CO_2 Emissions Due To Fuel Consumption

1st NAVEEN C S
Dept of computer science Engineering
PES University,Bengaluru,India
SRN : PES1201801744
naveencs927@gmail.com

2nd ASHAY G
Dept of Computer Science Engineering
PES University,Bengaluru,India
SRN : PES1201801767
ashaygowda@gmail.com

3rd BHARGAV V
Dept of Computer Science Engineering
PES University,Bengaluru,India
SRN : PES1201801796
bhargavv1245@gmail.com

4th MANOJ KUMAR P
Dept of Computer Science Engineering
PES University,Bengaluru,India
SRN : PES1201801923
manoj13812@gmail.com

I. Introduction and background

The worldwide atmosphere conditions is changing rapidly which has become a danger and perhaps the best test looked by the global community.Transport is among one of the primary factor causing one fifth of ozone harming substance emanations and furthermore a huge portion of air pollution.Presently, there are 1.42 billion vehicles in activity around the world, including 1.06 billion traveler vehicles and 363 million business vehicles.The dataset gives model-specific fuel utilization evaluations and estimates carbon dioxide outflows for new light-obligation vehicles for retail deal in Canada for year 2019 and through this we try to know the CO_2 emission for exhaust emissionss of carbon dioxide (in grams per kilometer) for combined city and expressway driving

The carbon dioxide emissions which would spread across the ocean and land called the uptake. If the CO2 emissions from the different sources and the uptake are at the same rate, then the concentration stays steady. Because of burning of the fossil fuels our emissions are higher currently than our uptake. Over the last six decades, we've quadrupled our rate of carbon dioxide emission. Stopping the increasing in our emissions isn't enough to stop the increase in concentration because carbon dioxide is a heat-trapping gas that will warm our planet causing melting ice, rising sea levels, changing our weather patterns, more droughts and hurricanes. In the latest international scientific reports on climate change, scientists emphasize the fact that it's not enough to level off emissions, we need to rapidly reduce them.In the future, researchers can analyze the impacts of efforts to intentionally improve air quality as well as reduce greenhouse(CO2, specifically) gas emissions.

This study has helped a lot to learn about the modern car's fuel consumption and CO_2.This itself clarifies the intention of the work we have done on the project.Its all about keeping the environmental pollution under check.The major questions we answer through these projects like the CO_2 emissions of the vehicles and calculating the CO_2 through the features of the vehicles and comparing these calculations with different models.The specific problem we are focused on the solving is the CO2 emissions due to light duty vehicles.We tend to train and calculate CO2 emissions the exhaust emissions of carbon dioxide (in grams per kilometre) for combined city and express driving through various regression models.

We have accounted to the fact that light duty vehicles generally are travelled in city as well as highway roads.Fuel consumption City and highway fuel consumption ratings are shown in our dataset as litres per 100 kilometres (L/100 km) - the combined rating (55percent city, 45percent highway) is shown in L/100 km and in miles per imperial gallon (mpg).We have used data analytics tools to observe that here most models emits CO_2 in the rate of 250 which is likely to be eco-friendly than higher rates.Through multivariate analysis we have determined the factors affecting the CO_2 emissions.

We have used different regression models and compared the accuracy and determined the best model to estimate the CO_2.Through this we can decide the cars which are good for the environment.This has various applications.We can directly know the model of the cars which are well under that 250 mark of carbon di-oxide emission value that we desire.This is method is used in our project for regression and classification problem in out project.

Through this project we intend to reduce the CO_2 footprint on earth.Although we know that electric cars the future and leading the game in this field is Tesla but it is not to be forgotten that these fossil fueled cars not going anywhere in the near future.Due to this analytical study in carbon emission we can know the fuel consumption and the carbon emission of different vehicles.Through this as a future work we can try to develop automated system to recognize the emission of these vehicles and help them to fix this issue.

II. Previous work

A. Brief review of the most relevant predecessor work

The most relevant predecessor work is the paper we have referred to called as Fuel consumption models applied to

automobiles using real time data : A comparison of statistical models by Ahmet Gurcan Caparaz,Pinar Ozel,Mehmet Sevkli,Omer Faruk.

This paper was published in the 6th International Conference on Sustainable Energy Information Technology (SEIT) IN 2016.The main purpose of choosing this Research paper is that this paper is very close to our problem we are to solve and the domain of both our problems also point out to the same substance.

The main and the precise aim of the paper.This paper mainly aims on the comparison of performance between the three statistical models which are Support Vector Machine(SVM) ,Artificial Neural network and multiple linear regresion.Unlike us they have done it on more of a realtime data than us doing it on the a data which is pre collected and is taken from kaggle.These researchers have taken data from the on board diagnostics ,bluetooth interface and a smartphone.

The statistical models used in this paper are :

- Multiple Linear Regression
- Artificial Neural Network
- Support Vector Machines

The results that are generated by the model stretch depending on the total consumption and instant consumption correlation.Support Vector Machines (SVM) provides better results for all automobile kinds comparing Artificial Neural Network(ANN) and linear regression. Support Vector Machines (SVM) outperforms other models for both accuracy and consistency looking at average and standard deviation of R values. Even when we just use speed, acceleration and slope as input values Support Vector Machines (SVM) delivers better results than other statistical model.Support vector machines gives a value of 0.09R on testing data with only a 0.009 standard deviation after a thousand runs. Neural Networks also gave fair results but raced infront with excellent results in the testing data which can be proven by the standard deviation of R values.

B. Limitations and Assumptions

The data was collected when drivers were in normal course.This limitation cannot be addressed by us as we didn't get a data set which consisted of non-normal course drivers.Another limitation of their work is they have done three statistical models which rather is not a limitation and we have done more than three models which are Multiple Linear Regression,Decision Tree Regression,Polynomial Regression,Support Vector Regression.

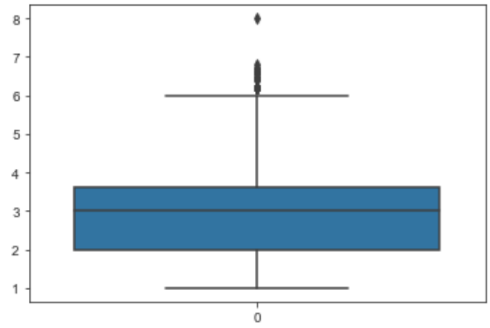
Our data set provide model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada.This data is from Canada rather it is being infered to all countrieshis not to go without mention is that light duty vehicles are the ones with low carbon emission rates.Here we are not addresing the category of heavy duty vehicles which contribute to most of the high carbon emission rate.

The specific set of vehicles which we are adressing is the light-duty vehicles

III. Proposed solution

A. Preprocessing

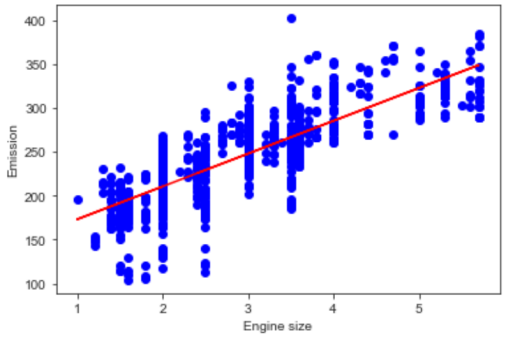
As a part of Preprocessing to our data we found no missing values.But we have found some outliners.We have removed the outliners as they will harm the data set an tend to give biased results.There were no inconsistent, incomplete, duplicate or incorrect data.Preprocessing contains techniques such as dimensionality reduction, range transformation, standardization.We have not done dimensionality reduction as all the dimensions aka attributes were important for the final machine learning models.



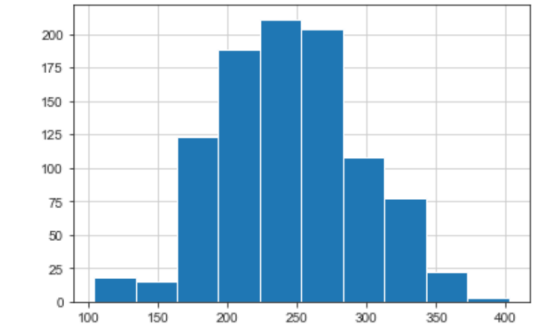
There is no range transformation as they were not required.Standardization was not done as a part of preprocessing but as a part of multilinear analysis this was done by default.Principal component analysis(PCA) does not help in the visualization of data because we have not done any operations based on dimensionality reduction as all rows were important to the output.

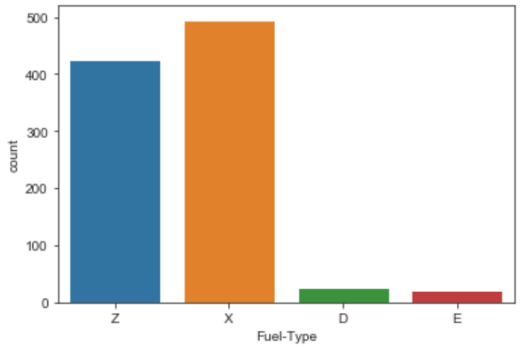
When it comes to Exploratory data analysis(EDA) and Visualization technique such as scatter plots, Correlation coefficients,probability plots and regression plots.To take deeper insights at data we also plotted histograms,bar charts.We have brought out insights and relations between various attributes of the data set as a part of Exploratory data analysis(EDA) and Visualization.

The scatter plots were done to realize the relation or rather to say in terms of data analytics Correlation between the variables aka attributes.This was followed by the Correlation coefficient which gave strength in quantitative terms and direction of the linear relationship between two or more variables. This helped us to realize the overall crux of the data,and to determine what types of data is building up our database.

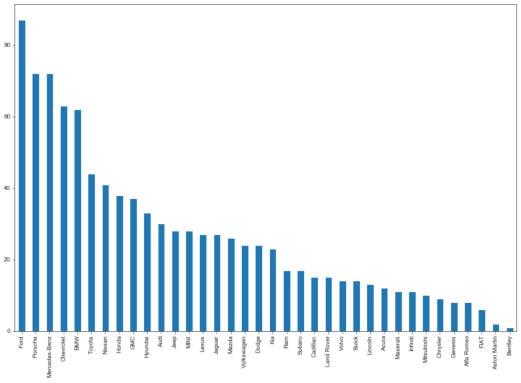


The categorical data which had to be taken into account for drawing a deeper insights to data.We plotted bar charts for fuel type used by the light weight vehicles.We plotted histograms for $C0_2$ emissions,Engine-Size and cylinders.

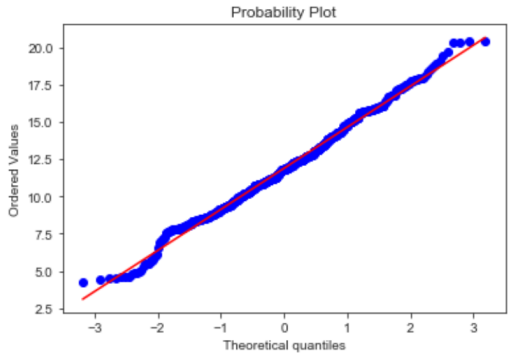




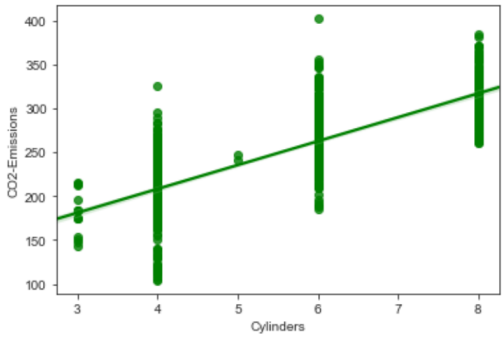
The histogram for the models in the decreasing order based on the counts.These gives us some pattern and the insights on what type of values are these attributes made of.



There are also probability plots.These are plotted to see if the data set follows normal distribution or not.



There are also regression plots.These are plotted to distinguish a line between 2 parameters and helps to visualize their linear relationships.



All the Exploratory data analysis(EDA) and Visualization techniques have done in jupyter notebook and pictures of some of them are attached and these pictures give us a detailed and precise information about the data and the hidden patterns in it and also the various well described summmary about the dataset.

B. Building models

Ours is basically a regression problem and for these problems we decided to train and test some of the well known supervised learning machine learning models.These models are :

- Multiple Linear Regression
- Decision Tree Regression
- Polynomial Regression
- Random Forest Regression

We have chosen the dataset which has the target variable or dependent variable as the continuous output which gives the predictions based on the continuous output. We are going to understand and build the models mostly based on the regression.

We have used the regression models to build and predict the CO_2 emissions based on the training given to the training set. Most of our models and predictions are done using the libraries like scikit learn which is most popularly used for building the models like Linear regression model, Decision Tree Regressor etc. We have used the scikit learn in the pre processing for splitting the data into the train test split. Since the models we have built used the regression to predict out the continuous values of CO_2 emissions we have discarded the categorical attributes which do not have much correlation with the CO_2 emissions.

Multiple Linear Regression:

In the multiple linear regression firstly we have split the dataset into the train test so that we can test with some part of the data within the dataset. Here we have the built-in linear model called Linear Regression imported from the module sklearn. We are going to fit the data which is to be trained and would give insights to the model and predictions so that when it is given a test data without the target variable it should be able to predict the output based on the trend the model has followed from the training dataset. In general terms the

independent variables should be able able to predict out the results of the target(dependent variables). R2 score is like a test for the accuracy where we can check out the proportion of correctly predicted to the total number of predictions. We are actually interested in increasing the accuracy of the test whereas interested in decreasing the Mean Squared Error.

Decision Tree Regression:

Firstly we are interested in splitting up the dataset into the train test dataset that helps in testing the predictions inside the datset. Here we have a builtin model called the Decision Tree imported from the module sklearn.tree. We are going to fit the data which is to be trained and give insights of the model and predictions so that when it is given a test data without the target variable it should be able to predict the output based on the decision taken from the decision tree. Decision tree not only does the predictions based on the classification but also based on the regression which imports a module from the sklearn. R2 score is likely to test for accuracy where we can check out the proportion of the test data to the total data.

We are actually interested in increasing the accuracy of the test result, on the other hand decreasing the mean squared error(MSE) of the model Decision Tree Regressor.

Polynomial Regression:

Polynomial regression is a special case of linear regression , where we are going to add some polynomial features before performing the linear regression model. Using the scikit learn we are going to do these two steps of adding polynomial features and applying linear regression to it in a pipeline. The polyfit function is the function which is a polynomial interpolation and not polynomial regression. Polyfit is performing a univariate polynomial fit for some vector a to some other vector b. We have performed the polynomial expansion of the feature X-train represented in a higher order polynomial terms for the multivariate fit.

Basically in this model we are transforming the X-train to the n degree polynomial where n is the number of independent variables(continuous attributes) using the function in scikit learn called fit-transform(). After transformed it into the polynomial feature we have applied fit so that it can fit with the output training data.

Then we have tested the test split using the above built polynomial regression with the test data. Applying the same transform and fit function to the test split and computing the accuracy compared to the polynomial featured trained data.

Same as for the other models we are intersted in predicting the higher accuracy between the train and test data and low MSE.

Random Forest Regression:

A random forest is an group method which has two or more(cluster decision trees. Predictions are done by taking average of all decision trees. We will use the random forest regression when we have complex relation between the features and the labels.

Using the scikit learn i.e sklearn.ensemble we are importing RandomForestRegressor. We are calling the RandomForestRegressor with the first argument being the number of forests of the random forest and assigned it to some variable. We will be applying the fit with respect to that variable which has RandomForestRegressor is been called. After fitting the model we have to find the same with the test data and have to find the r2 score(accuracy of test data with respect to train data) and also MSE.

Same as for the other models we are intersested in predicting the higher acuracy between the train and test data and low MSE.

C. Evaluations

We have constructed four different machine learning models related to our dataset and tried to find the best model that would best test the test split data from the train data. We have found out two parameters to have comparisions between the models: r2 score and Mean Squared Error(MSE). The model with the highest r2 score and the lowest MSE is the best model.

IV. Experimental Results

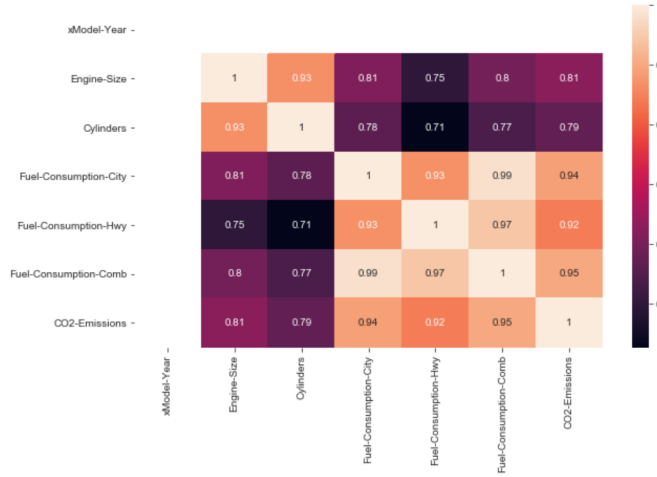
The first and foremost results we are going to talk about are results due to Exploratory data analysis(EDA) and Visualization technique such as scatter plots, Correlation coefficients,probability plots and regression plots.

The scatter plots tell us that all the numerical attributes such as 'Engine-Size','Cylinders','Fuel-Consumption-City','Fuel-Consumption-Hwy','Fuel-Consumption-Comb' are positevly correlation with CO_2 emissions.So now with this insight we can built all the machine learning models which require the target variable to be independent and others to be independent.This graph basically finds out the correlation between the engine size and CO2-emmisions using scatter plot. We can observe that the CO2-emmisions is linearly depending on the engine size. Hence we have to look on an engine with least possible engine size considering all other aspects being efficient for a model(vehicle).

The above fact is also prove by the correlation coefficients calculated between the two or more variables.

The plotted bar charts show that most of fuel type is found to be X and Z. From this histograms we observe here that most models emits CO2 in the rate of 250 which is likely to be eco friendly than higher rates. We should try to reduce to even better than this achieve eco friendly nature by finding out the necessary correlations where CO2 emissions should be at a certain minimum rate and also We can observe here that most of the models engine size lies between 2 and 3 units and We can observe here that most of the models engine size lies between 3 and 4 units.

One of the histograms also show number of unique models in the decreasing order based on the counts.The probabailty plot tells that Data is approximately normally distributed.The curve is the proof for this normal distribution



Heatmaps are utilized to show connections between two variables, one plotted on each axis. By seeing how cell colors change over each axis, you can notice if there are any examples in an incentive for one or the two variables. The above heat map shows the relationship of different variable plotted against each axis.This a very good way to reprent the relationship between the variables.

We have constructed four different models for our dataset and found out the best model that would best test the test split data using the train data. We have used two parameters to have a better comparision between the models i.e r2 score and Mean Squared Error(MSE). We have considered the model with the highest r2 score and the lowest MSE as the best model.

Firstly we built a decision tree regression model using Decision Tree Regressor which is imported from the sklearn(scikit-learn module). We have applied the fit to the X-train and y-train using the DecisionTreeregressor. Same done with the test data to fit the data and have found out the accuracy of the test data(wrt train data).

The r2 score value we have found for this model is 0.996111958169301

The MSE for the model is 0.02

Next we have built the Multiple Linear Regression model using LinearRegression which is imported from the sklearn.linear-model. We have applied the fit to the X-train and y-train using the LinearRegression. We have performed the same procedure for the test data to fit the data and have found out the accuracy of the test data(with respect to the train data).

The r2 score value we have found for this model is 0.9996494557501069

The MSE for the model is 0.0014

Next we have built the Polynomial Regression model PolynomialRegression which is imported from the sklearn.linear-model. We have applied the fit to the X-train and y-train using the PolynomialRegression. We have performed the same with the test data to fit into the model and interested in finding the accuracy of the test data(with respect to the train data).

The r2 score value we have found for this model is 0.9996402566640442

The MSE for the model is 0.0014

Then we have built the Random Forest Regression model RandomForestRegressor which is imported from the sklearn.ensemble(scikit-learn module). We have applied the fit to the X-train and y-train using the PolynomialRegressor. We have performed the same with the test data to fit into the model and interested in finding the accuracy of the test data(with respect to the train data).

The r2 score value we have found for this model is 0.9979848209338302

The MSE for the model is 0.012

Based on the results obtained from the above model, wehave got the higher r2 score and the lower MSE is for Multiple Linear Regression(MLR)[r2 score:0.9996494557501069, MSE:0.0014] is assumed to be the best model for our datset to perform testing of new data with the similar kind of scenario or for a source based on our dataset given.

The above written models work well when the dependent attribute in the dataset have continuous values in it(Regression analysis).

The above models which have been implemented for our datset fails when it is asked to classify the CO2 emissions based on the order like good features, medium features and bad features.

V. Conclusions

A. Contribution of each team member

Literature review + initial solution approach : Paper1 was done by Bhargav V and second paper was done by Ashay G and third paper was done by Manoj Kumar P Ashay G and Bhargav V.

Model design and testing

- Design model/ refine model parameters - Run cross validation tests and make a note of the results - Run comparisons, test any other models that need to be tested Wrap up the model building/ testing and work on presenting and interpreting results.

Manoj Kumar P and Naveen C S : - Preprocessing of data - Exploratory data analysis(EDA) and Visualization - Preparing other materials for the project - Record the video presentation

The research paper was collectively by all members of the team.The solution approaches to the problem was also equally contributed by the team.

References

[1] Fuel Consumption Models Applied to Automobiles Using real time data Ahmet Gurcan Caparaz,Pinar Ozel,Mehmet Sevkli,Omer Faruk Source : SEIT Year : 2016

[2] Global Research on Carbon Lebunu Hewage Udara Willhelm Abeydeera ,Jayantha Wadu Mesthrige 2 and Tharushi Imalka Samarasinghalage Source:Department of Building and Real Estate from hong kong polytechnic university Year : 2019

[3] Emissions and Fuel Consumption Modeling for Evaluating Environment Effectiveness of ITS Startegies Yuan-yuan Song,En-jian Yao,Ting Zuo and Zhi-feng Lang1 Year : 2014