Project Proposal: Predicting Fuel economy and vehicle origin

Name	Email	Contributions
Aditi Goyal	adigoyal@ucdavis.edu	Exploratory data analysis ,data visualization, data transformation, report writing,and web app.
Brandon Hom	bwhom@ucdavis.edu	Regression analysis, model tuning, interpretation of regression results,machine learning models(logistic and linear regression),and web app.
Tammie Tam	tastam@ucdavis.edu	Regression diagnostics (i.e residual plots), conclusion, web app, report writing,and interpretation of regression diagnostics

Introduction/Background

Fuel economy is defined as how much a car can travel per volume of fuel, which is usually measured as $\frac{miles}{gallon}(mpg)$. Cars that have a lower value for mpg are considered to have good fuel economy and vice versa. Logically, one would prefer to have a car with good fuel economy, since money would not have to be frequently expended on gas. This also ties into global warming. Cars with poor fuel economy end up contributing to the global warming problem, as more gas is used to cover travel certain distances [1]. Of course individuals still need vehicles to get them to their destinations, but the effects of cars on global warming can be mitigated by using cars with good fuel economy instead. By being able to predict a car's fuel economy based on a set of features, one can make a more informed decision when purchasing a car, a decision that can have a positive impact on both spending and global warming.

Dataset Description

The data set on fuel economy was obtained from kaggle: https://www.kaggle.com/uciml/autompg-dataset. Although it is data from the late 1900s, the features within the data set can provide us insight on what variables have impact on fuel economy. The dimensions of the data set are 398 by 9, meaning that we have 9 features within our data set. `Model year` and `car name` are not useful for our purposes, so they will be dropped during the analysis. `Weight`, `acceleration`, `horsepower`, `displacement`, `cylinders` are all numerical features, while `origin` is categorical. `Weight`, `acceleration`, and `horsepower` are self-explanatory. `Cylinders` are indicative of the power of an engine, where more cylinders means more power but more consumption of gas. `Displacement` refers to how much volume of air and fuel moved through the cylinders of the engine.

Key questions

- What numerical variables have the most impact on mpg? For example, if increasing weight contributes the most to MPG, individuals should be wary about purchasing heavy cars, since it will lead to more consumption of fuel.
- The origin column has Europe, Japan and USA encoded. Do cars from these regions have similar or different fuel economy? If so, can we use the features of this dataset to classify a car as coming from

one of these regions? This can especially be useful if a customer is deciding on buying a car from any of these regions but, for instance, only knows of the weight, horsepower and acceleration. For example, if the average mpg of USA cars is 18, this can be used as a second estimate of the mpg.

Methodologies

The model that seems appropriate for predicting mpg is linear regression model; linear regression models can be useful when predicting a numerical variable. We suspect that as weight, acceleration and horsepower increase, mpg will also increase in a relatively linear fashion. To test this, we plan on using a multiple linear regression model, where the coefficients of the model will indicate the significance of that variable. For example, if we the coefficient for weight and acceleration is 5 and 6 respectively, this informs us that acceleration has a greater impact than weight.

To assess the appropriateness of a linear regression model, we will also perform diagnostics. Plots of the residuals vs fitted values and residuals vs predictor variables will be made to check for constant variance. The normality assumption will be checked with a QQplot and histogram of the residuals.

Finally, there may also be multicollinearity present within our data. To address this problem, we plan on using forward and backward stepwise regression to tune our model by removing predictor variables that do not any significant change to the insurance costs.

To build a more general predictive model, we plan on taking a machine learning approach using R's `caret` package. That is we will split the data into training and test set and build a linear regression and logistic regression model using the training set. For the linear regression model, we plan on using only the impactful features found during the regression analysis involving the entire dataset. The linear regression model can be evaluated by computing the `RMSE` value for both training and testing; if the RMSE is high for a linear model, we may try fitting other models such as polynomial regression.

We plan on building three logistic regression models, since we have 3 categories and logistic regression is a binary classifier. We can then evaluate our model with a confusion matrix to see how well these models perform. If, for example, the accuracy is poor when USA is being distinguished from Japan and Europe in the two models, this will indicate that the cars are too similar.

References

1. Transportation Technologies and Innovation. Union of Concerned Scientists. (n.d.). Retrieved November 12, 2021, from https://www.ucsusa.org/transportation/technologies.
2. McGregor, H. V., Gergis, J., Abram, N. J., & Phipps, S. J. (2016). The Industrial Revolution kick-started global warming much earlier than we realised.
3. Learning, U. C. I. M. (2017, July 2). Auto-mpg dataset. Kaggle. Retrieved November 12, 2021, from https://www.kaggle.com/uciml/autompg-dataset.