Fuel Consumption prediction

Sai Nikhil Boyapati
19981120-9692

Abstract— **Fuel consumption is a very important sector in automobile industries. The main goal of this project is to find important factors that influence consumption of fuel. Different Machine learning algorithms are chosen for the implementation. Based on the results random forests performed better than the remaining algorithms.**

Keywords—Fuel Consumption, Data Correlation, Machine Learning

I. Introduction

The main objective is to look at a number of aspects that describe a fuel usage scenario and use ml approaches to find a relationship between those attributes and fuel efficiency. Across most of the world's metropolitan areas, a combination of gases from coal burning and urban automobile use becoming the major source of air pollution [2].

Because fuel is a non-renewable source of energy, it is important to discover measures to limit its use. Alternative fuel cars, such as hybrid and electric automobiles, are being tested, although they have yet to gain widespread acceptance.
Various vehicle design innovations, like as aerodynamic design, engine and transmission efficiency optimization, braking systems, automatic gearbox, and cruise control, are among the technology innovations accessible in a semi-autonomous vehicle, all of which result in massive fuel savings [1]. But at the other side, fuel efficiency depends largely on speed and riding technique

As a result, anomalies in driving style must be identified using situational parameters, and only relevant recommendations should be delivered to drivers. The applicable guidelines for reducing speed and non-fuel-efficient/aggressive behavior can help the operator maintain optimal fuel efficiency.

This project mainly focuses on developing a machine learning model for fuel consumption. The fuel data consists of mpg, acceleration, displacement, engine. We have chosen machine learning algorithms such as Linear Regression, Decision Trees, and Random Forests. We will evaluate the performance of the machine learning techniques using the different types of metrics like mean squared error and r2 score. The comparison of machine learning techniques will be based on metrics results.

Motivation for selection of algorithms:

The linear regression algorithm was chosen to predict the value of a dependent variable based on an independent variable. In the implementation we have divided the data based on the dependent and independent variables. The stronger the linear relationship between the independent and dependent variables, higher accuracy in the prediction.

The reason for choosing random forests because it is a versatile algorithm. Random forest is an appropriate model for our needs, whether we have a regression or classification task. It also can deal with binary, categorical, and numerical attributes. Little pre-processing of data is required and also no need to rescale or transform the data.

In decision tree algorithm, it is not necessary to standardize or normalize the data that is collected. It is capable of dealing with both continuous and categorical variables. It is also saves time like random forests in pre-processing steps while making the model. Outliers are usually treated by Decision Tree and handled automatically.

II. Related work

The authors [3] have proposed an approach on Driver Behavior on Fuel Consumption. The goal of this study is to implement two machine learning methods in evaluating fuel consumption. First method is to implement unsupervised spectral clustering algorithm using the collected data. The dynamic input from the driving environment and natural driving data are combined in the second method to create a conceptual model between various driving behaviors and their associated fuel consumption parameters. The authors used LSTM model to predict consumption of fuel.

The authors [4] have proposed an approach on Vehicular fuel consumption estimation using real-world measures. The goal of this study is to implement a machine learning method in evaluating fuel consumption using data collected from a fleet of 27 vehicles. The authors improved accuracy by introducing engine speed. The authors used Artificial neural networks and support vector machine in predicting consumption of fuel for 27 vehicles.

The authors [5] have proposed an approach on Predicting Gasoline Vehicle Fuel Consumption in Energy. The goa l of this study is to use different types of machine learning techniques like linear regression, naïve bayes, neural networks, random forests and LightGBM in fuel consumption prediction. The performance of machine learning models is evaluated by different types of metrics like R squared, mean squared error and mean absolute error.

III. METHOD

In the data set each feature is selected by performing feature selection techniques like Data Correction, this will help us to make the machine learning effective. During the data preprocessing correlation helps in removing the features which have more correlation values. The following are the steps which are followed in this project.

- The data is collected from different types of cars and have extracted them which are required in this project.
- Selecting and applying machine learning techniques.
- The performance of the machine learning model is evaluated by comparing the performance metrics like r2 score, mean absolute error.
- Based on the performance metrics values the appropriate algorithm is chosen.

A. *Experimentation Environment*

Python is a versatile programming language which can be used for a variety of tasks. Python is widely used in data science and is a common choice because it is both dynamic and open source. There are many in-built libraries in this language which we can manipulate and visualize the data in different ways. Most common libraries used in this project are numpy, pandas, matplotlib, seaborn and Sklearn [7].

B. *DataSet Description*

For this project we tried various sources but there is a very limited data available on the internet. Therefore, we chosen the data which we have gathered from the various sources. In this project different types of attributes have been collected which are related to cars in fuel consumption. The data is used for training and testing the machine learning model.

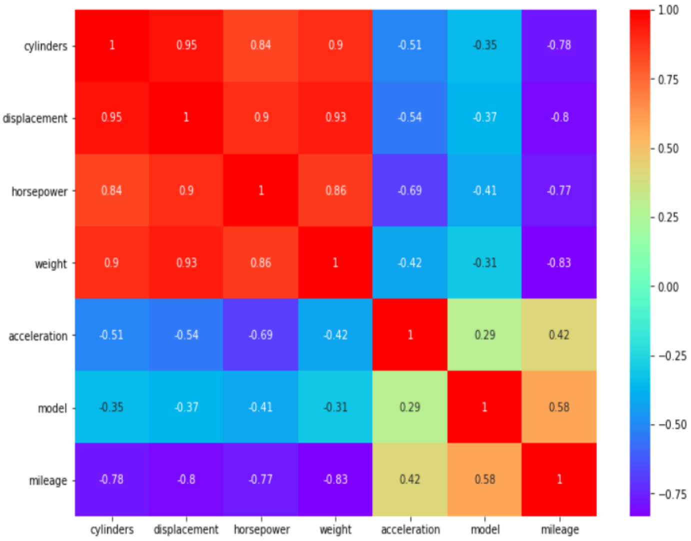
This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This dataset is downloaded from UCI machine learning repository. The following are the features which are present in the dataset.

- cylinders - Number of cylinders
- displacement - Car engine displacement
- horsepower - Car engine performance
- weight - car weight
- acceleration - car acceleration
- model - car model
- origin - car origin(country)
- car - car name
- mileage - mileage per liter

C. *Feature selection*

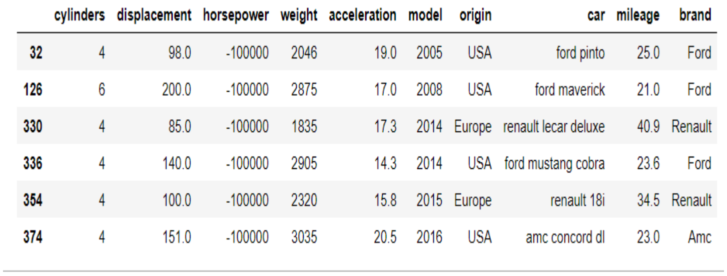
In this project data correlation have been used in feature selection process. The evaluation of machine learning model on all features which have low correlation value may impact on the accuracy of the model. Therefore, data correlation method helps us in feature selection.

In the following figure heatmap is used to visualize the correlation of all features.



D. *Data Precessing*

This is a very important method in building machine learning model. For this project we found redundant values in different features. In the below figure for the horsepower column there are negative values exists. In general, for any type of car there is no negative values for horsepower. There are few ‘nan’ values in different columns. We have used statistical method like median for columns which have null values and redundant values.



	cylinders	displacement	horsepower	weight	acceleration	model	origin	car	mileage	brand
32	4	98.0	-100000	2046	19.0	2005	USA	ford pinto	25.0	Ford
126	6	200.0	-100000	2875	17.0	2008	USA	ford maverick	21.0	Ford
330	4	85.0	-100000	1835	17.3	2014	Europe	renault lecar deluxe	40.9	Renault
336	4	140.0	-100000	2905	14.3	2014	USA	ford mustang cobra	23.6	Ford
354	4	100.0	-100000	2320	15.8	2015	Europe	renault 18i	34.5	Renault
374	4	151.0	-100000	3035	20.5	2016	USA	amc concord dl	23.0	Amc

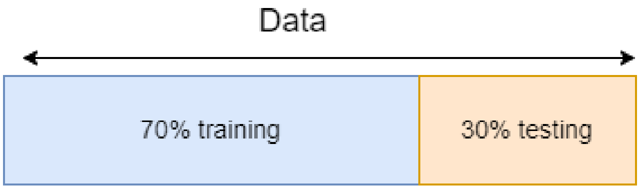
Categorical Encoding:

Categorical encoding is the process of converting the data into binary format. Each category is represented by a binary variable with a value of 0 or 1. The absence of that category is represented by 0 while the presence of that category is represented by 1.

Dummy variables are the name for these newly produced binary attributes. The number of dummy variables used is determined by the category variable's levels.

Train test split:

Train test split method is used in this project before building the machine learning model. The data is divided into training and testing i.e., 70% of the data is used for training and 30% of the data is used for testing.



Performance Metrics:

To measure the performance the of the machine learning model performance metrics is used. In this project two types of performance metrics such as r2 score and mean square error are used [6].

I. Mean square error:

This is a metric used in regression problems. It states the determining the squared difference between the actual and expected value.

$$\text{MSE} = \frac{1}{n}\sum(y - Y)^2$$

II. R2 squared:

It is also a most commonly used metric in regression problems. It refers to the difference between the dataset's samples and the model's predictions.

$$\text{R2 score} = 1 - \frac{SS_r}{SS_m}$$

Where SS_r is the squared sum error of regression line and SS_m is the squared sum error of mean line

IV. Results & Analysis

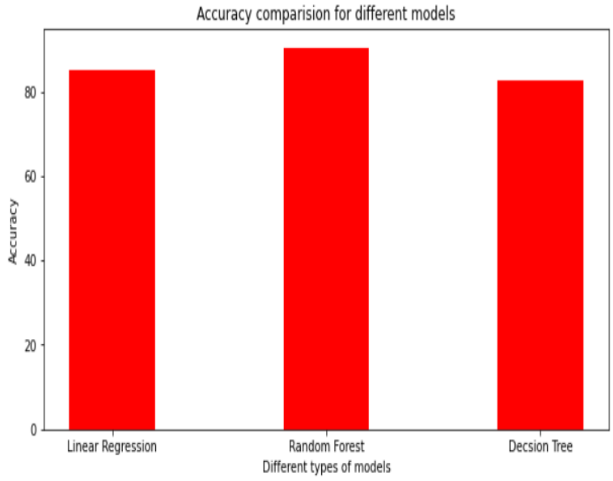
A. *Analysis*

The reason for random forests outperforming decision trees and linear regression is as follows:

- A specific number of attributes is prioritized by the decision tree model. During the training phase, moreover, the random forest selects features at random. As a result, it is not overly reliant on any one set of features.
- When compared to linear regression algorithm risk of overfitting in random forests is low hence it will lead to increase in accuracy. The outliers in linear regression are very sensitive hence these will lead to some decline in accuracy whereas the random forests are robust to outliers.

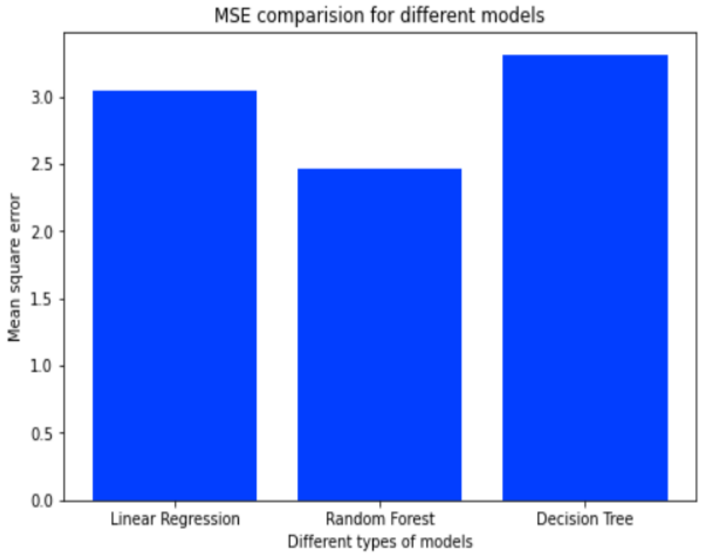
B. *Results*

Accuracy:



From the figure above based on the r2 score we have chosen random forests performs better when compared to the rest of the algorithms.

Mean square error:



From the above figure mean square error value is more for linear regression and decision trees when compared to random forests. The larger the value means the larger the error.

V. Discussion

In the modern world as there is a lot of money is used only for fuel, so we need to use the fuel very efficiently while using different vehicles. Therefore, mileage plays a very important role in fuel economy.

As per the results and analysis section, random forests is the efficient algorithm for predicting mileage for the cars in the data set after comparing the r2 score and mean square error. Random forests produced the least error in predicting mileage when compared to linear regression and decision trees. There are also different metrics while training the algorithms they can also play a vital role in training some algorithms.

VI. CONCLUSION

Fuel consumption is very important sector in oil industries and also automobile field. With the help of this topic fuel revenue analysis helps us to get estimated revenue details. Different types of machine learning algorithms have been evaluated to find important factors for fuel consumption. One can also do research on this topic which can be very helpful for the society.

REFERENCES

[1] Jain, Neetika, and Sangeeta Mittal. "A machine learning pipeline for fuel-economical driving model." International Journal of Intelligent Computing and Cybernetics (2021).

[2] P. Ping, W. Qin, Y. Xu, C. Miyajima and K. Takeda, "Impact of Driver Behavior on Fuel Consumption: Classification, Evaluation and Prediction Using Machine Learning,"

[3] Ping, Peng, et al. "Impact of driver behavior on fuel consumption: Classification, evaluation and prediction using machine learning."

[4] Moradi, Ehsan, and Luis Miranda-Moreno. "Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling." Transportation Research Part D: Transport and Environment 88 (2020):

[5] Yang, Yushan, et al. "Predicting Gasoline Vehicle Fuel Consumption in Energy and Environmental Impact Based on Machine Learning and Multidimensional Big Data."

[6] https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/ (Accessed: 13-03-2022)