Fuel consumption rating

Abstract:

Fuel consumption has implications to climate change and the general environment hence the need for accurate prediction and classification of motor vehicles. This report aims to examine the impact of fuel consumption by analysing a dataset spanning from 2010 to 2014, with the objective of elucidating the effects observed during this time frame. The objective of this study is to forecast the carbon dioxide emissions of automobiles and categorise them according to their categorical attributes.

Introduction:

In light of contemporary efforts to mitigate global warming by reducing carbon emissions, there is growing apprehension regarding fuel consumption, given its significant contribution to the undesired carbon content in the atmosphere. According to Lindsey (2021), the majority of the warming effect caused by anthropogenic greenhouse gases can be attributed to carbon dioxide, which accounts for approximately two-thirds of the total. The primary source of these gases is fossil fuels.

The present report aims to offer an in-depth analysis of the potential application of machine learning techniques in predicting the CO2 emission of vehicles and categorising them based on their respective categorical variables.

1.0. Steps required to train a machine learning model.

1.1. The Collection of Data

The initial stage in the process of training a machine learning model involves the collection of data that is pertinent to the specific type of model that is to be trained and the particular problem that the model intends to solve. The efficacy of the machine learning model is dependent upon the data it is trained on, Therefore, it is imperative that the data provided is trustworthy to ensure optimal performance of the model. The accuracy of the model is contingent upon the quality of the data that is inputted into the machine.

This concerns the process of rectifying any flaw or inconsistencies present in the dataset, commonly referred to as data preparation. The limitations of the dataset comprise of missing values that require imputation with reasonable values, elimination of replicated entries that could be superfluous to the machine learning model and may lead to an increase in training time with no discernible advantage, and the reorganisation of the dataset.

1.3. The Exploratory Data Analysis

This involves conducting preliminary investigations on data in order to identify patterns, anomalies, verify hypotheses, and assess assumptions through the use of summary statistics and graphical depictions.

1.4. Selecting a model

Numerous machine learning models are available for various tasks. The objective is to identify the most appropriate option or options that are well-matched for the given problem.

1.5. Training of the model

During the training phase, the machine learning model is provided with preprocessed data and it tries to determine optimal weight and bias values based on labelled examples. The model tries to discern regularities within the data and generate a function that can effectively map forthcoming test data to a precise forecast or prediction.

1.6. Model evaluation

In order to assess the efficacy of the model, it is imperative to conduct an evaluation on data that has not been previously utilised in the training phase. It is done in order to measure or check the model's ability to generalize and make accurate predictions on unseen data.

1.7. The process of hyperparameter tuning or optimisation

This is the process of finding the optimal values for the hyperparameters of a machine learning model. Hyperparameters are parameters that are set before training the model and affects its behaviour and performance.

1.8. The Process of Making Predictions and Deployment

After building the model and optimising it to its optimal performance, it is subsequently employed for the intended purpose of prediction, which was the primary objective of the process.

Methodology

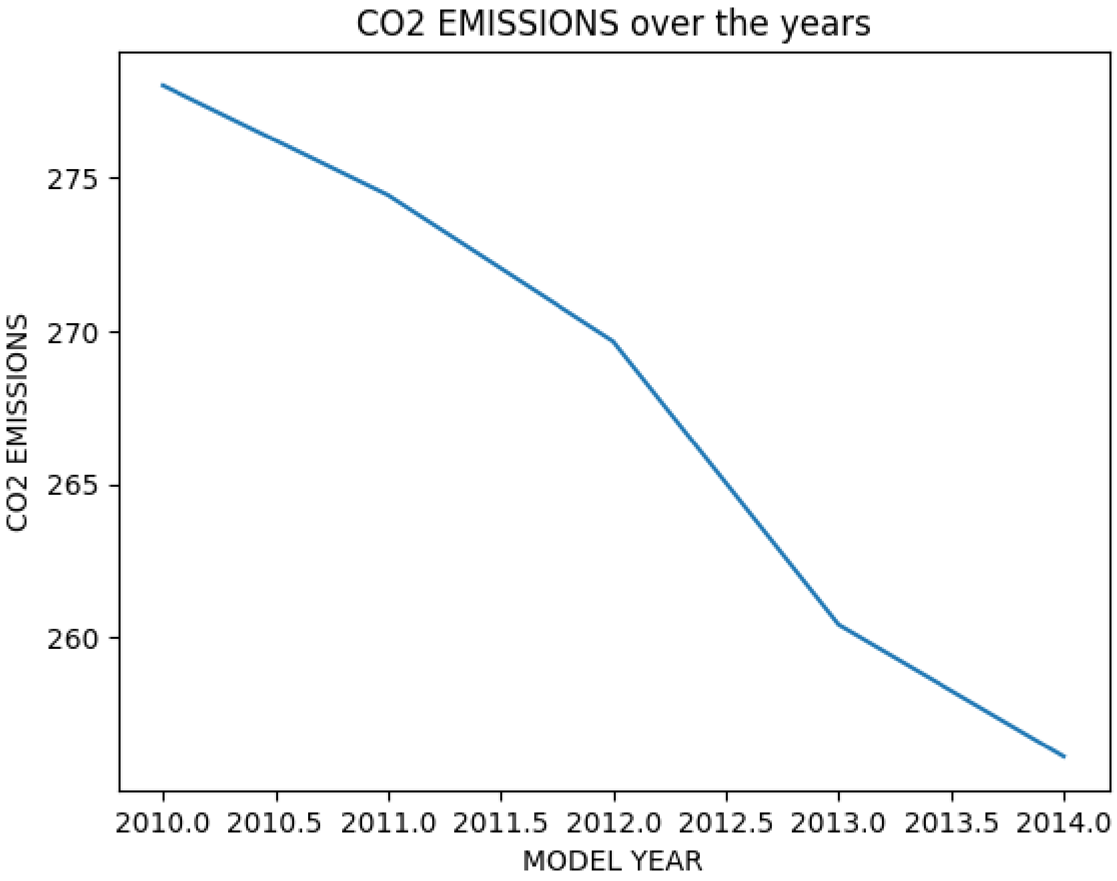
2.0. Building models using continuous variables

A predictive model was constructed for CO2 emissions based on each continuous variable within the dataset spanning from 2010 to 2014. The variables under consideration in this study were Model year, Engine Size, Fuel consumption, HWY (Highway consumption), and Comb (combined rating of consumption, highway and city). Based on the metrics utilised to evaluate the models, it was determined that Model Year exhibited the poorest performance, as evidenced by a mean absolute error of 50.81 and a root mean squared error of 63.52. The attribute that exhibited the most favourable performance was Comb, as evidenced by its mean absolute error of 18.11 and root mean squared error of 28.33.

Upon conducting exploratory data analysis on the continuous variables, it was determined that the Model Year attribute did not significantly contribute to the identification of CO2 Emissions. Consequently, the decision was made to eliminate this attribute from the analysis. Utilising the residual attributes, a machine learning algorithm was constructed that exhibited favourable performance in comparison to those models derived solely from singular attributes. The statistical evaluation of the model yielded a mean absolute error of 15.25 and a root mean squared error of 23.51.

3.0. Checking for improvement in CO2 Emissions from 2010 - 2014

A line plot graph was created to display the relationship between Model Year and mean CO2 Emissions from 2010 to 2014. The graph indicates a decreasing trend in CO2 Emissions over the years. The average value of CO2 emissions was recorded as 278.02 g/km in 2010, which was reduced to 256.12 g/km by 2014.



4.0. The optimal variables for classification.

It was determined that the variable, Fuel type exhibited the highest classification accuracy of the dataset, with an average of 75 percent accuracy following optimisation. The second highest percentage was Transmission, accounting for 33 percent, while the lowest percentage was observed in Model, constituting merely 5 percent.

5.0. Checking for overfitting

The model's accuracy score on the unseen test data that was relatively high, indicating that the model did not exhibit overfitting.

6.0. Performance measures that were utilised.

For my regression, I utilised the performance metrics, r2 score and root mean squared error. This is because these measures help to assess the degree of deviation between the model's predicted values and the actual values.

The performance measures employed in my classification task included accuracy, recall, f1-score, and precision. The metric of accuracy provides a concise representation of a model's performance in terms of correctly classifying data, yet it is insufficient in conveying a comprehensive understanding of the model's overall efficacy. Hence, the recall metric was employed to determine the proportion of positive classifications, alongside other relevant metrics.

7.0. **Whether the models can be deployed**

The regression model is deployable due to its high performance on the test data, as evidenced by a r2 score of 0.87. The classification model achieved a moderately high score, however, I feel it's unsuitable for deployment because accuracy level of 75 percent is not really great.

8.0. The categorical variable that describe the groups best

The categorical variable that described the groups best was Transmission, as it demonstrated the lowest Davies Bouldin score of 4.28.

Conclusion

The process of building a model was outlined. The study undertook a review of the methodology employed in performing linear regression for the purpose of predicting CO2 emissions, as well as the process of creating clusters.