Project 1- ID5059

Model Summary

Firstly, I did some data cleaning by removing NA rows of important columns. Then, filtered out data with price less than 100 and over 100K, because they were either fake adverts or outliers.

I first ran a basic linear model to check the covariates that were important. I used state, odometer, year and type as my covariates. I used Analysis of Variance (ANOVA) tests and used the R-squared values as a measure to check select these covariates were explaining more data and were significant. In other words, I used it for some feature selection. I took out some other covariates using ANOVA which were significant in predicting price to use in the tree model. Although this might not be correct, as p-values can't be trusted when assumptions of linearity aren't met, but it gives us some direction. This model gave a 30% R-squared value, means it explained 30% of the data. This is because the data set is too large and the assumption that there is linearity between the covariates and response doesn't hold. Also, it violates other assumptions such as non-constant variance and normality. It also gives negative predictions which is also a cause of concern.

To this end, we needed a non-linear component added to our model. I used b-splines with degree 5 for the continuous covariates. After checking the R-squared values for degrees 1 to 10. This model was explaining the data about 53%. Although, there isn't huge deviance from assumptions of the linear model, there is still the issue of negative predictions. This can be alleviated using Poisson models, which don't have negative predictions.

Finally, I made 2 Tree regression models, one with manufacturer covariate and the other with model replacing it. I used type, condition and odometer as other covariates. The first model after some tuning of **"cp"** parameter and **Max depth,** I was able to achieve 65% R-squared value. The 2nd model, with model covariate replacing the manufacturer covariate, the model showed extremely improved performance, 77%. The reason I was reluctant in using this covariate in other models because it increases the complexity of the model greatly, takes a long time to run the model and difficult to explain the tree.