Module 2 Project | Predict the MPG of cars

Trang Tran

CPS Analytics, Northeastern University

ALY6020 | Predictive Analytics

Professor Prashant Mittal

Nov 12, 2023

Introduction

The Environmental Protection Agency (EPA) annually supplies crucial fuel economy data to federal agencies such as the Department of Energy (DOE), the Department of Transportation (DOT), and the Internal Revenue Service (IRS). This data, sourced from comprehensive vehicle testing at the EPA's National Vehicle and Fuel Emissions Laboratory and submissions from manufacturers, forms the basis for fuel economy-related programs. This analysis focuses on applying Multiple Linear Regression (MLR) to EPA-provided data, specifically examining the MPG (miles per gallon), the variable "RND_ADJ_FE," representing lab-tested fuel efficiency in this dataset.

Data Description and Preprocessing

The original ‘cars’ dataset contains 1129 observations with 39 columns of features. Whereas there are 21 numerical variables (Figure 1) and 18 categorical variables.

```
# Print only numerical columns
print(list(cars))
cars.select_dtypes(include = 'float')
```

```
['Model Year', 'Represented Test Veh Make', 'Represented Test Veh Mode
l', 'Test Veh Displacement (L)', 'Vehicle Type', 'Rated Horsepower', '#
of Cylinders and Rotors', 'Tested Transmission Type', '# of Gears', 'Dr
ive System Description', 'Transmission Overdrive Desc', 'Equivalent Tes
t Weight (lbs.)', 'Axle Ratio', 'N/V Ratio', 'Shift Indicator Light Use
Desc', 'Test Procedure Description', 'Test Fuel Type Description', 'Tes
t Category', 'THC (g/mi)', 'CO (g/mi)', 'CO2 (g/mi)', 'NOx (g/mi)', 'PM
(g/mi)', 'CH4 (g/mi)', 'N2O (g/mi)', 'RND_ADJ_FE', 'DT-Inertia Work Rat
io Rating', 'DT-Absolute Speed Change Ratg', 'DT-Energy Economy Ratin
g', 'Target Coef A (lbf)', 'Target Coef B (lbf/mph)', 'Target Coef C (l
bf/mph**2)', 'Set Coef A (lbf)', 'Set Coef B (lbf/mph)', 'Set Coef C (l
bf/mph**2)', 'Aftertreatment Device Cd', 'Aftertreatment Device Desc',
'Police - Emergency Vehicle?', 'Country_of_Origin']
```

	Test Veh Displacement (L)	# of Cylinders and Rotors	Axle Ratio	N/V Ratio	THC (g/mi)	CO (g/mi)	CO2 (g/mi)	NOx (g/mi)	PM (g/mi)	
0	2.000	4.0	2.95	27.4	0.006645	0.246809	278.759525	0.002467	NaN	0.0
1	2.000	4.0	2.95	27.4	0.007389	0.184334	296.472289	0.004871	NaN	0.0
2	2.000	4.0	2.81	25.2	0.000920	0.082900	195.260000	0.001900	NaN	0.0
3	2.000	4.0	3.91	37.0	0.004932	0.233703	337.165480	0.005107	NaN	0.0
4	2.000	4.0	2.81	34.5	0.003621	0.228178	289.089847	0.006814	NaN	0.0
...	...	...	...	...	...	...	...	...	...	
1124	1.984	4.0	3.09	39.4	0.004400	0.050000	389.000000	0.033800	0.000	0.0
1125	1.395	4.0	3.87	29.1	0.175700	0.737000	284.000000	0.025200	NaN	0.0
1126	1.984	4.0	4.17	26.5	0.006100	0.050000	306.000000	0.011300	0.000	0.0
1127	1.984	4.0	3.23	28.6	0.004100	0.400000	279.000000	0.002900	0.001	0.0
1128	1.984	4.0	3.53	37.3	0.000000	0.016000	212.000000	0.002400	0.000	0.0

1129 rows × 21 columns

Figure 1: Numerical Variables

- ***Handling missing data***

We performed a comprehensive missing data check, including missing values for each row and missing values for each column. We removed rows with too many missing values, in this case, if greater than 10 missing values per row. Next, we checked on the missing values per column (Figure 2) and dropped the two columns ‘PM (g/mi)’ and ‘N2O (g/mi)’ that have a large number of missing values. Finally, we remove all the missing values on other columns since the quantities are acceptable. The final cleaned data set consists of 906 rows and 37 columns.

Figure 2: Missing values per column

```
------------
Model Year                        0
Represented Test Veh Make         0
Represented Test Veh Model        0
Test Veh Displacement (L)         0
Vehicle Type                      0
Rated Horsepower                  0
# of Cylinders and Rotors        15
Tested Transmission Type          0
# of Gears                        0
Drive System Description          0
Transmission Overdrive Desc       0
Equivalent Test Weight (lbs.)     0
Axle Ratio                        0
N/V Ratio                         0
Shift Indicator Light Use Desc    0
Test Procedure Description        0
Test Fuel Type Description        0
Test Category                     0
THC (g/mi)                       70
CO (g/mi)                        69
CO2 (g/mi)                       15
NOx (g/mi)                       75
PM (g/mi)                       773
CH4 (g/mi)                       93
N2O (g/mi)                      341
RND_ADJ_FE                        2
DT-Inertia Work Ratio Rating     30
DT-Absolute Speed Change Ratg    30
DT-Energy Economy Rating         30
Target Coef A (lbf)               0
Target Coef B (lbf/mph)           0
Target Coef C (lbf/mph**2)        0
Set Coef A (lbf)                  0
Set Coef B (lbf/mph)              0
Set Coef C (lbf/mph**2)           0
Aftertreatment Device Cd         15
Aftertreatment Device Desc       15
Police - Emergency Vehicle?       0
Country_of_Origin                 0
```

- ***Encoding Categorical Variables***

  Among the categorical variables shown below (figure 3), we decided to choose three variables 'Country_of_Origin', 'Vehicle Type', 'Drive System Description' to implement on the Multiple Linear Regression model. To incorporate these categorical variables into our model effectively, we performed one-hot encoding. This process results in the creation of binary "dummy" variables that represent each category within the original categorical variables. The new dataset has a total of 45 columns. Afterward, we continued to remove unused remaining categorical columns from the data frame. Thus, the dataset for training includes only 34 columns.

```
# Print only the categorical columns
cars4.select_dtypes(include = ['object'])


listofcat = list(cars4.select_dtypes(include = ['object']))
listofcat
```

```
['Represented Test Veh Make',
 'Represented Test Veh Model',
 'Vehicle Type',
 'Tested Transmission Type',
 'Drive System Description',
 'Transmission Overdrive Desc',
 'Shift Indicator Light Use Desc',
 'Test Procedure Description',
 'Test Fuel Type Description',
 'Test Category',
 'Aftertreatment Device Cd',
 'Aftertreatment Device Desc',
 'Police - Emergency Vehicle?',
 'Country_of_Origin']
```

Figure 3: Categorical variables

- ***Defining Features and Target***

  With the dummy variables in place, we can now distinguish between Features (33 columns) and Target variable (MPG or called 'RND_ADJ_FE’).

- ***Scaling the Features***

  Standardizing the predictors is a crucial step in any model due to the wide range of scales in the factors. This scaling process ensures that all predictors are on a common scale. In this step, we utilize the StandardScaler (z-scores method) to accomplish this standardization.

Implementation of the Multiple Linear Regression Models

First, we split the data set into the train and test sets with a ratio of 80%-20% respectively. Second, we fit the MLR model on the train set and run the predicting process on the test set. The R-squared value in the test set is 0.868 (>86.8%) showing a good predictive power of this MLR model. More than 86.8% of the variance in the target variable (MPG or 'RND_ADJ_FE') is explained by the model (figure 4).

```
Mean Absolute Error: 2.0372591809940985
Mean Squared Error: 7.563182933198569
Root Mean Squared Error: 2.7501241668693015
R-Squared value: 0.8687733855491435
```

Figure 4: MLR results on the full dataset

Besides, we try another approach on a simplified model by selecting only 6 variables 'CO2 (g/mi)', 'NOx (g/mi)', 'CH4 (g/mi)', 'DT-Inertia Work Ratio Rating', 'DT-Absolute Speed Change Ratg' as features of the MLR model. The result of R-squared 85.6% is pretty similar to the previous MLR model, even though with only six predictors.

```
Mean Absolute Error: 2.292907086346667
Mean Squared Error: 8.268505136763979
Root Mean Squared Error: 2.8755008497240926
R-Squared value: 0.8565355426080945
```

Figure 5: Result of a simplified model

References

ALY6020.70767.202415. Module 2. Slides-Sample-Programs-Other Documents.

P_Assignment2_EPA_Cars_Data.ipynb