Technical Report - Final project, Data Science 5k

Topic – A study in fuel efficiency and CO2 emissions as well as building machine learning models to predict fuel efficiency

Problem - Vehicle MPG varies by make/model.

EPA and other agencies spend large sums of money and time conducting fuel efficiency tests.

Hypothesis – Vehicle MPG varies by make/model (alternately MPG varies by engine displacement/transmission type) and a machine learning model can be built to predict this.

Replacing costly and time consuming tests with a machine learning model would certainly benefit agencies such as the EPA.

Null Hypothesis – Vehicle MPG does not vary by make/model (MPG does not vary by engine displacement/transmission type) or a machine learning model cannot replace laboratory testing for fuel efficiency.

Prediction task – Predict the miles per gallon based on several variables.

Tasks:

1) Evaluate the datasets, sort and clean them.
2) Analyze pertinent features for inclusion in machine learning model
3) Continue to run tests on various features till best fit model is identified.

Goals and success metrics:

Goal – identify the best model for predicting MPG based on make/model (engine displacement/transmission type). Success metrics – test to find best fit model
Risks and limitations:

Risks – Making huge assumptions of MPG without full scope of variables could lead to a skewed model.

Limitation – lack of access to underlying factors that affect fuel consumption.

Datasets:

Vehicle MPG database (vehicles.csv): shape (40704 x 83) columns: Too many to list here (many irrelevant to this analysis and therefore will be cleaned during data wrangling)

This series of notebooks walks through the steps for data cleaning exploratory data analysis machine learning. The data used is fuel efficiency data from the EPA and spans 33 years (1985-2018). The data itself contains over 40,000 observations and 83 features.

The folder structure for this project is as follows

1) data_sets: contains all data used in this project
2) Presentation: houses the final presentation and associated image files
3) PythonNotebooks: contains all notebooks used in this project

The python notebooks are as follows:

1) Data_Cleaning - basic data cleaning
2) EDA - first exploratory data notebook
3) EDA2 - second EDA notebook
4) Linear_Regression - a notebook for machine learning
5) User_Pred - a notebook with a final machine learning model for users to make MPG predictions based on input

Initial Findings from Exploratory Data Analysis

It became clear quite quickly that there were many features that were completely irrelevant in determining fuel efficiency.

Conversly, it also became clear that CO2 emissions are nearly perfectly inversly correlated to fuel efficiency. Basically, this means that using CO2 as a feature for predicting MPG lead to an overfit model 100% of the time.

Using just CO2 emissions as the feature, I was able to acheive 90% accuracy.

Though you will see those models remaining in my Linear Regression workbook, you'll not that for my deployed model, I changed my approach.

For the final model, I chose to drop CO2 as a feature from the model. I also chose to pre-filter the data using year.

This helped ensure that comparisons were made against like types. The assumption here being that a vehicle produced in 2018 is technologically far superior to a vehicle produced in 1985. This technological advantage would be seen in higher MPG.

As it turned out, this was a very prudent approach and allowed me to achieve realistic predictions of MPG based on four key variables: Displacement Cylinders Vehicle Class Fuel Type

More fine tuning can be done. A random forest regression can be attempted. These are growth areas for future development.

Also of note, the model feature needed (and still needs) extensive cleaning. I took some effort to clean the model feature, but more work might have resulted in model having a bigger impact. The biggest issue with the model feature were the large number of 'dirty' entries leading to several instances of the same model being represented separately. As an example, there as F-150, F-150 2WD, F-150 4WD, F-150 Super Crew, F-150 Crew Cab, etc... Being able to collapse all these down into a single model is the goal.

I only partially achieved that goal due to time constraints.

Going forward, in the hopes of building the best model possible, it would be advisable to find a way of incorporating other key variables that were not available in this dataset.

Specifically, horsepower and gross vehicle weight would have been good to have. But, the list of unavailable variables, which may impact MPG is quite large.