ASSIGNMENT 2 REPORT

**CS 4372.501**

**Computational Methods for Data Scientists**

**Names:**

Jeremy Pacheco: [jsp200000@utdallas.edu](mailto:jsp200000@utdallas.edu)

Rhed Santiago: [rts170002@utdallas.edu](mailto:rts170002@utdallas.edu)

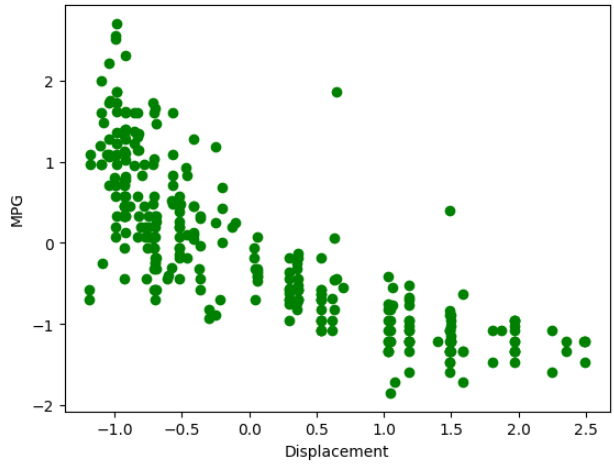
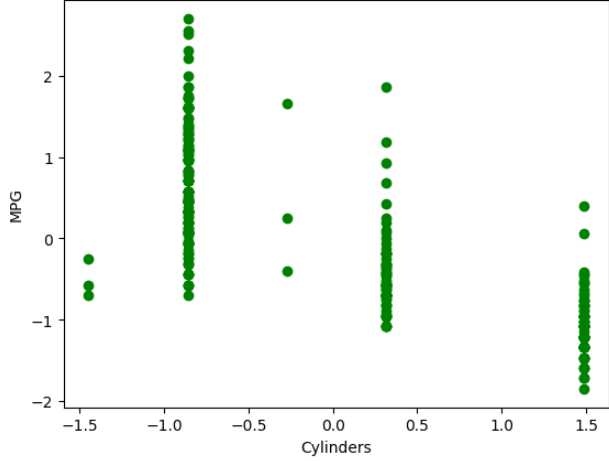
**Number of Free Late Days Used:** **1**

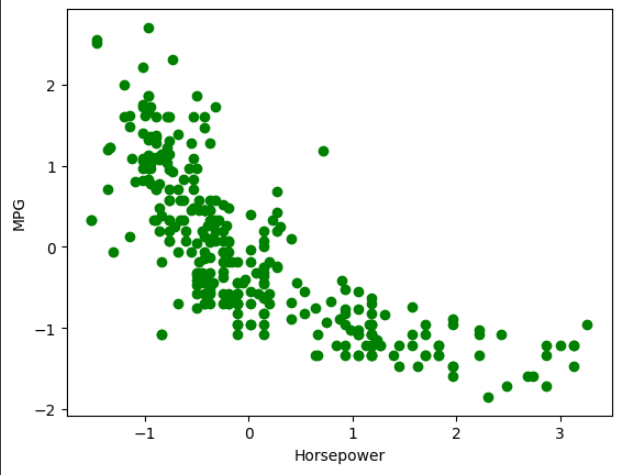
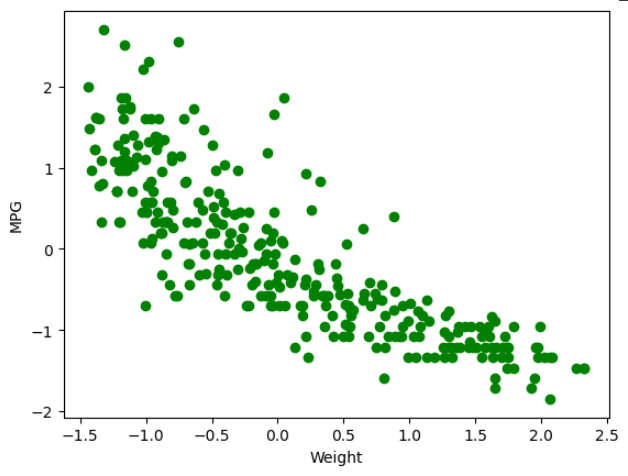
**Sources:**

<https://scikit-learn.org/>

<https://xgboost.readthedocs.io/en/stable/python/python_api.html>

**PLOTS**





**Cylinders v. MPG**

The plot shows that Cylinders and MPG are inversely correlated with one another. As a side-note, we can also observe that Cylinders is a bit more of a discrete attribute.

**DIsplacement v. MPG**

The plot shows that Displacement and MPG are inversely correlated with one another.

**Horsepower v. MPG**

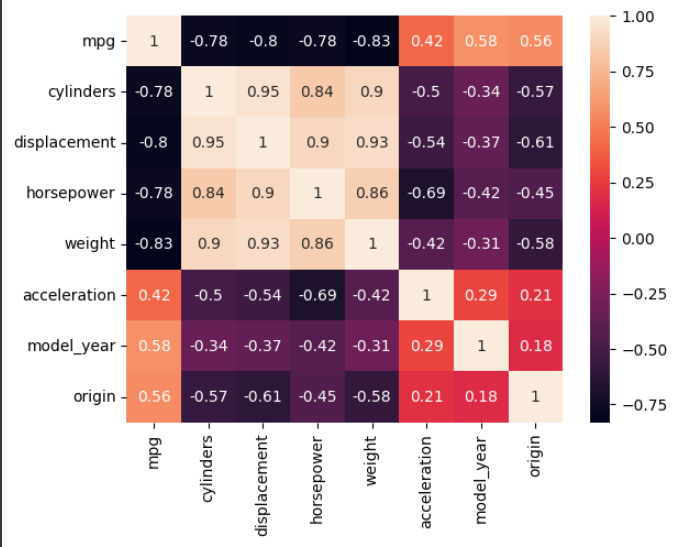
The plot shows that Horsepower and MPG are inversely correlated with one another.

**Weight v. MPG**

The plot shows that Weight and MPG are inversely correlated with one another.

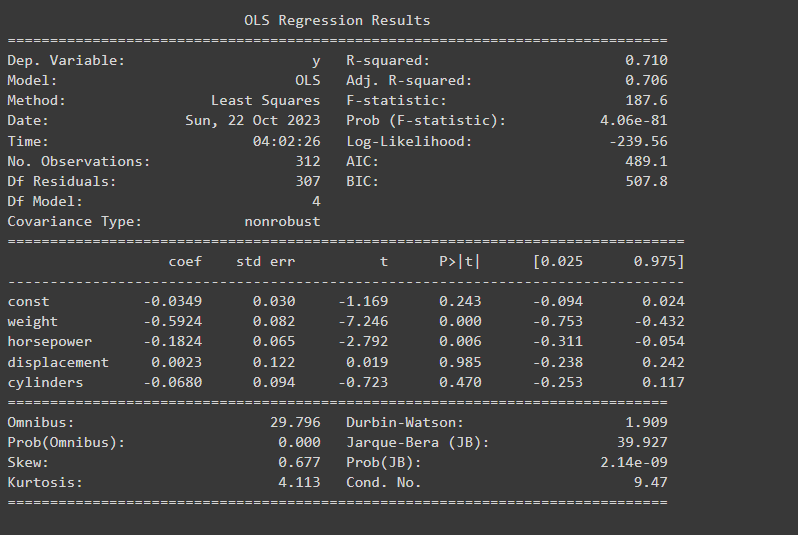
Overall, it is quite evident that there is strong correlation with these four attributes, individually with MPG.

**HEATMAP SUMMARY**



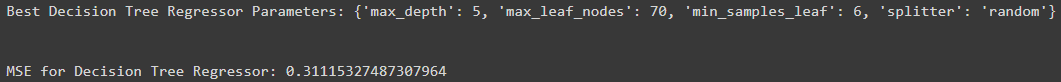
The heatmap tells us that cylinders, displacement, horsepower, and weight have a stronger correlation to mpg compared to the other options. Because of this, we ended up choosing these 4 to be our feature variables.

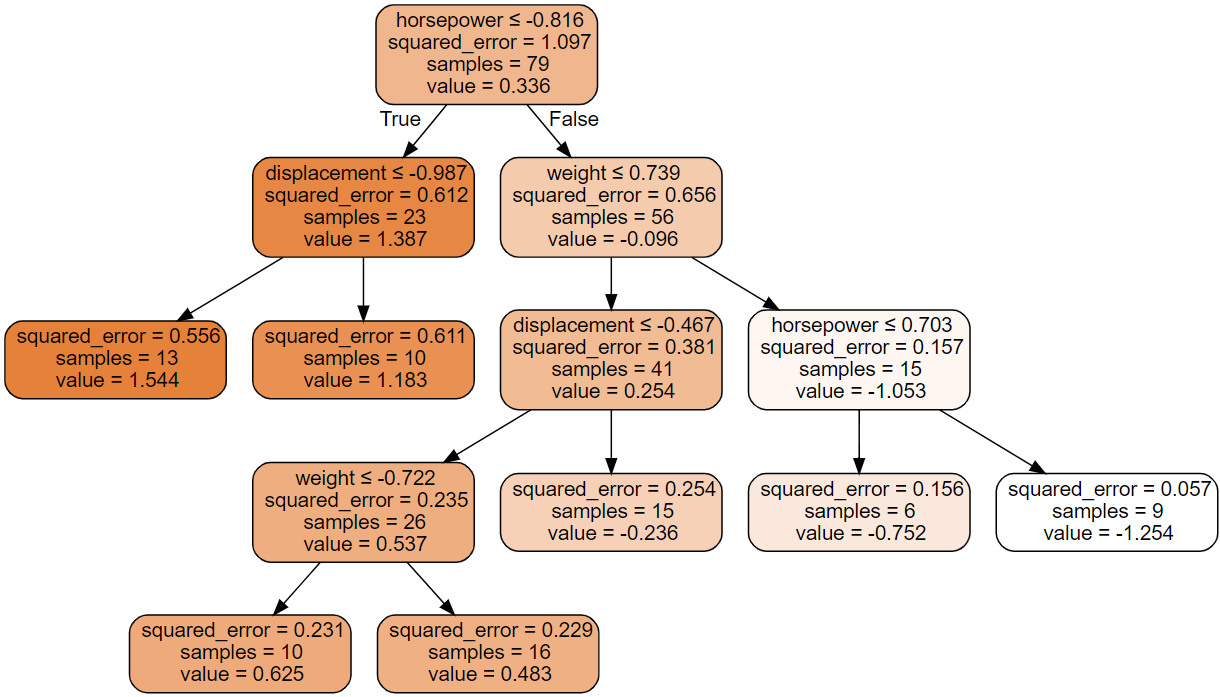
**OLS SUMMARY**



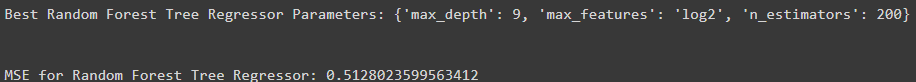
The OLS Summary tells us that our R-squared value is 0.710, which implies that the features we have correlate well with our y value for the output. The F-statistic also tells us the significance of our model, and we arrived at a decent value of 187.6. The coefficients tell us the weights of each feature, and numbers larger than others means they have a more significant impact.

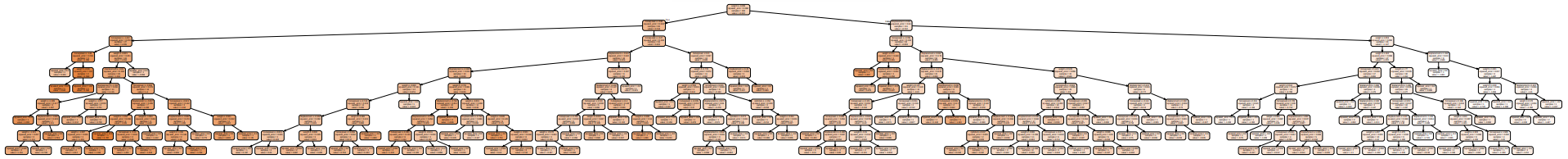
**DECISION TREE REGRESSOR**





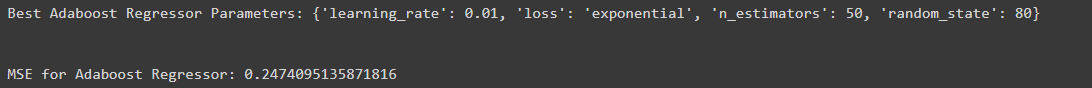
The best weights are shown in the picture, and I believe they are because we get a very good MSE output. Also, we can see that the squared error decreases nicely as the tree reaches the leaves. This also tells us that horsepower is the best splitting feature.

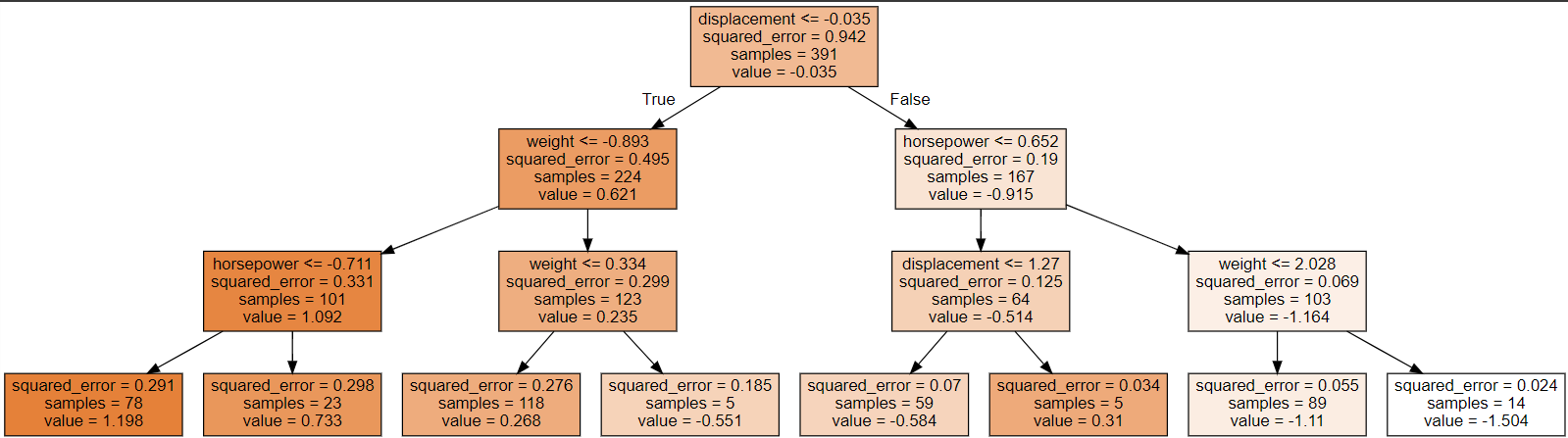
**RANDOM FOREST REGRESSOR**

****

The best weights are shown in the picture,and they are for this tree because we get a decent MSE value. The Random Forest Tree shown has its squared error decreasing as well, the deeper it reaches the children.

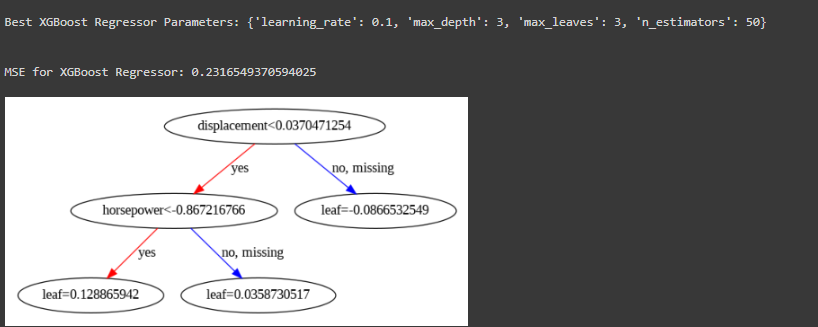
**ADABOOST REGRESSOR**

****

****

The best parameters are shown in the picture, and we get a very good value for MSE, as it is a really low number. We can also see in the predicted values, the leaves, that the squared error is very, very low, meaning that we used really good parameters for this.

**XGBOOST REGRESSOR**

****

The best parameters are shown in the picture above, and we also get a really nice value for MSE, as it is one very close to 0. We can see here that displacement is the best splitting node for this case, and for the leaf, we end up with a really low value for error.

**Overall**, we believe that AdaBoost seemed to be the best tree in our case because the leaf nodes have really good values for error, and the tree seems to split nicely. Lastly, we also have a really low value for MSE, and we want one that is close to 0. Ours was less than 0.25, which is a really good sign for our model.