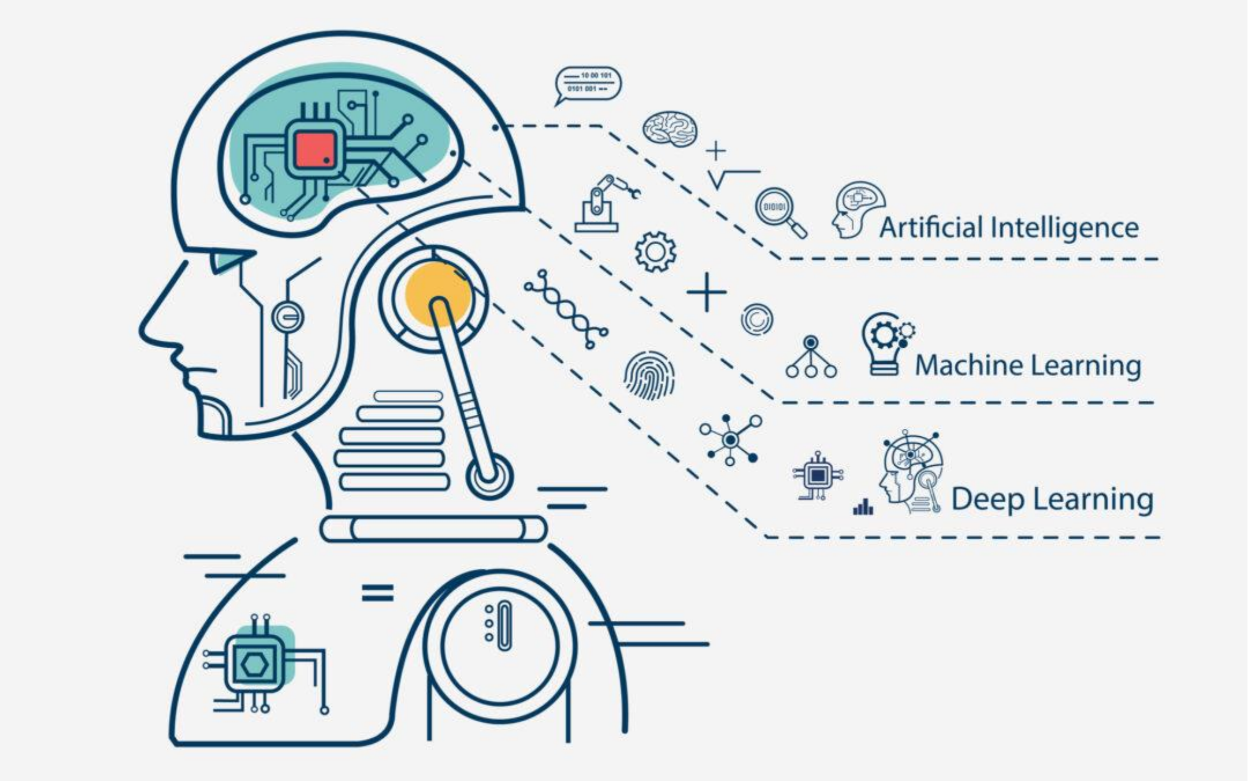
Smart Bridge Externship Project



Car Performance Prediction Using IBM Watson Studio

Submitted By

Contents

Introduction

Overview

Machine learning refers to a class of computer algorithms that learn from examples rather than being explicitly programmed to perform a task. It learns to formulate a general rule from a set of concrete examples. Thus, like human learning, the computer becomes capable of improving its performance from acquired knowledge. The difference is that, at the current state of our knowledge, the computer needs many more learning examples than people do.

To realize true potential, we create a system that is capable of predicting the desired output using supervised machine learning. The system or model is defined by a datasets as in csv file undergoes pre-processing, followed by scaling and conversion of categorical values to numerical form for machine understanding. It is then trained and tested and accuracy is noted to see the efficiency with which the output can be precise to actual data.

Purpose

The purpose of the project is to illustrate the use of machine learning as an easy means to predict possible output in case for favorable events. Practically, if we define characteristics of a thing as features or called datasets and feed it to computer, it trains itself based on it and if similar data provided to it, it will provide possible outcome value as a result. Its important as it can help in determining favorable conditions that can be helpful for an individual or a company if looked at future aspects.

The purpose of machine learning is to discover patterns in your data and then make predictions based on often complex patterns to answer business questions, detect and analyse trends and help solve problems. Machine learning in business and other fields is effectively a method of data analysis that works by automating the process of building data models.

Literature Survey

Existing Problem

The desired problem set defines features of a car engine in a csv file that can be used to predict mileage of a car per gallon of litres. Thinking theoretically, it would take long measures to generate a formula with complex calculations and will present a lot of hard work after which even the best of result is not guaranteed. Also, generation of such a formula would lead to leverage of hundreds of possible assumptions due to large set of parameters and the prediction through them may be absurd if a mistake occurs.

The process is time consuming and may be fatal with even a single mistake. It would require research that may be physical because there are no such algorithms exist that may account for it. Even if it is easy to create a solution, the problem may arise that it might not predict result outside the provide input and may not be valid for other similar datasets with similar features and predictions. In such case, it is rather unsatisfactory to be use in any way.

Proposed Solution

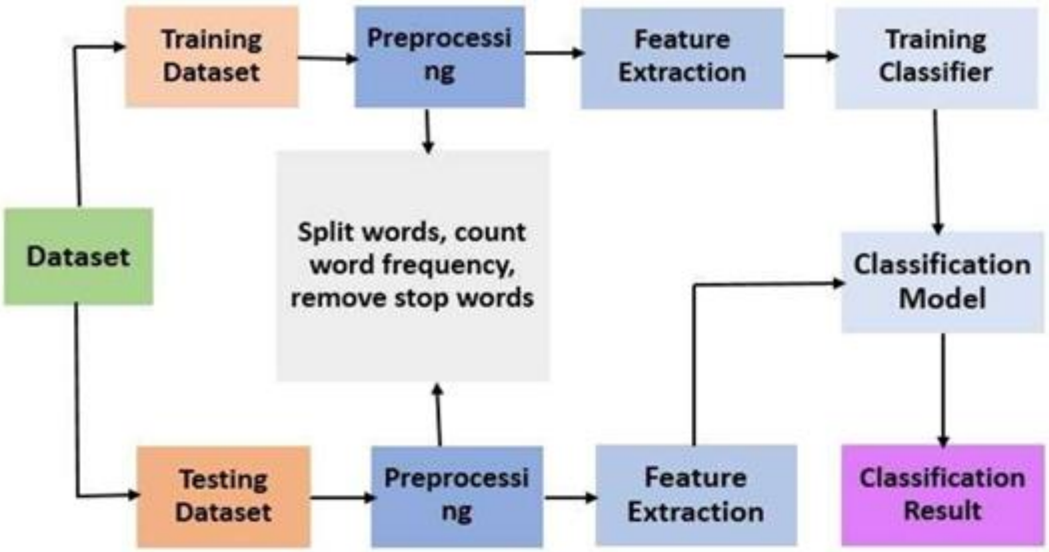
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Machine learning uses predictive analytics, or predictive modeling. Its goal and usage is to build new and/or leverage existing algorithms to learn from data, in order to build generalizable models that give accurate predictions, or to find patterns, particularly with new and unseen similar data.

The goal is to get best possible solution without much effort to leverage our audacity to work. Therefore, the dataset or features of our car engines acts as an input to system and using an Algorithm we define them. These are then trained and tested for best efficient solution that can provide suitable output as required.

Theoretical Analysis

Block Diagram



The following diagram showcases how Machine Learning is used in creating a Model using Datasets provided to it. It follows

Dataset is distributed into train and test datasets.

It is followed by pre-processing for the trained datasets which includes

- Checking of null values if exist
- Removing of not required value or feature
- Scaling to limit range for features
- Encoding to convert categorical to numerical form

These feature are then extracted and and trained using a required algorithm. Thus the model is built which is tested using test datasets and to determine efficiency.

Software Designing

The Software designing includes using dataset to create a model using a required algorithm which is trained to predict possible output.

For this we use

Jupyter Notebook –

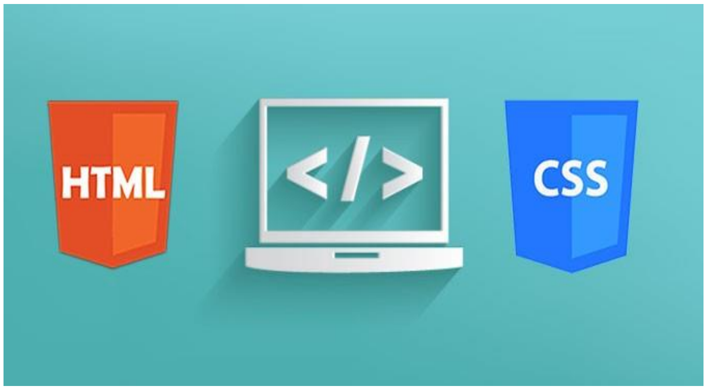
To import dataset, pre-processing, Scaling, applying Algorithm and thus model creation.

IBM Watson Studio Cloud –

For creating an AI model trained for higher accuracy at cloud.

HTML and CSS File –

For visualization of model at local host in GUI form for enhanced look.



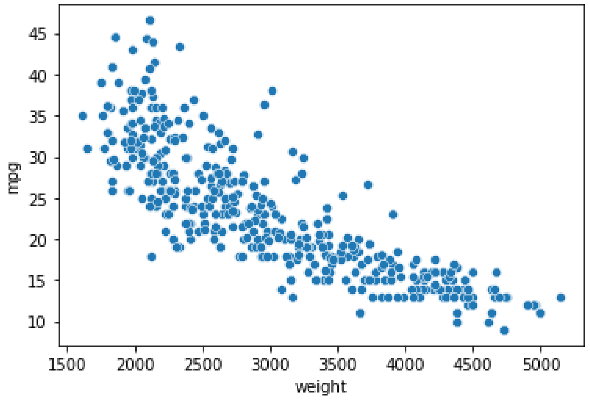
Experimental Investigations

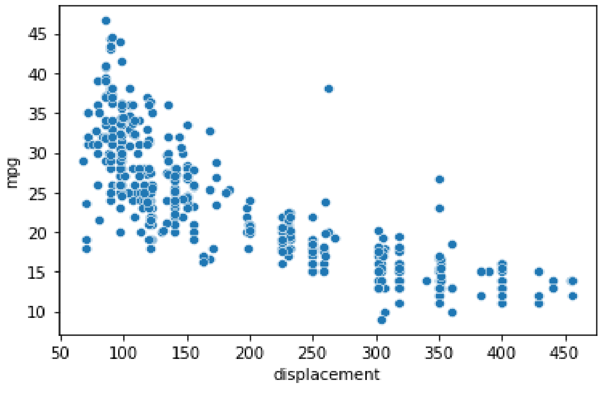
Investigations done to understand the nature of datasets provided with features describing the performance parameters of a car engine based on

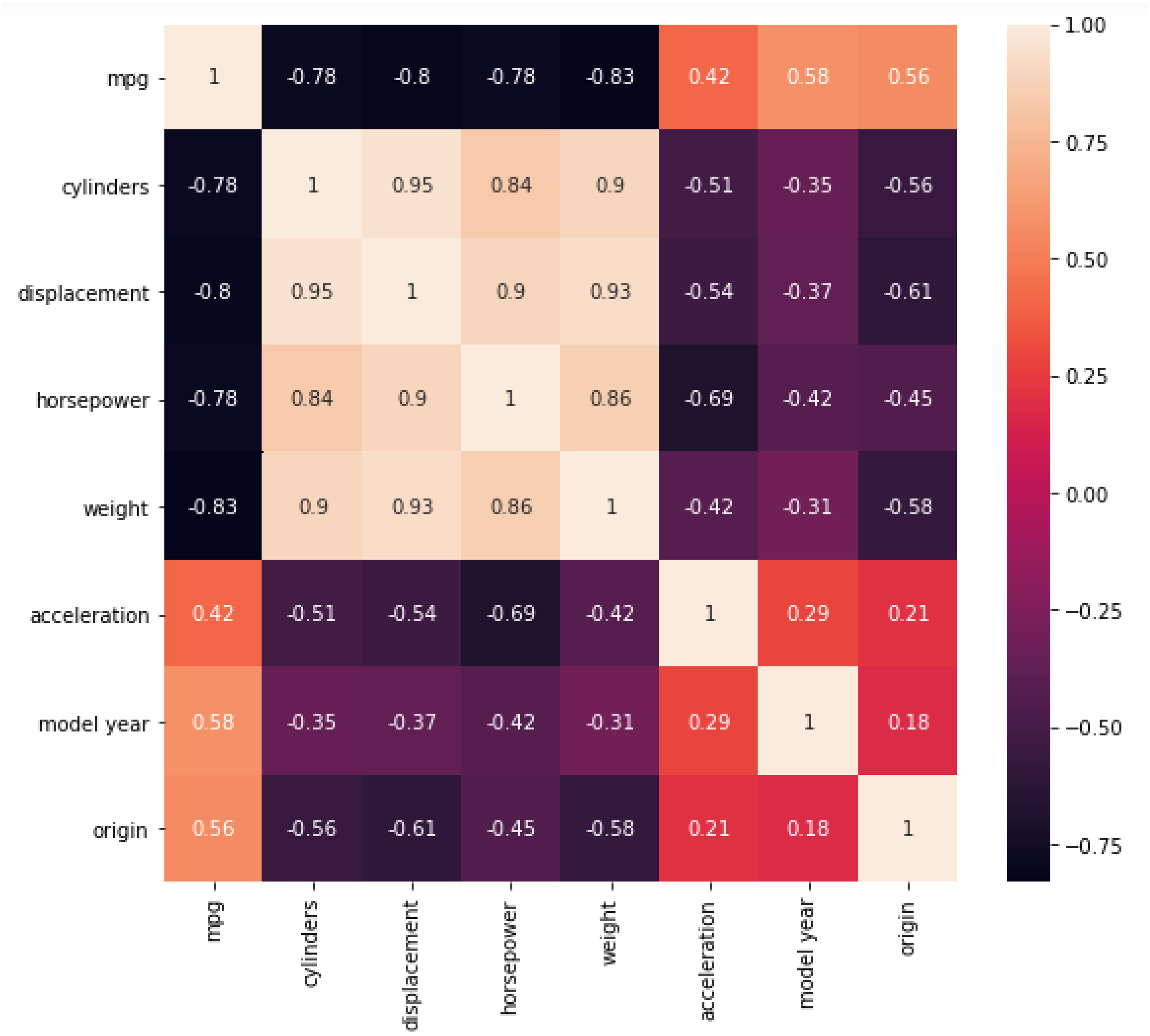
- Car Name
- Model Year
- Displacement
- Origin
- Weight
- Horsepower
- Accelaration
- No. of Cylinders

The Output to find is “mpg or miles per gallon” which is efficiency of car driven. Since Output is continuous, so we go for Regression algorithms.

The dataset is pre-processed i.e. removal of null values, scaling, outliers detection etc. It is followed by visualization of features for understanding the dataset efficiently.





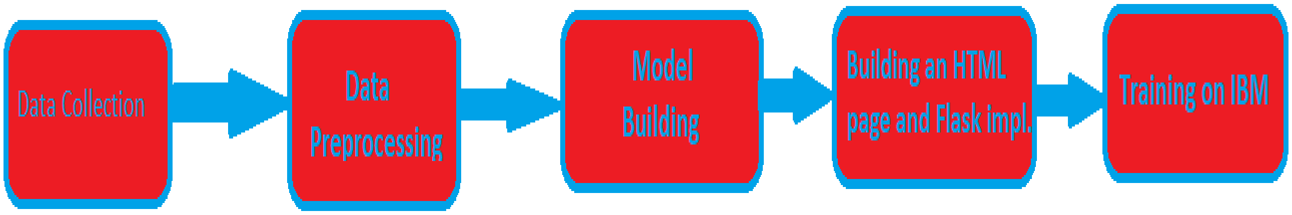


	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
mpg	1	-0.78	-0.8	-0.78	-0.83	0.42	0.58	0.56
cylinders	-0.78	1	0.95	0.84	0.9	-0.51	-0.35	-0.56
displacement	-0.8	0.95	1	0.9	0.93	-0.54	-0.37	-0.61
horsepower	-0.78	0.84	0.9	1	0.86	-0.69	-0.42	-0.45
weight	-0.83	0.9	0.93	0.86	1	-0.42	-0.31	-0.58
acceleration	0.42	-0.51	-0.54	-0.69	-0.42	1	0.29	0.21
model year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1	0.18
origin	0.56	-0.56	-0.61	-0.45	-0.58	0.21	0.18	1

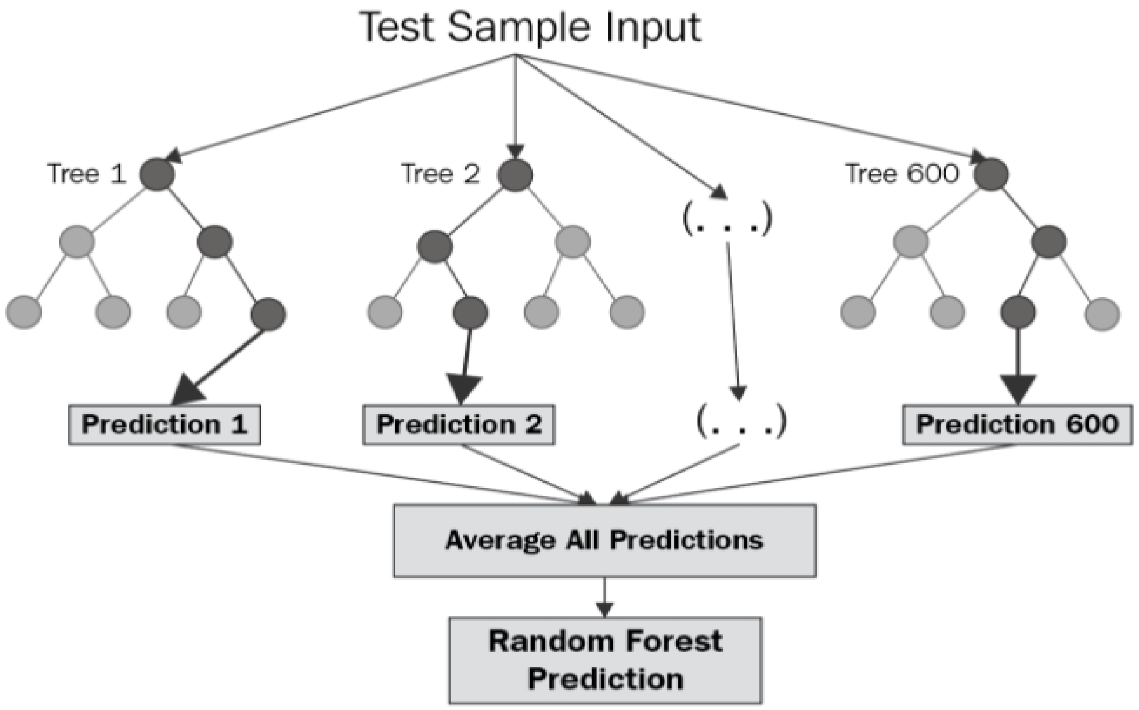
After that co-relation is being founded between required features to categorise them. Since, the features are highly correlated and does not show a linear relation, Some Regression Algorithms can’t be applied and even if trained with them the efficiency would be low.

We used Random Forest Regression which is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

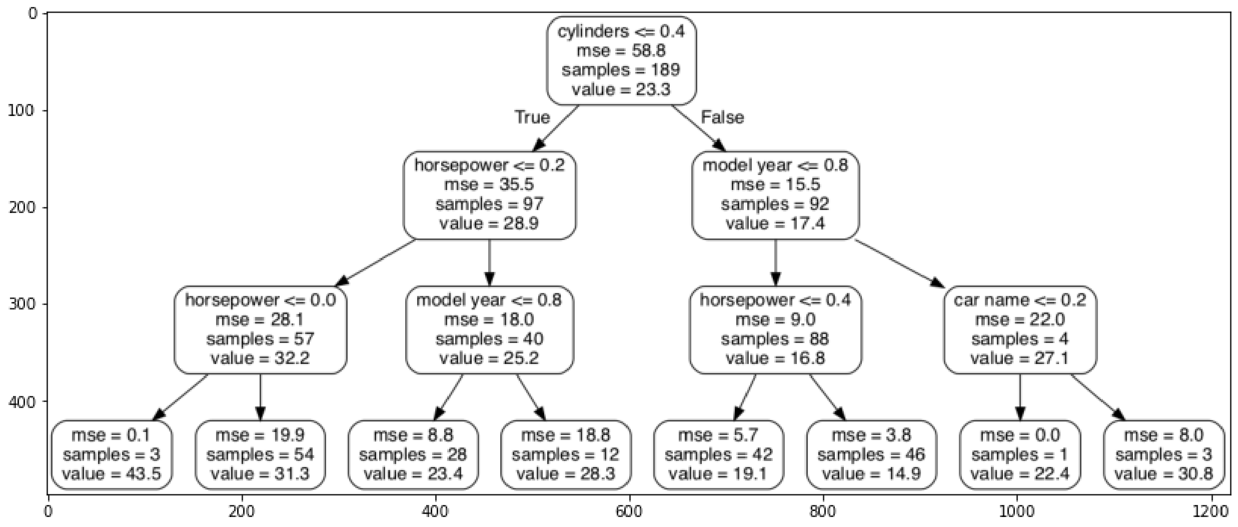
Flowchart



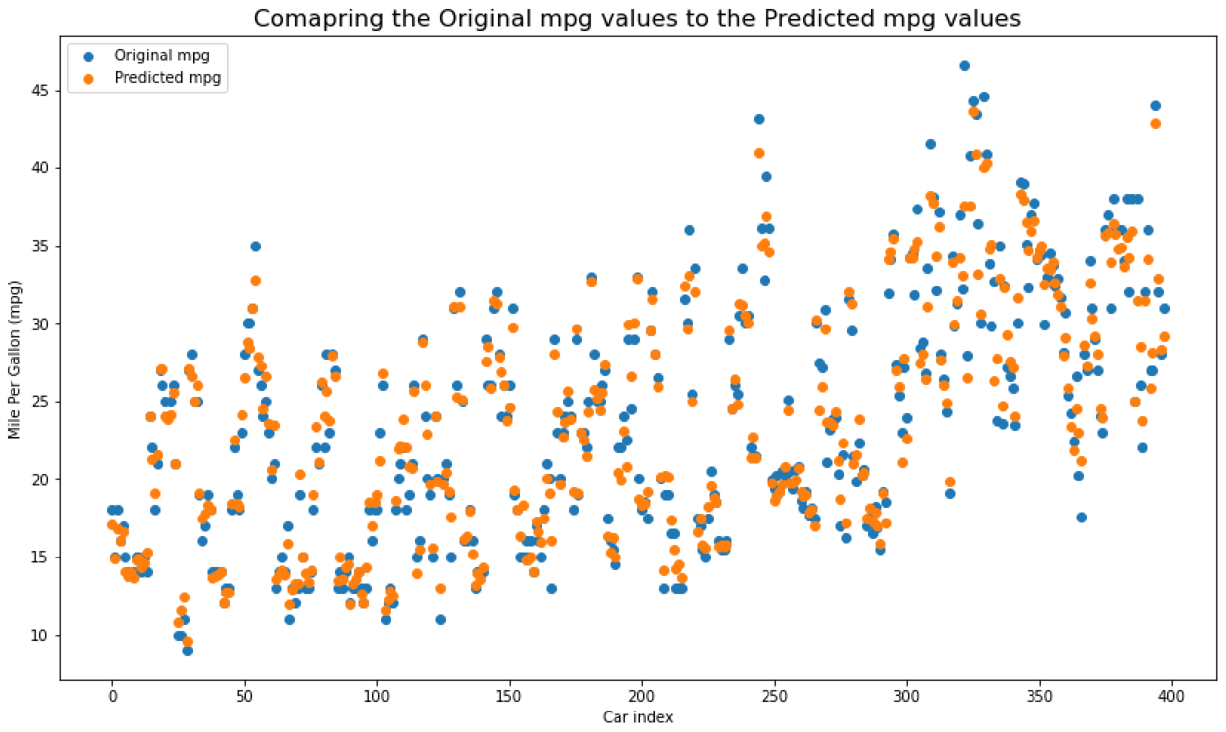
A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.



Results



cylinders <= 0.4
mse = 58.8
samples = 189
value = 23.3
True
False
horsepower <= 0.2
mse = 35.5
samples = 97
value = 28.9
model year <= 0.8
mse = 15.5
samples = 92
value = 17.4
horsepower <= 0.0
mse = 28.1
samples = 57
value = 32.2
model year <= 0.8
mse = 18.0
samples = 40
value = 25.2
horsepower <= 0.4
mse = 9.0
samples = 88
value = 16.8
car name <= 0.2
mse = 22.0
samples = 4
value = 27.1
mse = 0.1
samples = 3
value = 43.5
mse = 19.9
samples = 54
value = 31.3
mse = 8.8
samples = 28
value = 23.4
mse = 18.8
samples = 12
value = 28.3
mse = 5.7
samples = 42
value = 19.1
mse = 3.8
samples = 46
value = 14.9
mse = 0.0
samples = 1
value = 22.4
mse = 8.0
samples = 3
value = 30.8
0
100
200
300
400
0
200
400
600
800
1000
1200



Comapring the Original mpg values to the Predicted mpg values
● Original mpg
● Predicted mpg
Mile Per Gallon (mpg)
45
40
35
30
25
20
15
10
0
50
100
150
200
250
300
350
400
Car index

Advantage and Disadvantage

1. It reduces overfitting in decision trees and helps to improve the accuracy
2. It is flexible to both classification and regression problems
3. It works well with both categorical and continuous values
4. It automates missing values present in the data
5. Normalising of data is not required as it uses a rule-based approach.
6. Handle Missing Values.
7. Robust to Outliers.
8. Less impacted by noise.

However, despite these advantages, a random forest algorithm also has some drawbacks.

1. It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
2. Longer Training Period as it combines a lot of decision trees to determine the class.
3. Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.
4. Biased while dealing with categorical variables.
5. Not suitable for linear methods with a lot of sparse features.

Applications

The applications are not limited to any industry. Because of its high efficiency, it is used in various sectors for better operations

- Banking Industry
- Credit Card Fraud Detection
- Customer Segmentation
- Predicting Loan Defaults on LendingClub.com
- Healthcare and Medicine
- Cardiovascular Disease Prediction
- Diabetes Prediction
- Breast Cancer Prediction
- Stock Market
- Stock Market Prediction
- Stock Market Sentiment Analysis
- Bitcoin Price Detection
- E-Commerce
- Product Recommendation
- Price Optimization
- Search Ranking

Conclusion

To our knowledge, we applied Random Forest Regression Algorithm to our dataset consists of parameters based on car features and engine parameters that determine its fuel efficiency. After training and testing it, we found that the overall accuracy of our trained model is over 89% which is quite efficient and thus could be helpful for any source who wants to understand cars features.

To improve efficiency, certain other parameters can also be added or removed as per requirement. No. of estimators and decision tree numbers can be varied also for it.

Future Scope

The model we obtained is quite efficient and useful in predicting a car performance. Several Automobile Industries can refer it to produce new cars with certain features to attract customers. This model will also help dealers that re-sell cars to show efficient data in easy form for customers to understand such that a certain value can be put on cars too.

Using Machine Learning a variety of such models can be trained based on inputs given to them. Various algorithms can be applied based on relationship obtained between required input and output parameters. Higher the efficiency of model, higher the probability to get an accurate possible result.

Bibliography

http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html

https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees

https://www.mygreatlearning.com/blog/random-forest-algorithm/

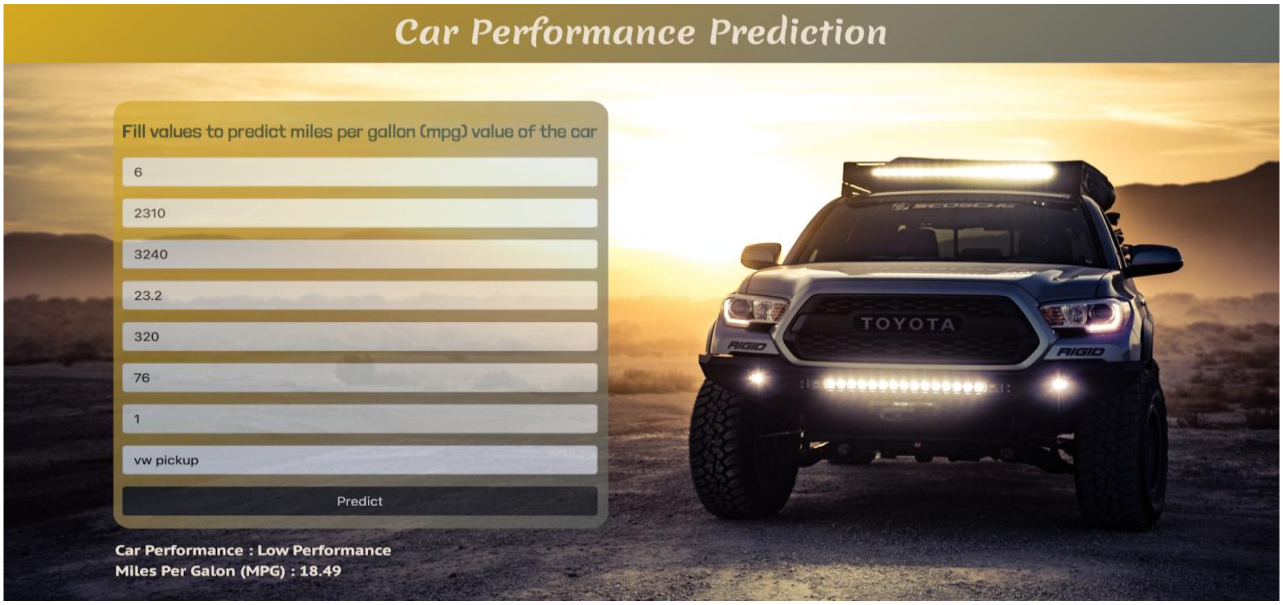
https://www.sciencedirect.com/science/article/pii/S0168169920303355

Appendix

Source Code

Refer to attached files in the folder for code.

UI Output



Car Performance Prediction

Fill values to predict miles per gallon (mpg) value of the car

6

2310

3240

23.2

320

76

1

vw pickup

Car Performance : Low Performance

Miles Per Galon (MPG) : 18.49