Data Analysis with Fuel Economy Data

Sai Sahiti Gudla
Department of Computer Science and Engineering
PES University

Electronic City
Bangalore

SRN:PES2201800481
Email: sahiti7705@gmail.com

Github:

https://github.com/saisahitigudla/Data-Analytics-Project

Deepika K
Department of Computer Science and Engineering
PES University

Electronic City
Bangalore

SRN:PES2201800659
Email: deepikak01022001@gmail.com

Github:

https://github.com/deepika02012001/Data-Analytics-project

Ajay
Department of Computer Science and Engineering
PES University

Electronic City
Bangalore

SRN:PES2201800724
Email: ajay5uu.ak@gmail.com

Github:

https://github.com/Ajay820/datascienc

***Abstract*—This study evaluates methods of machine learning and statistical analysis for predicting fuel consumption in vehicles through fuel economy data from EPA. This is useful since reduced fuel consumption means reduced environmental impact as well as reduced fuel costs. Through this dataset we can answer questions regarding the improvement in fuel economy over a period of time and the model of vehicles that have improved by using alternative sources of fuel.**

I. Introduction

Fuel economy is a measure of how far a vehicle will travel with a gallon of fuel; it is expressed in miles per gallon. This is a popular measure used for a long time by consumers in the United States; it is used also by vehicle manufacturers and regulators, mostly to communicate with the public. As a metric, fuel economy actually measures distance traveled per unit of fuel.

Fuel consumption is the inverse of fuel economy. It is the amount of fuel consumed in driving a given distance. It is measured in the United States in gallons per 100 miles, and in liters per 100 kilometers in Europe and elsewhere throughout the world. Fuel consumption is a fundamental engineering measure that is directly related to fuel consumed per 100 miles and is useful because it can be employed as a direct measure of volumetric fuel savings. It is actually fuel consumption that is used in the CAFE standard to calculate the fleet average fuel economy (the sales weighted average) for the city and highway cycles. Fuel consumption is also the appropriate metric for determining the yearly fuel savings if one goes from a vehicle with a given fuel consumption to one with a lower fuel consumption.

The regulation of vehicle fuel economy requires a reproducible test standard. The test currently uses a driving cycle or test schedule originally developed for emissions regulation, which simulated urban-commute driving in Los Angeles in the late 1960s and the early 1970s. This cycle is variously referred to as the LA-4, the urban dynamometer driving schedule (UDDS), and the city cycle. The U.S. Environmental Protection Agency (EPA) later added a second cycle to better capture somewhat higher-speed driving: this cycle is known as the highway fuel economy test (HWFET) driving schedule, or the highway cycle. The combination of these two test cycles (weighted using a 55 percent city cycle and 45 percent highway cycle split) is known as the Federal Test Procedure (FTP). This report focuses on fuel consumption data that reflect legal compliance with the CAFE requirements and thus do not include EPA's adjustments for its labeling program, as described below. Also discussed below are some technologies—such as those that reduce air- conditioning power demands or requirements—that improve on-road fuel economy but are not directly captured in the FTP.

Compliance with the NHTSA's CAFE regulation depends on the city and highway vehicle dynamometer tests developed and conducted by the EPA for its exhaust emission regulatory program. The results of the two tests are combined (harmonic mean) with a weighting of 55 percent city and 45 percent highway driving. Manufacturers self-certify their vehicles using preproduction prototypes representative of classes of vehicles and engines. The EPA then conducts tests in its laboratories of 10 to 15 percent of the vehicles to verify what the manufacturers report. For its labeling program, the EPA adjusts the compliance values of fuel economy in an attempt to better reflect what vehicle owners actually experience. The certification tests yield fuel consumption (gallons per 100 miles) that is about 25 percent better (less than) EPA- estimated real-world fuel economy. Analysis of the 2009 EPA fuel economy data set for more than 1,000 vehicle models yields a model-averaged difference of about 30 percent.

The unfortunate consequence of the disparity between the official CAFE (and proposed greenhouse gas regulation) certification tests and how vehicles are driven in use is that manufacturers have a diminished incentive to design vehicles to deliver real-world improvements in fuel economy if such improvements are not captured by the official test. Some examples of vehicle design improvements that are not completely represented in the official CAFE test are more efficient air conditioning; cabin heat load reduction through heat-resistant glazing and heat-reflective paints; more efficient power steering; efficient engine and drive train operation at all speeds, accelerations, and road grades; and reduced drag to include the effect of wind. The certification tests give no incentive to provide information to the driver that would improve operational efficiency or to reward control strategies that compensate for driver characteristics that increase fuel consumption.

II. TABLE OF CONTENTS

1. Knowing the data attributes
2. Data Preprocessing
3. Data Visualization
4. Training and Testing Data
5. Developing the model
6. Prediction
7. Conclusions
8. References

III. DATASET

Fuel Economy Data: This information is provided by the US Environmental Protection Agency , Office of Mobile Sources, National Vehicle and Fuel Emissions Laboratory.

A. *EPA Fuel Economy Testing:* https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy

B. *DOE Fuel Economy Data:*

https://www.fueleconomy.gov/feg/download.shtml/
Green vehicle guide documentation in the EPA website best describes the dataset.

We use 2010 and 2020 datsets for comparison

IV. ATTRIBUTES

1. **Model year**
2. **Vehicle manufacturer name**
3. **Vehicle mfr code**- manufacturer code

Mfr Code	Manufacturer Name
20	DAIMLERCHRYSLER
30	FORD MOTOR COMPANY
40	GENERAL MOTORS
70	ASTON MARTIN
90	FIAT AUTO S.P.A.
108	ROVER GROUP LTD. (AR)
120	BMW
178	DAEWOO
190	DAIHATSU MOTOR COMPANY LTD.
196	MITSUBISHI MOTOR MANUF OF AMERIC
200	MERCEDES BENZ
220	FERRARI
230	FIAT
260	HONDA
265	HYUNDAI
290	ISUZU
305	JAGUAR CARS INC
338	KIA MOTORS CORPORATION
350	LOTUS
380	NISSAN
407	PANOZ AUTO-DEVELOPMENT CORP.
420	PORSCHE
440	ROLLS-ROYCE MOTOR CARS LTD.
460	LAND ROVER GROUP LTD.
470	SAAB
490	MITSUBISHI
491	MITSUBISHI MOTOR SALES AMERICA
492	MITSUBISHI MOTORS AUSTRALIA LTD
540	SUZUKI MOTOR CORPORATION
560	MAZDA MOTOR CORP.
570	TOYOTA
576	NEW UNITED MOTOR MFG INC
590	VOLKSWAGEN
600	VOLVO
640	AUDI
660	FUJI HEAVY IND – MAZDA
540	LAMBORGHINI

4. **Represented test vehicle make**
5. **Represented test vehicle model**
   bidx - basic engine index number identifying a unique basic engine or sub-basic engine (if engine is divided into two or more groups)
6. **test vehicle id -** manufacturer defined vehicle identification number within EPA's computer system (not a VIN number)
7. **vehicle type**- 'C' for passenger vehicle and 'T' for truck (classification of the tested vehicle)
8. **test vehicle configuaration # -** configuration number identifying a unique configuration within a vid
9. **test vehicle displacement(l)-**displacement of test vehicle
10. **actual tested testgroup**
11. **# of cylinders and rotors**
12. **Engine code**
13. **police -** indicator for police vehicle (Y or N)
14. **rhp -** rated horsepower
15. **police -** indicator for police vehicle (Y or N)
16. **rhp -** rated horsepower
17. **ec1 -** exhaust emission control system code
18. **ec2 -** exhaust emission control system code
19. **ec3 -** exhaust emission control system code
20. **ec4 -** exhaust emission control system code
21. **ec5 -** exhaust emission control system code
    Code Description
    002 - Engine modification
    005 - Thermal reactor
    008 - Exhaust recycle
    010 - Air pump
    011 - Pulsating air system
    016 - Oxidation catalyst
    017 - Reduction catalyst
    018 - Three-way catalyst
    019 - Closed loop control of air/fuel ratio
    020 - Three-way catalyst and closed loop control of air/fuel ratio
    021 - Closed loop air injection
    099 – Other
22. **evc** - evaporative emission control system code
    101- Crankcase
    102- Canister
    103- Tank
    104- None
    105- Canister and charcoal air cleaner
    199 - Other
23. **tested transmission type code -** transmission code
    C4 - Manual 4-Speed (Creeper) (M-4)

M3 - Manual Three-Speed
M4 - Manual Four-Speed (No Creeper)
M5 - Manual Five-Speed
SA - Semi-Automatic
A3 - Automatic 3-Speed (No Lockup)
L3 - Lock-Up/Automatic/3-Speed
A4 - Automatic 4-Speed (No Lockup)
L4 - Lock-Up/Automatic/4-Speed
C5 - Manual 5-Speed (Creeper) (M-5)
S2 - Semi-Automatic Two Speed
S3 - Semi-Automatic Three Speed
S4 - Semi-Automatic Four Speed
S5 - Semi-Automatic Five Speed
AV - Automatic Variable Gear Ratios
M6 - Manual Six Speed
A5 - Automatic 5-Speed (No Lockup)
L5 - Lock-Up/Automatic/5-Speed
C6 - Manual 6-Speed (Creeper) (M-6)
A6 - Automatic 6-Speed (No Lockup)
S6 - Semi-Automatic Six Speed

24. **tested transmission type**
25. **number of gears**
26. **transmission lockup**
27. **drv -** drive system code
    F - Front wheel drive
    R - Rear wheel drive
    14 - 4-wheel or all-wheel drive
28. **drive system description**
29. **transmission overdrive code-** overdrive code
    1. - No gear ratio <1
    2. - Top gear ratio <1
    3. - Electronically operated overdrive
    4. - Computer-controlled automatic electronic overdrive
    5. - Computer-controlled automatic electronic overdrive with lockout switch
30. **trans_desc -**Transmission descriptor (it is constructed according to the following logic:
    Transmission descriptors are constructed based on the data submitted on G2 record of the General Label subsystem
    If Shift Indicator Light is (Y) -SIL
    If Engine Management System is (Y or L)-EMS
    If Number of Transmission Mode is an x which is V, C or a number between 2 and 9 -xMODEM
    (If x is 1, this item is not to be used)
    If Variable Lockup Point is an x which is V, C or a number between 2 and 9 –xLKUP
    If Declutching/Freewheeling is Y or L-DC/FW
    Any combination of the above, with a blank in between and in the order as shown and not to exceed 15 columns,makes up a transmission descriptors field.
    If the last three (xMODE, xLKUP, DC/FW) are the only descriptors, it is to be displayed as xMODE xLKUP FW.
31. **etw -**equivalent test weight
32. **cmp -**compression ratio
33. **axle -**axle ratio
34. **n/v -**n/v ratio (engine speed versus vehicle speed at 50 mph)
35. **a/c-** indicates air conditioning simulation
36. **aftertreatment device cd**
37. **aftertreatment device description**
38. **sil** - shift indicator light use cd for standard transmissions (y = yes or n = no,, indicates if test was performed used the sil to upshift during the test driving cycle.)
39. **shift indicator light use description**
    prc-test procedure code
    2-CVS 75 & later (EPA city w/o canister loading)
    3-HWFE (highway test)
    21 -Fed fuel 2-day exhaust (C4H10 canister load)
    25 -Calif fuel 2-day exhaust (C4H10 canister load)
    31 -Federal fuel 3-day exhaust (C4H10 canister load)
    35 -Calif fuel 3-day exhaust (C4H10 canister load)
    41 -Federal fuel 2-day exhaust (heat fuel tank to load canister)
    45 -Calif fuel 2-day exhaust (heat to load) test
40. **tcp**-test purpose code
    01 - Emission data
    08 - Manufacturers' development
    31- Fuel economy
    32- Analytical fuel economy
41. **tnum**-test number, a unique identifier for a set of test data performed performed at the manufacturer or EPA test lab.
42. **Test originator**
43. Test procedure cd
44. Test procedure description
45. **Test Fuel**- fuel type code
    06- Unleaded (at EPA 96 RON)
    09- Diesel (at EPA #2 Diesel)
    22- Special unleaded (91 RON)
    23- Carb Phase II Gasoline
    33 - Methanol(M85)
    39 - Ethanol
    40 - CNG
46. **Test fuel type description**
47. **Analytically derived fe**
48. **Averaging method code**
49. **Averaging method description**
50. **avcd**- averaging code for weighting test fuel economy in cases like testing with and without use of shift indicator light
51. **wt**- weighting factor for averaging mpg values
52. **thc**- HC(hydrocarbon emissions) Test level composite results
53. **co**- CO(carbon monoxide emissions) Test level composite results
54. **co2**- Co2(carbon dioxide emissions) Test level composite results
55. **nox**- NOX(nitrogen oxide emissions) Test level composite results
56. **pm**- particulate matter (for diesel powered vehicles) Test level composite results
57. **fe_unit mpg**- mpg(fuel economy, miles per gallon) Test level
58. **rnd_adj_fe**
59. **target-a** - electric dynamometer target coefficient a
60. **target-b** -electric dynamometer target coefficient b
61. **target-c**- electric dynamometer target coefficient c
62. **set-a** - electric dynamometer set coefficient a
63. **set-b** - electric dynamometer set coefficient b
64. **set-c** - electric dynamometer set coefficient c

V. CITING THE DATASET

A. ASSESS THE DATASET

We find the number of samples in dataset , number of columns, duplicate rows , datatypes of columns, features with missing values ,number of non-null unique values for features in each dataset,etc.

B. Data Preprocessing

The purpose of data preprocessing stage is to minimize potential errors in the models as much as possible. Generally a model is only as good as the data passed into it and ensures that model has accurate dataset. Some columns have missing values and dataset is a max of numeric columns and categorical columns .In this particular dataset, there are 64 columns and many categorical columns are not necessary for analysis like some group id and model codes. We perform cleaning by dropping the extraneous columns, renaming the columns, querying and dropping the columns, fix the data types, work with missing and null values. We find the count of null values, duplicate values and replace them with appropriate values. The null values are in large number which just cannot be drooped since those are values of some unnecessary categorical attributes which can be replaced by any frequent value since that does not affect our dataset or analysis. We also replace some categorical values with integer values for better understanding. Some columns of 2010 dataset are different from 2020 dataset, which we modify accordingly.

C. Data Visualization- EDA

The purpose of EDA is to enhance our understanding of trends in the dataset without involving complicated machine learning models. A correlation map is a visual tool that illustrates the relationship between different columns of the dataset. We find the correlation between the 'vehicle displacement' and 'rated horsepower' alongwith correlation between 'vehicle displacement' and 'equivalent weight of the vehicle'. Draw various conclusions from the boxplots and other graphs. We need to keep in mind about the outliers in the dataset using boxplots, violin plots, bar plots and scatter plots. The predicted mpg values will oftentimes be lower than the actual number. We perform normalization and standard scaling for clear picture of data. We find that there is not much correlation between the variables . Moreover the correlation if further reduced over time. The 2010 dataset shows medium correlation between variables but, 2020 dataset has less of a correlation between the same variables. This shows that the vehicles have improved in such a way that the features are independent such that their values do not affect or least affect the improvement of a vehicle. In the histogram plots of attributes, we find that in 2010 dataset, there are some left or right skewed plots. But, in 2020 dataset, there are more left or right skewed plots indicating less of a correlation.

D. MODELLING

We split the data into training and testing set and also the validation set , inspect the data and start analyzing with different models. We test with different models and find the best fitting model for the dataset using the accuracy value. To use linear regression its necessary to remove the correlated variables to improve our model. We can predict what a car's mpg will be. Since there are multiple algorithms we can use to build our model, we will compare the accuracy scores after testing and pick the most accurate algorithm. We perform the splitting of our data into train, test and validation .We perform some modifactions in the dataset at this stage, since some categorical columns ought to be ignored as they are not much of a help for the modelling and analysis.

During training we take care of errors and fraudulent transactions. In testing phase, we see how the model performs against data where we know the outcome. Through validation we check that the model isn't overfitting to our specific dataset.We found that random forest regression was the most accurate method that can be used for the datasets. For both the datasets, random forest algorithm was most accurate in the performance list. But we also try multiple regression to see the amount of accurateness and how its not a good fit.

We test the model by using sklearn's built-in methods to get RMSE. From that we select lowest number since the algorithm has predicted closest to actual value.We run validation testing on these to ensure there is no overfitting or least as possible. We also perform multiple linear regression to check its performance on the datasets.

VI. CONCLUSIONS

This model could be trained with newer car data and be used to predict competitor's future mpg ratings for upcoming cars, allowing companies to potentially resources currently used on R&D today on making more efficient, more popular vehicles that outshine competitors. We also draw some conclusions from already available data for improvement.

We see that cars are more in count among cars, trucks and both. So , of course we can find a little biased analysis here.

We see that random forest regression was quite helpful in predicting the mpg of the models. We find the different alterantive sources of fuels used such as CNG, hydrogen 5, electricity,Cold CO,etc. It's found that number of unique models using alternative sources of fuel has increased in 2020 by nearly 30 percent more compared to 2010, which tells the improvement in economy and vehicles. We also find which of the drive system and the corresponding model, has improved over time, by generating a bar chart of drive system type vs. increase in average mpg of vehicles.

We also find what are the features that are associated with b etter fuel economy looking at the summary statistics. Explor e trends between mpg and other features in this dataset, sele ct all vehicles that have the top 50% fuel economy ratings to see characteristics. For all the models that were produced in 2010 that are still being produced now, we find by how muc h has the mpg improved and which vehicle improved the mo st using the available mpg values.

Out of all vehicle types including both trucks and cars, Mitsubhisi Motoras Co improved its models the most. Out o f all car types Mitsubhishi model improved the most and out of all truck vehicle types, Honda model improved the most now compared to 2010.

Acknowledgment

We would like to thank our teachers for giving a chance to perform this project. During the process , we learned a lot about various machine learning techniques that have caused drastic changes in terms of technology growth for the large amounts of data that is being produced daily. We were able to find the changes in vehicle fuel economies over the time. 2020 has a better view in improved economy compared to 2010. This shows that in future, the economy increases and improves further with better vehicle features and better rated attribute scores.

Contribution of Team Members:

1. Sai Sahiti Gudla- Data preprocessing and Visualization
2. Ajay- Data cleaning and data preprocessing
3. Deepika K- Data analysis and modelling, visualization, conclusions.

APPENDIX

How Vehicles Are Tested

Fuel economy is measured under controlled conditions in a laboratory using a series of tests specified by federal law. Manufacturers test their own vehicles—usually pre-production prototypes—and report the results to EPA. EPA reviews the results and confirms about 15%–20% of them through their own tests at the National Vehicles and Fuel Emissions Laboratory.

Estimating MPG with Laboratory Tests

In the laboratory, the vehicle's drive wheels are placed on a machine called a dynamometer. The "dyno" simulates the driving environment much like an exercise bike simulates cycling.Engineers adjust the amount of energy required to move the rollers to account for wind resistance and the vehicle's weight. On the dyno, a driver runs the vehicle through standardized driving routines called cycles or schedules. These cycles simulate "typical" trips in the city or on the highway.Each cycle specifies the speed the vehicle must travel during each second in the test.

Measuring Fuel Use

For vehicles using carbon-based fuels (e.g., gasoline, diesel, natural gas, etc.), a hose is connected to the tailpipe to collect the engine exhaust during the tests.The carbon in the exhaust is measured to calculate the amount of fuel burned during the test. This is more accurate than using a fuel gauge.A different method is used for vehicles that run on non-carbon fuels, such as fuel cell vehicles and electric vehicles.

Which Vehicles Are Tested

Manufacturers do not test every new vehicle offered for sale. They are only required to test one representative vehicle—typically a preproduction prototype—for each combination of loaded vehicle weight class, transmission class, and basic engine.Some vehicles are exempt from these requirements:

Motorcycles

Large vehicles prior to 2011: Vehicles with a gross vehicle weight rating (GVWR) over 8,500 pounds, such as larger pickup trucks and SUVs Large vehicles from 2011 onward:

Pickup trucks and cargo vans with GVWR over 8,500 pounds

Passenger vehicles, such as SUVs and passenger vans with GVWR of 10,000 or more.

Detailed information about factors affecting fuel economy:

<https://www.fueleconomy.gov/feg/factors.shtml>

References

[1] Winston Harrington 1, “Journal of Environmental Economics and Management” , Volume 33,Issue 3, July 1997 , Pages 240-353

[2] Gloria Helfand , Ann Wolverton, “Evaluating the Consumer Response to Fuel Economy : A Review Of the Literature”, NCEE Working Paper Series , Working Paper #09-04 , August, 2009.

[3] Tnazila Khan , H. Christopher Frey , “Evaluation of Light-Duty Gasoline Vehicle Rated Fuel Economy Based on In-Use Measurements”, January 1,2016.

[4] L.J.Slater , “Regression Analysis”, The Computer Journal , Voume 4 ,Issue 4,1962, Pages 287-291, <https://doi.org/10.1093/comjnl/4.4.287>, January 1,1962.

[5] Epa.gov. Carbon Dioxide Emissions- Climate Change –US EPA;2015.<http://www3.epa.gov/climatechange/ghgemissions/gases/co2.html> [accessed 20.01.2016].

[6] Prabhjeet Kaur, Nitin Sharma, Sukhpreet Singh. (2020). A Survey on Machine Learning Prediction Algorithms Based on IoMT. International Journal of Advanced Science and Technology, 29(05), 7475-7488. Retrieved from <http://sersc.org/journals/index.php/IJAST/article/view/18240>

[7] Ying Yao, Xiaohua Zhao, Chang Liu, Jian Rong, Yunlong Zhang, Zhenning Dong, Yuelong Su, "Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones", Journal of Advanced Transportation, vol. 2020, Article ID 9263605, 11 pages, 2020. https://doi.org/10.1155/2020/9263605