**Anirudh Vunnam**

**06/07/2022**

**Introduction**

For the price of cars using the available independent variables, various data science techniques must be used. This should aid management in determining how prices vary in relation to the independent factors. They can then adjust the car's design, commercial strategy, and other factors to fulfil specified price targets. Here we will be trying answer the following two important questions:

Which variables are important in forecasting a car's price?

How well those variables accurately describe a car's pricing?

**Data Cleaning**

The following is a summary of the data: 205 rows, 26 columns, and no null values. The target variable is "Price," whereas the other variables are independent variables. Categorical and numerical variables are separated again among the independent variables.

Numerical variables: (wheelbase, carlength, carwidth, carheight, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg)

Categorical variables: (symboling, fueltype, aspiration, doornumber, carbody, drivewheel, enginelocation, enginetype, cylindernumber, fuelsystem, car\_name)

Variables are also in the correct format, except for "symboling," which should be a categorical variable (so that dummy variable are created for the categories). We also pre-processed the data in the field "CarName" and added a new variable named "car company." vw and vokswagen should be written volkswagen, porcshce should be porsche, toyouta should be toyota, Nissan should be nissan, maxda should be mazda, and so on.

**EDA**

The heatmap reveals several interesting facts:

Correlation of the dependent variable "Price" with the target variable "Price":

* Wheelbase, carlength, carwidth, curbweight, enginesize, horsepower (note how all these variables describe the size/weight/engine power of the car) are all substantially (positively) associated with price.

Treemap chart

Description automatically generated

* 'Citympg' and 'highwaympg' are adversely associated with price (-0.70 approximately). This suggests that automobiles with high mileage fall into the 'economy' car category and are therefore less expensive (think Maruti Alto/Swift-type cars, which are designed to be affordable for the middle class, who value mileage over horsepower/car size, etc.).
* Many independent variables are significantly connected (see top-left section of matrix): wheelbase, carlength, curbweight, enginesize, and other'size/weight' measurements are all positively correlated.

**Analysis**

The (numeric) target variable must be linearly connected to at least one other numeric variable in order to execute linear regression. The model with all the variables has an r-squared of 91.88%

The model with 15 variables has an r-squared of 77.6 percent, although this is based on training data. 75.2 is the adjusted r-squared. RFE with 6 features is giving about 65.4% r-squared while the adjusted r-squared is 63.8%.

As the number of features are reduced the r-square value or the accuracy of the model also reduces hence it shows the importance of all the variables.

**Conclusion**

Despite the fact that this is the simplest model we've ever created, the final predictors appear to have significant correlations. Some of these elements can be removed, but this will have a major impact on the adjusted-r2 score.

One can select the number of features based on the r2 score they desire. There are a couple drawbacks to this method, and there are more advanced ways for determining the appropriate number of features.

**Reference:**

Canvas. (2022). Retrieved June 7, 2022, from https://canvas.northeastern.edu/ <https://northeastern.instructure.com/courses/110240/assignments/1345639>

**Appendix**

A picture containing text, window

Description automatically generated

Various properties of the cars

Chart, line chart

Description automatically generated

Adjusted r-square against all features

Chart, box and whisker chart

Description automatically generated

Boxplot of price vs car company