Adam Brudz-Rodríguez

Danielle Senechal

May 14, 2020

Final Project MAT 374

Dr. Larose

Danielle and I originally had intended to use the data from her senior research project for her computer science major and expand on it to meet the criteria of this project. Early on we realized this would not be feasible because the data was not categorical. Instead of having something like age, location, gender, each column was a pixel of an image of a skin lesion with the last column identifying the kind of lesion it was. The data also had no predictor variables because the data set is meant for image recognition which is not compatible with this project.

Based on that realization we did our final project on the correlation between a car’s year, miles per gallon, engine fuel type, horsepower and its manufacturer suggested retail price (MSRP). We are ultimately trying to see to what degree do those variables effect a car’s MSRP. We also look at other variables such as make, model, year engine fuel, horsepower, how many cylinders, transmission type, size, style, and MPG, among others. One issue that we ran into early on was the amount of data we had, particularly with make and model. Our original data included makes not sold in the United States (i.e Renault) and makes sold in the United States but are not popular or realistic to the average buyer. Due to this when we first went through it our graphs and other visuals were overloaded with data and information that if someone looked at it, it would most likely be overwhelming. To deal with this we removed a lot of makes and their respective models to help alleviate the problem.

With that being said we do have some outliers however these are more “realistic” in perspective compared to the outliers we would have had if we kept the data as is. Some of these outliers include Rolls-Royce, Bentley, Lamborghini, and Ferrari. The outliers in this case are ultimately high-end exotic cars as well as Tier I luxury cars. This demonstrated in the below graph with the high-end exotic and Tier I luxury cars having a much lower frequency than other mainstream and premium/luxury makes. A picture containing fence

Description automatically generated

When it comes to the data we removed 715 entires of duplicated data and identified no NAs since there is no need to remove them because they aren’t being used as predictors. Outliers in the year and MSRP plot are visible, and in our bar plots outliers can be seen as previously shown above. When it came down to transforming and standardizing the data we did so by utilizing z-score standardiation. Number of doors also had to be reclassified into categories instead of numerical. For our partition we did 80% training and 20% testing which was later validated on the MSRP, make, horsepower, and style variables.

A screenshot of a map

Description automatically generatedWe lastly modeled our data using a decision tree for MSRP using year, horsepower, highway MPG, fuel type, and vehicle size which is shown below.

The following is a key to understanding the levels on the Classification Tree. Since the data was standardized using z-scores, the decision tree displays the values as the standardized data. The categories for MSRP correspond to: [-1.581, -0.682] = $1,995-$18,645, [-0.682, -0.271] = $18,645-$26,257, [-0.271, 0.104] = $26,257-33,202, [0.104, 0.659] = $33,202-$43,481, and [0.659, 3.708] = $43,481-$99,950. In regards to the two year branches, they correspond to: Year < -1.4 = Years before 2000, and Year < 0.4 = Years before 2014. Lastly for the three horsepower branches, they correspond to: Engine.HP < -0.2 = 220 horsepower, Engine.HP < 0.92 = 156 horsepower, and Engine.HP < 0.83 = 311 horsepower. Engine.Fuel.Type is for the type of fuel required by the vehicle, and the decision tree uses flex-fuel and regualr unleaded as deciding factors.

In our personal opinions we believe that our models represent that MPG, year, horsepower and engine fuel type are predictors are do effect a car’s manufacturer suggested retail price.