Power and drive train specifications impact on fuel economy

Braden G. Millard

Northwest Missouri State University, Maryville MO 64468, USA

Abstract. **Keywords:** fuel economy · power train · drive train · transmission ·

1 Introduction

There are many things to consider when buying a new or used vehicle. With the rising cost of fuel and oil, maximization of the dollar is a priority for many [6]. For many others however decreasing their personal carbon foot print is equally or of greater priority[2]. In this analysis I will be using data collected by the United State Department of Energy regarding vehicle specifications and miles per gallon. Three different machine learning models will be built to predict miles per gallon when given attributes. The models that will be built is a multiple linear regression, lasso regression and a random forest model.

1.1 Goals of this Analysis

The goal of this analysis is to find the factors that greatly affect the gas mileage of cars and build machine learning models to predict miles per gallon.

2 Methodology

The steps used in this analysis were broken down into five phases. They are 1)Data collection 2)Data Cleaning 3)Data Exploration 4)Model Training, and 5)Results.

2.1 Data Collection

The data set for this project was collected via the U.S. Department of Energy's website [3]. The data set used is the "vehicle" data set. This data set contains all of the fuel economy data for vehicles built between 1984 and 2023. In total this data set includes one hundred and forty one car brands and four thousand six hundred twelve models. For a total of forty five thousand twenty six total records.

2.2 Data Cleaning

The data for this project was published data, therefore Microsoft Excel was used. A majority of the data cleaning was removing any unnecessary attributes. The data set was shrunk down from eighty three attributes to nine attributes. The attributes kept are as follows:

MPG - combined MPG
cylinders - number of cylinders
displ - volume of engine
drive - AWD, RWD, FWD, 4WD
fuleType - fuel type
make - brand of the car
model - model of the car
trany - type of transmission
Year - year the car was made

For this analysis combined MPG will be the dependent variable. Cylinders, displacement, drive, fuel type, transmission, and year will be the independent variables. All text fields (drive, fuel type and trany) were converted into dummy attributes containing Boolean variables for ease of analysis.

2.3 Data Exploration

Exploratory data analysis is used to understand the cleaned data set. For this project pivot tables, line graphs, and basic statistics (min, max, mean).After exploratory analysis it was found that there was a trend to the relationship between mpg and both the number of cylinders and engine displacement. It was also found that the average diesel engines average four miles per gallon, or twenty percent higher then it's gas counterparts. It was also observed that after the year 2006 there has been a positive increase in the average mpg per year.

There were some assumptions made throughout the data collection and cleaning process. They are as follows:

Automatic transmissions with different gears all affect mpg the same
Manual transmissions with different gears all affect mpg the same
All grades of gasoline affect mpg the same

2.4 Model Building

Three models were built to predict the expected Miles Per Gallon when given the input data. Python was used to build the pipeline [5]. The clean data was exported to a CSV and then inputted into the model. The data was then split to create a training and testing data set. Eighty percent of the data was put into the training data and the remaining twenty were set aside in the test data set. Within the pipeline three different analysis methods were used. The training

data was used to create a multiple linear regression model, lasso regression and a random forest.

The multiple linear regression model, lasso regression and the random forest model were built using the Scikit-learn machine learning library[4]. The results of these models were then cross validated using Scikit-learn's GridSearchCV[1]. Due to the size of the input data the results were split into three groups per model.

2.5 Data validation

To see how accurate our models were they needed to be validated. To do this we used our fit models and the testing data set we created earlier in the pipeline. The independent variables in the test data set were ran trough the three models. The results of each run were then compared to the recorded miles per gallon. Scikit-learn's metrics mean absolute error were used for this step. The output is the mean absolute error for each model. The results of this analysis were displayed by mean absolute error, displayed below in Table 1.

Table 1. Mean Absolute Errors

Model	Mean Absolute Error
Multiple Linear Regression	2.046 mpg
Lasso Regression	2.388 mpg
Random Forrest	1.384 mpg

3 Results and Analysis

The goal of this analysis was to see the impact different drive train and power train features have on gas mileage and build a model to help predict gas mileage. After cleaning the data set, all of the attributes were compared to miles per gallon. The results are below in Figure 1.

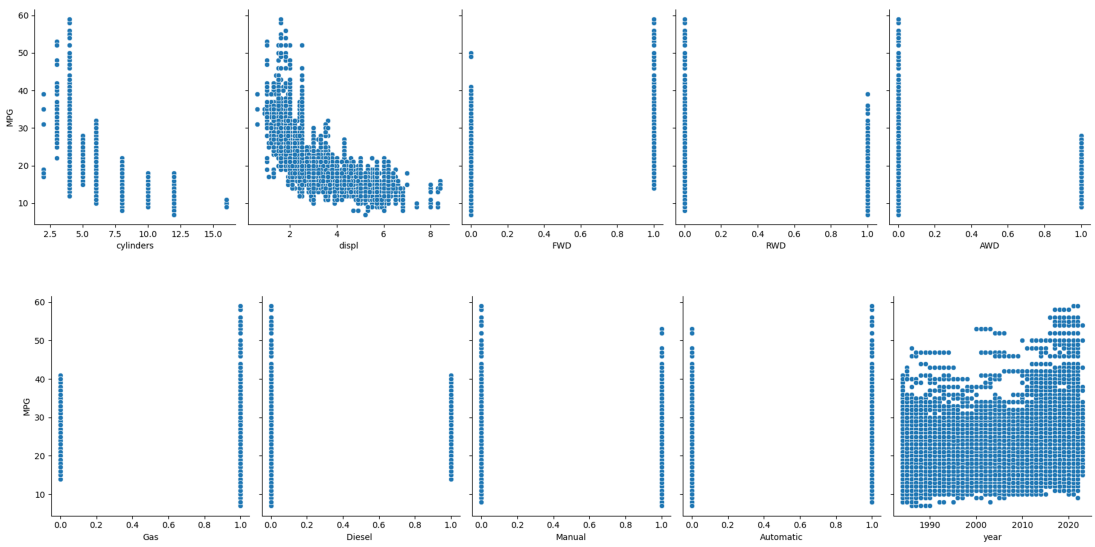

Figure 1. Attributes compared to MPG

At first glance it is observed that both the number of cylinders and the engine displacement have similar trends. With the higher displacement and more cylinders being detrimental to the total fuel economy of the car. It is also observed that newer vehicles tend to perform better then older cars.

Three models were then built to analyze the data, multiple linear regression, lasso regression and a random forest model. When looking at the multiple linear regression results there were some interesting findings. It was found that weather the car was a manual or an automatic was statistically insignificant as the p value found was higher then .05. The Adjusted R-squared Value for this model was .69 so we can conclude that this model accounts for 69 percent of what effects miles per gallon. Which can help explain the model mean absolute error of 2.046 mpg. The lasso regression model performed the worst of the three, with a high mean absolute error of 2.388 mpg and a low R-squared Value of .58. Overall the Random Forest model performed the best, with a mean absolute error of 1.384 mpg and an overall R-squared Value of .84. Therefore if we wanted to predict the fuel economy of a vehicle then the attributes should be passed through the trained random forest model.

4 Conclusions and Future Work

In this report we analyzed the impact of different drive train and power train features on fuel economy. The results showed a correlation between engine displacement, number of cylinders and type of drive train compared to miles per gallon. Three different machine learning models were used to predict the miles

per gallon when given attributes. The random forest model performed the best with an MAE of 1.384 mpg and an R-squared Value of .84. One major limitation of this study is that other major characteristics were not present in the data ie horsepower, gear ratios, torque. For this data to be present it would need to be scraped and then paired with the EPA data. The EPA data for this project is updated yearly so a annual review of this analysis is feasible.

References

1. sklearn.model$_s$*election.gridsearchcv.scikit*(2022), *https* : *//scikit* − *learn.org/stable/modules/generated/sklearn.model*$_s$*election.GridSearchCV.html*
2. Alec Tyson, B.K., Funk, C.: Gen z, millennials stand out for climate change activism, social media engagement with issue (2021), https://www.pewresearch.org/science/2021/05/26/gen-z-millennials-stand-out-for-climate-change-activism-social-media-engagement-with-issue/
3. EPA: Fuel economy data. Office of Energy Efficiency and Renewable Energy (2022), https://www.fueleconomy.gov/feg/download.shtml
4. scikit learn: Machine learning in python (2022), https://scikit-learn.org/stable/
5. Millard, B.: Analysis files and code. GitHub (2022), https://github.com/bradenm97/NWMSUMastersCapstone
6. Stevens, P.: Rising fuel costs are a massive problem for business and consumers (2022), https://www.cnbc.com/2022/05/19/fuel-is-a-problem-for-business-and-consumers-why-prices-are-so-high.html