COMPONENT 2
PREDICTING CARBON MONOXIDE EMISSIONS

Abstract

To predict carbon mono-oxide emitted from motor vehicles, the supervised machine learning algorithm: regression and classification were employed. This helped in training a model that can possibly predict these emissions, classify the dataset, and form groups of clusters.

Results were interpreted which helped us determine if these models should be deployed or not.

Introduction

One of the effects of climate change is the emission of carbon monoxide from fuel consumed vehicles into our natural environment. Hence, a need for correct prediction and classification of motor vehicles. Sometimes, it is challenging to detect hidden trends from a data but, thanks to machine learning algorithms that has made it possible for models to trained. This helps these models to learn on how to use built-in information to keep making decisions on how these problems can be resolved.

STEPS REQUIRED IN TRAINING A MODEL

Data Collection

The data to be collected is dependent on the problem one wants to solve. Sometimes, data can be available in databases or gathered from different API sources.

Data Cleaning and Pre-processing

This is a very crucial part in training a model. Data is loaded and it is advisable that a copy of it be made to retain original details. Noticeable errors such as missing values, outliers, null entries are revealed and resolved if present. The data is also being explored and visualized using different graphs and charts. The relationships existing between the variables are revealed.

Data cleaning and pre-processing helps to improve the overall quality of data and it is the stage where all the core preparations are being done; ‘dividing the dataset into numeric and categorical’, ‘data visualizations’, ‘splitting data into training and testing sets’, ‘Encoding and feature scaling’ (Yufeng, 2017).

Splitting Data into Training and Testing Sets

The data is further split into two sets- the training and testing sets. The training set is one used to train the model and to test the accuracy of the models, the testing set is evaluated (Stack Overflow, 2022).

Encoding and Feature Scaling

The categorical variables were encoded, and the numeric data were scaled. Encoding is essential as it helps to convert categories into numbers, a format the computer understands. Feature scaling is simply ‘Normalization’.

Choosing and Training a Model

This is the process of choosing relevant models based on the goal one may want to achieve. The best model is chosen after cross validation with an evaluation metrics. The dataset is then trained on the chosen model to carry out the specified task, find patterns and make predictions.

Model Evaluation and Hyperparameter Tuning

After training the model, the test set is evaluated to check model performance. Grid search cross validation is done on the selected model with some hyperparameters to determine best features.

This is where we introduce the test set which the model hasn't seen before to understand how much the model has learnt and how well it will perform. Finally, predictions are made with the test set data.

FUEL CONSUMPTION RATING

The 'Fuel Consumption Rating' dataset was loaded into the Jupyter Notebook in csv format. During data exploration, there wasn't need for data cleaning as no errors were identified. Histograms, pair plot and correlation matrix were used to show the distributions and variables' relationships. The variables were casted to their respective data types and divided into seven numeric and six categorical variables.

Regression

In building regression model, we attempted using all numerical continuous variables (independent variables) to build a model to predict CO2 Emission (target variable). Both were scaled using

'Mean Normalization' (MinMaxScaler) which aims at using the mean of observations to transform and preserve the shape of the original distribution (Nishesh Gogia, 2019).

First, after considering all numerical variables, subsets were selected with the help of the correlation matrix. I cross validated the multiple linear and decision tree regression models, using the coefficient of determination (r2_score).

NB: Mean squared or Root mean squared error can be considered as well.

Model comparisons were made at each instance of feature(s) selection. Centered on the highest r2_score, the decision-tree regression model was selected with the two features (Engine Size and FC_COMB (mpg). Hyperparameter tuning was conducted, and the model was evaluated using the test data.

Comparing the model performance, findings showed that the model performed better with all numerical variables. Nevertheless, due to high correlations among these variables, the model of the subsets was preferred and said to have predicted CO2 Emissions well.

The bar plot below shows that there is little or no decrease in C02 emissions over the years and as such, no noticeable improvements.

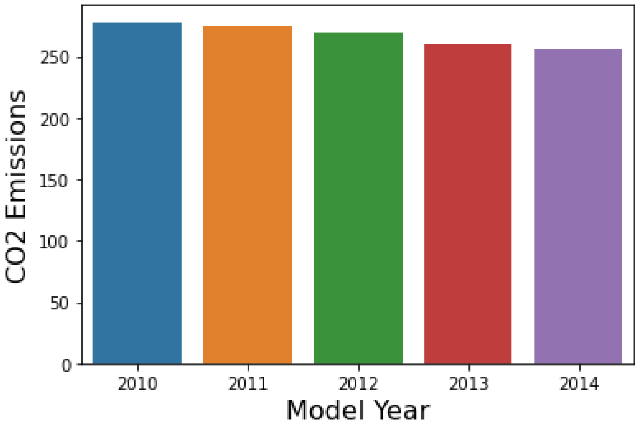

Fig1: Bar plot showing C02 Emissions over the years

Classification

Each categorical variable was selected as a target variable to determine which one best classifies the data. Except for 'Fuel Type', all other categories either due to large features or low accuracy, performed poorly at classification. The model selected is the decision tree classifier, based on evaluation matrix (accuracy = 0.99). Fuel type classified 96% of the data correctly.

Checking Overfitting

To check for overfitting, the accuracy of the train and validation test were evaluated. We can say that if the performance of the model on the train is better than the test, the model is likely to overfit. But the model was ross validated and this check for overfitting. Accuracy was chosen as the performance measure, as it best describes the percentage level of true predictions of our model.

Clustering

Groups were formed with the numeric data using the internal and external evaluation metrics, 'Davies Bouldin' and 'Normalized Mutual Info Score', in conjunction with KMeans clustering model.

First, the model was fitted on the scaled numeric data, sum of square distance and Davies Bouldin score was calculated and plotted against the number of clusters.

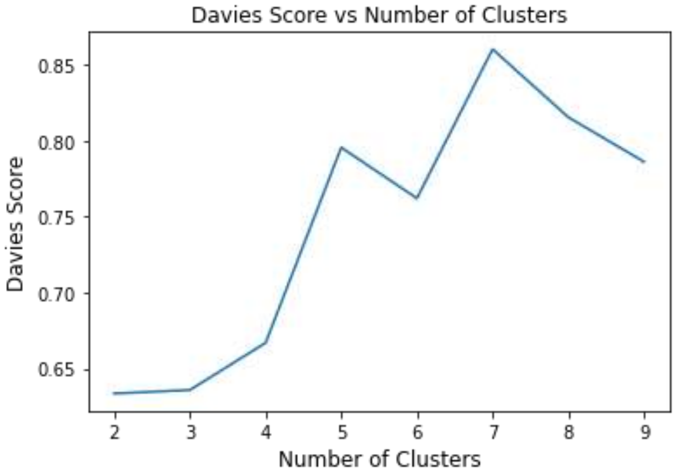

Fig2: Showing number of clusters to score

The lower the Davies Bouldin score, the better the model. On this basis, clusters 2,3,4 were chosen, preferably (K=4). The model was fitted again with K=4 and groups were formed.
Secondly, the normalized mutual score was calculated using the KMeans cluster labels on all the categorical variables.
Results of both metrics showed that both 'Fuel Type' and 'Model' best describes the groups.

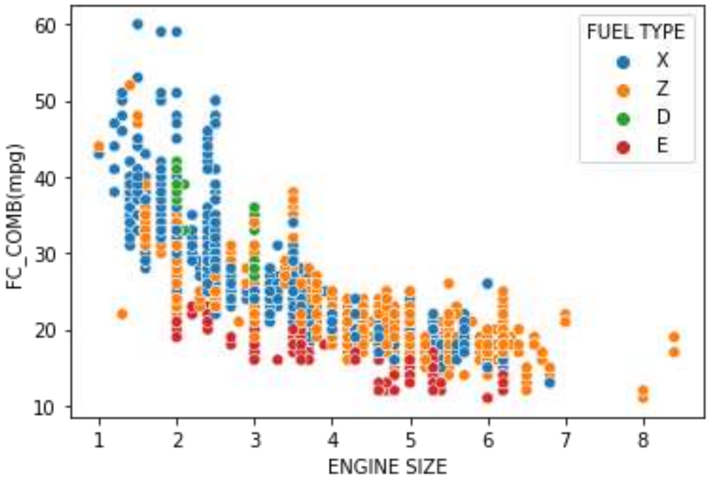

Fig3: Plot showing how Fuel types classifies data

In conclusion, training a model is of great importance as it makes life easy and aids in problem solving. Predicting CO_2 emissions will help save the environment and our world at large.

Reference

Nishesh, G. (2019), *Why Scaling is Important in Machine Learning?* Available Online: Why Scaling is Important in Machine Learning? | by Nishesh Gogia | Analytics Vidhya | Medium [Accessed: 14/12/2022]

Yufeng, G, 2017. *The 7 Steps of Machine Learning.* Available Online: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e [Accessed: 12/12/2022].

Simplilearn, Mayank Banoula (2022). *Machine Learning Steps: A Complete Guide!* Available Online: https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps [Accessed: 13/12/2022].

Stack Overflow (2022). *What are the Basic Steps for Training a Model?* Available Online: https://stackoverflow.com/questions/43883723/what-are-the-basic-steps-for-training-a-model [Accessed 14/12/2022].

CFI Team (2020). *Overfitting.* Available Online: Overfitting - Overview, Detection, and Prevention Methods (corporatefinanceinstitute.com) [Accessed 14/12/2022].