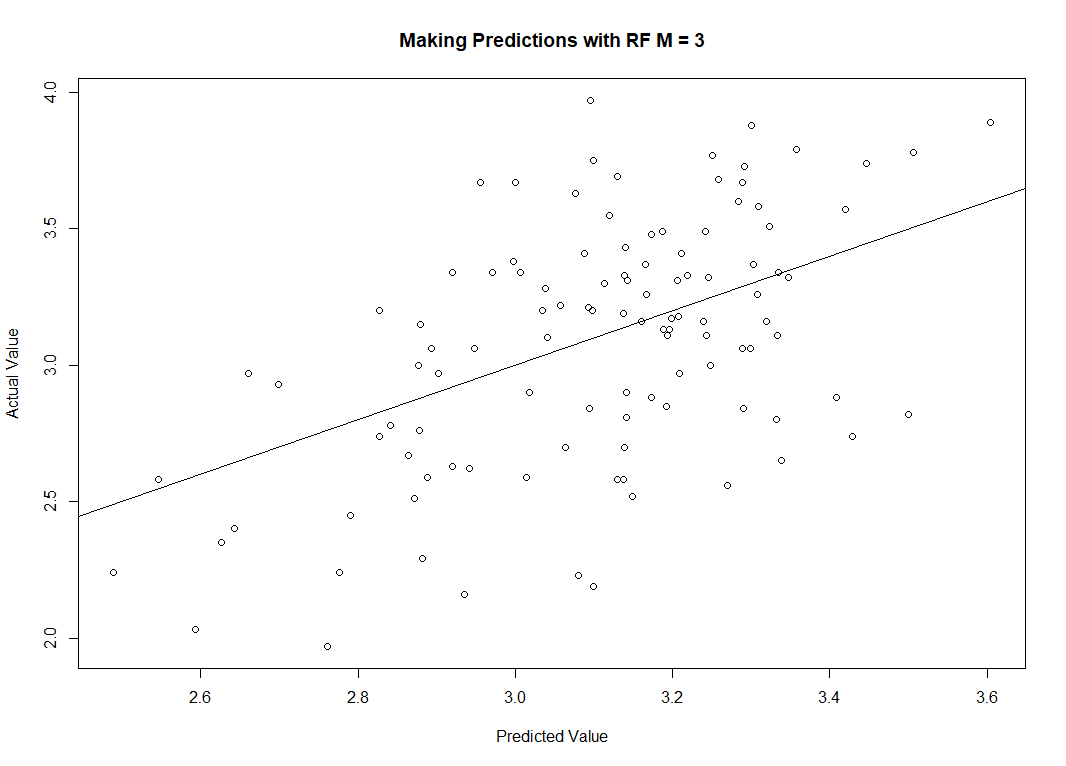
From the plot above, we can see that the deviance was lowest with trees that had 4 terminal nodes. Thus, we will prune our tree to have 4 terminal nodes. After creating our pruned tree, we made predictions and plotted them against the actual values from the test set. The following plot was generated. We also found that the test MSE = 0.241.



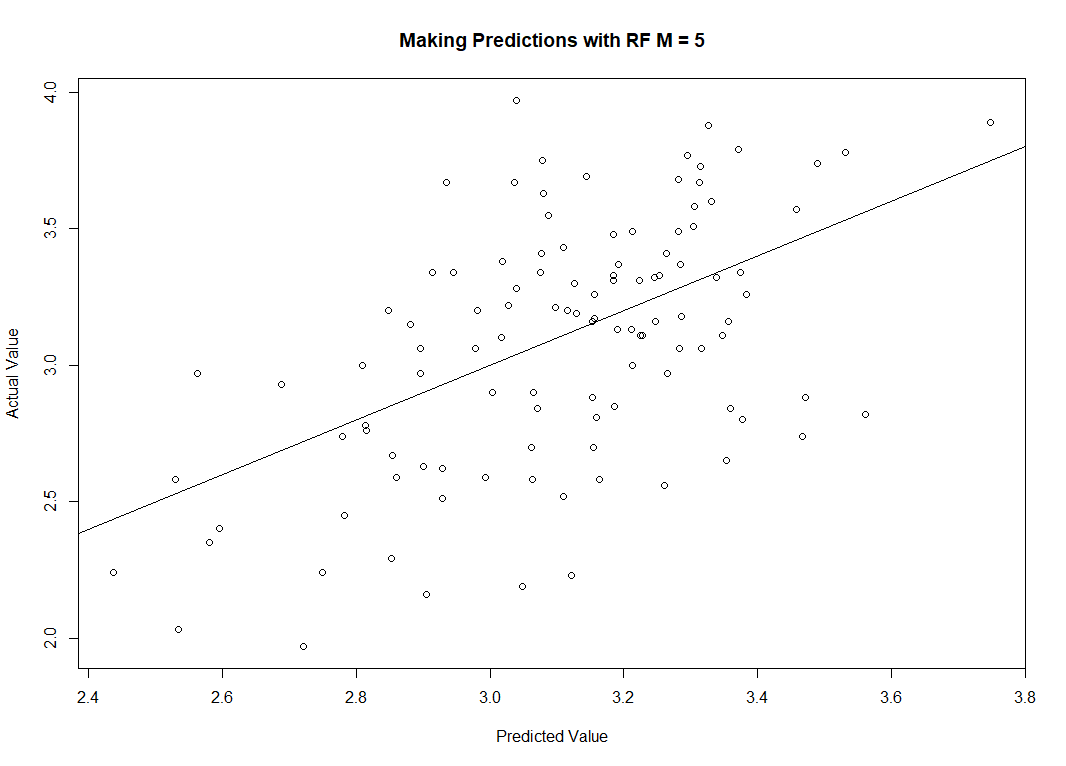
The next step in our analysis was to create a new model for predictions using bagging. When we created the model, we found that the test MSE = 0.142. The plot of actual values vs. predicted values has been included below.



Next we created a random forest model with m = 3. We found that test MSE = 0.148 for this model. A plot of actual values versus predicted values has been included below.



We also created a RF model with M = 5. The test MSE = 0.141 for this model. The plot of actual values versus predicted values follows.

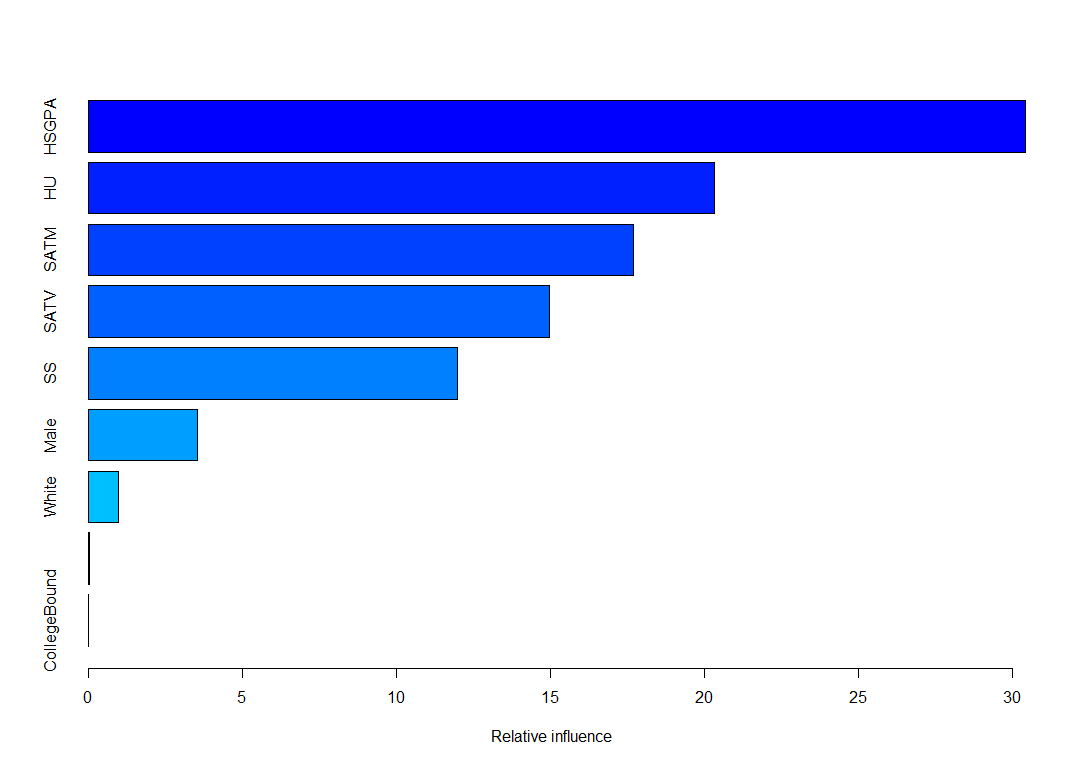


After creating different models using pruning, bagging, and random forests, we found that the random forest model with m = 5 performed the best because it had the lowest test MSE. After creating all of the models, we looked at the importance of each variable and found the following (note that the actual values within the table varied a little bit for each different model, but the order of importance didn’t change).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HSGPA | SATV | SATM | Male | HU | SS | FirstGen | White | CollegeBound |
| %MSE Increase | 11.74 | 5.77 | 2.33 | 5.77 | 9.43 | 0.27 | -0.56 | 6.44 | -2.61 |
| Included Node Purity | 5.70 | 4.10 | 3.45 | 0.77 | 5.53 | 2.04 | 0.21 | 0.86 | 0.19 |

We found that the three most important variables were HSGPA, HU, and SATV.

The final step of our analysis is to create a model which includes boosting. When we did this, we used 5000 trees. The output we obtained from this model is included below.



The output above is very similar to what we found when we looked at the importance of the variables in our previous step. The only difference is that the boosted model shows that SATM is more important or influential than SATV whereas the previous models had SATV as more important. Overall it still very close.

Overall, I find it interesting that HSGPA seems to be the most important/significant predictor in this lab because in our lab from last quarter it was also a significant variable. One difference I noticed in this lab when comparing to last quarter’s lab was that the variable “White” didn’t seem to be an important predictor in this lab, while in the last time we worked with this data it was.