DATA 1030 Final Report

Predicted CO2 Emission by Vehicles

Haiting Huang

Professor:Andras Zsom

Affiliation: DSI Brown

Github Repository:

https://github.com/HaitingHuang/DATA-1030Final-Project.git

Introduction

In recent years, low-carbon travel has been one of the mainstream topics in the world. The data analyzed in this report is CO2 Emission by Vehicles[1] collected from Kaggle, which took and compiled the original data source from the Canadian Government official site.

Previous citing of this dataset is included in *Hypothesis Test, Correlation and Linear Regression (Zilma Bezerra[2])*, which includes building a linear regression model using Numpy Polyfit with an accuracy 0.729. Moreover, an article titled *CO2 Emission EDA/RF Hyperparaeter Optimization (Tom Bache[3])* used this dataset to predict CO2 emissions based on the vehicle characteristics, and got RMSE of 14.9 is about a 6% relative error.

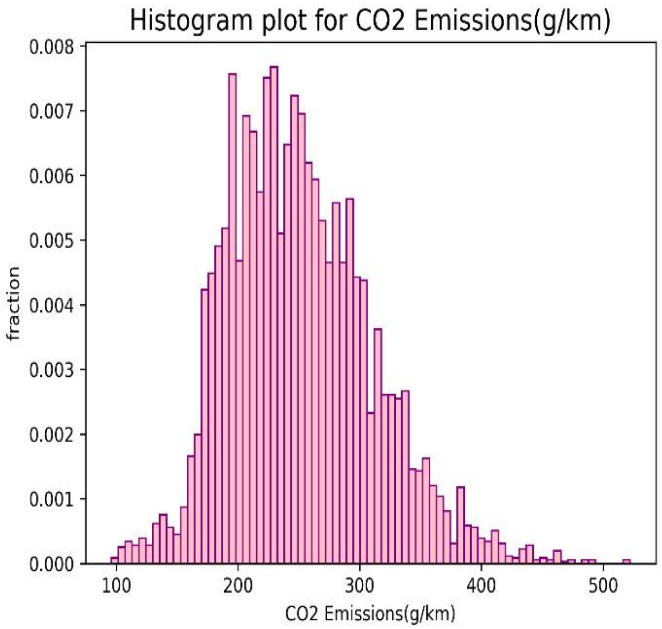
This data has 7,385 rows and 12 columns. After removing all duplicated observations, there are 12 features with 6282 data points left without any missing values. The categorical features in this dataset include Make, Model, Vehicle Class, Transmission and Fuel Type. The continuous features with their units are Engine Size (L), Cylinders (number of cylinders), Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), and CO2 Emissions (g/km), which are all in floats or integers.

Among all the features listed above, CO2 Emissions(g/km) will be treated as the target variable in this report. The rest of this article will introduce the regression method of using other features to predict CO2 emission, which can help consumers to choose more environmentally friendly and desirable cars as travel tools, and even help protect the earth's environment and reduce the greenhouse effect.

---

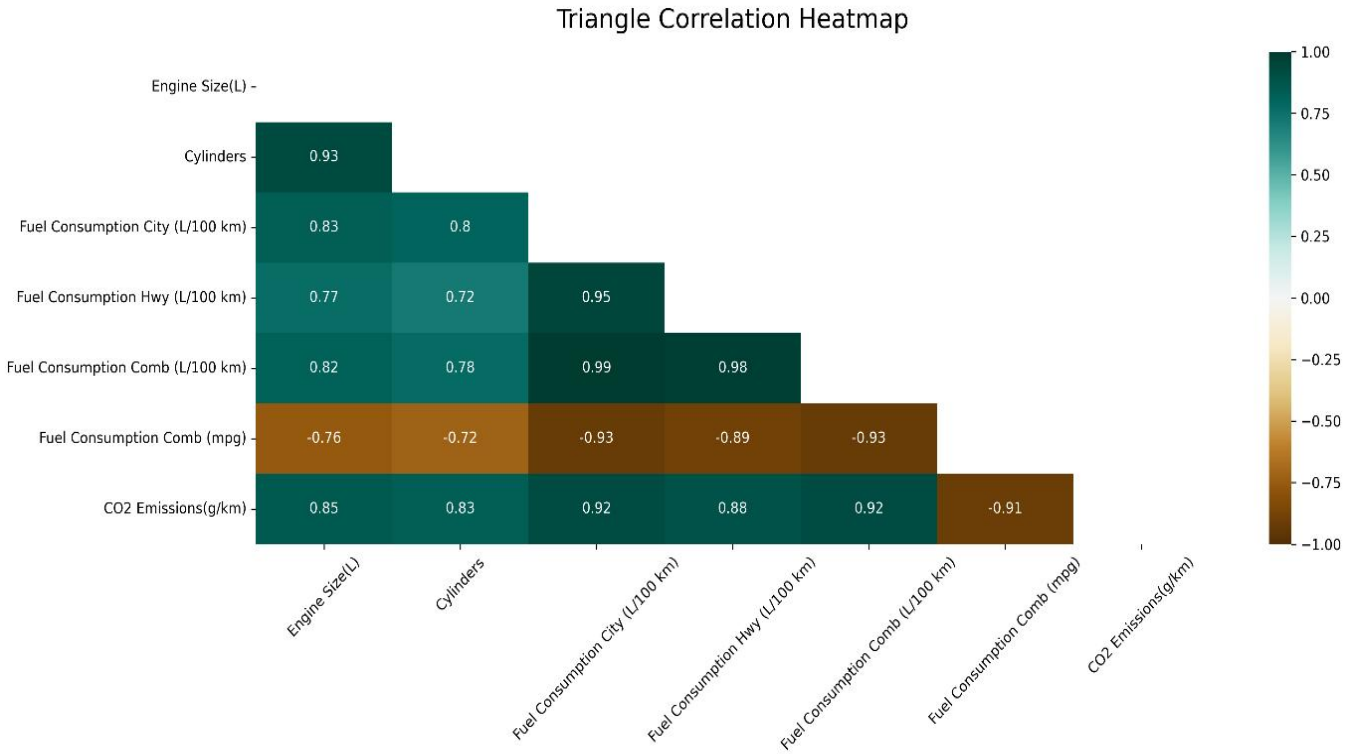
[1] https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles
[2] https://www.kaggle.com/code/zilmabezerra/hypothesis-test-correlation-and-linear-regression
[3] https://www.kaggle.com/code/tombache/co2-emissions-eda-rf-hyperparameter-optimization

Exploratory Data Analysis



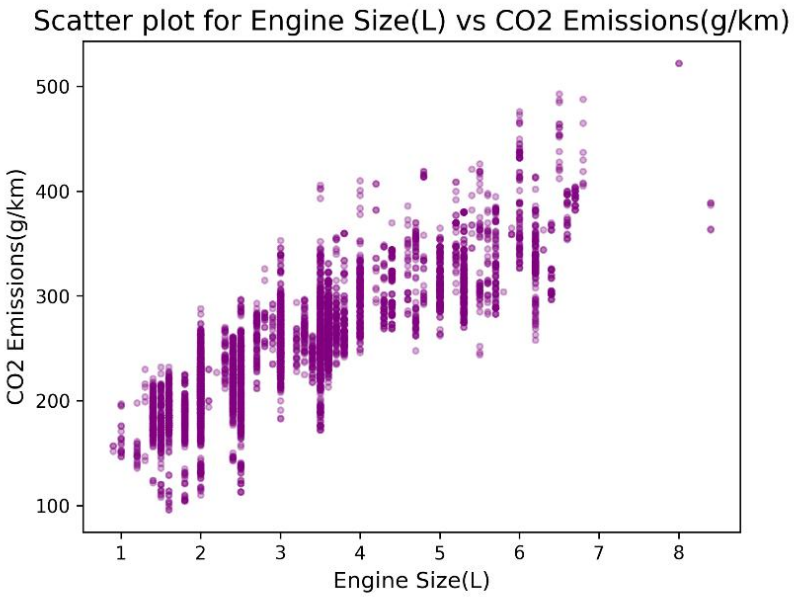
This figure shows a histogram of CO2 Emissions(g/km), and it approximately shows like a normal distribution

The histogram shown on the top indicates the distribution of the CO2 Emission(g/km) which is the target variable for this dataset. As shown in this figure, the distribution is approximately normal with a slightly right skewed with magnitude 1 which means there is not much difference between the maximum and the minimum, and there is a outlier appear over 500 from this histogram.



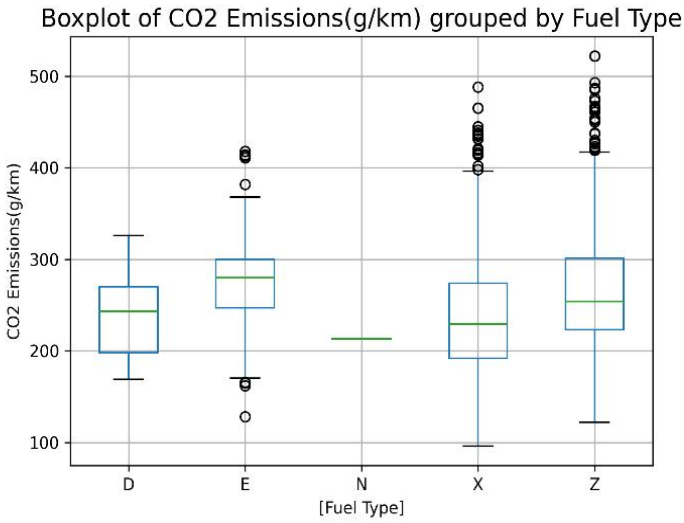
We can find Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), and Fuel Consumption Comb (L/100 km) were highly correlated from this correlation heatmap

The heat map shown on the top indicates correlations between features. As stated in the figure, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), and Fuel Consumption Comb (L/100 km) was highly correlated, which the correlation was more than 0.95, Therefore, Fuel Consumption City (L/100 km) and Fuel Consumption Hwy (L/100 km) are dropped for later preprocessing step, and kept Fuel Consumption Comb because it is the combination of these two features which have been dropped.



This figure shows a strong positive correlation between Engine Size(L) and CO2 Emission(g/km)

The scatter plot shown on the top captures a strong positive correlation between Engine Size and CO2 Emission. This is because when the engine size increases, more fuel will be needed to activate the engine, which leads to higher CO2 emission.



This figure shows different Fuel Type correspond to different CO2 Emission(g/km) that condense in different levels

Moreover, the boxplot on the top shows the relationship between Fuel Types and CO2 Emission. The figure indicates that CO2 Emission levels of five different type of fuels is concentrated between 200 and 300 g/km. The information of CO2 Emission with Natural Gas is insufficient due to data collection, so a complete box cannot be drawn, but from the boxplot the mean of CO2 Emission of Natural Gas is the least among the five fuels. Thus, if cars use natural gas as their main fuel, they may be associated with reducing the greenhouse effect and air pollution.

Methods

Because each row of data is independent and follows the same distribution, they can be treated as i.i.d. variables, so no group structure is needed. Besides, since CO2 emissions do not vary with time, it is not a time-series dataset. After making a series of EDA analyses, the result shows that CO2 emissions can be predicted using other features by regression. Since the dataset is not big enough, kfold split should be better applied, which makes the results more generalized and accurate than those of basic split. Considering the relatively small dataset, the train and test sets are separated into the percentages of 80% and 20%, so that the model will have enough train sets to learn, and enough test sets to ensure its accuracy and precision. For continuous features, StandardScaler instead of MinMaxEncoder is applied, because these continuous features do not have a range that satisfies the MinMaxEncoder requirement. For categorical features, OneHotEncoder rather than OrdinalEncoder is applied because it can transform categorical features to dummy arrays that provide better observations, and the categorical features in this dataset do not have ordinal levels. In order to simplify categorical features with a small number of observations, the eight most frequent categories were kept and the others were considered infrequent categories for each input feature. The features of the make and model of a car were removed because we'd like to focus on the generalized characteristics of a car. After applying OneHotEncode for categorical features and StandardScaler for continuous features, there are 24 columns in the modified dataset.

After data processing, a grid search of model parameters is conducted on the 5-fold split of training set to find the best model parameters. The machine learning pipeline repeats this procedure 10 times to find the average RMSE and its standard error on the validation split. The model performance estimate from a single run of the k-fold cross-validation technique might be noisy because results from various data splits might be extremely different. Repeated k-fold cross-validation offers a means to improve an estimated machine learning model's performance.

RMSE is an important metric for measuring the accuracy of predictions in CO2 emissions. RMSE is preferred over other metrics such as mean absolute error (MAE) because it penalizes large errors more than small ones. This makes it a better metric for assessing the accuracy of predictions in CO2 emissions, since it is important to accurately predict both high and low emissions.

The use of repeated 5-fold cross validation is a powerful tool for measuring the accuracy of a predictive model. However, it is important to understand that there is an inherent uncertainty associated with this technique. This uncertainty is because the accuracy of a predictive model depends on the subset of data used to train and validate the model, and this subset of data can vary significantly from one iteration of 5-fold cross validation to the next. We report the mean RMSE and its standard error as it is anticipated to provide a more accurate representation of the actual, unobserved performance of the model on the dataset.

Table 1. The machine learning algorithms and the corresponding tuning parameters and values we use to fit the data dataset.

Model	Tuning Parameter	Tuning values
linear regression with l1 regularization	alpha	-5 to 3 spaced evenly on log scale
linear regression with l2 regularization	alpha	-4 to 4 spaced evenly on log scale
linear regression with an elastic net	alpha	-4 to 4 spaced evenly on log scale
	L1 ratio	0.1 to 1 spaced evenly
Random forest	Max depth	1, 3, 10
	Max features	0.1 to 1 spaced evenly
K Nearest Neighbors	n_neighbors	3, 5, 10, 20, 50, 100

The table shows the machine learning algorithms we use to fit the data dataset. It also lists the tuning parameters and the values we use. We use three linear regression algorithms with difference regularization methods and two non-linear regression algorithms. Each model is tuned by multiple parameter values.

Results

The baseline model is using the mean CO2 emissions in the dataset as the estimated CO2 emissions. The RMSE of the baseline model is 60.07264. The following table shows the average RMSE score and its standard error for each machine learning algorithms using the 5-fold cross validation repeated 10 times.

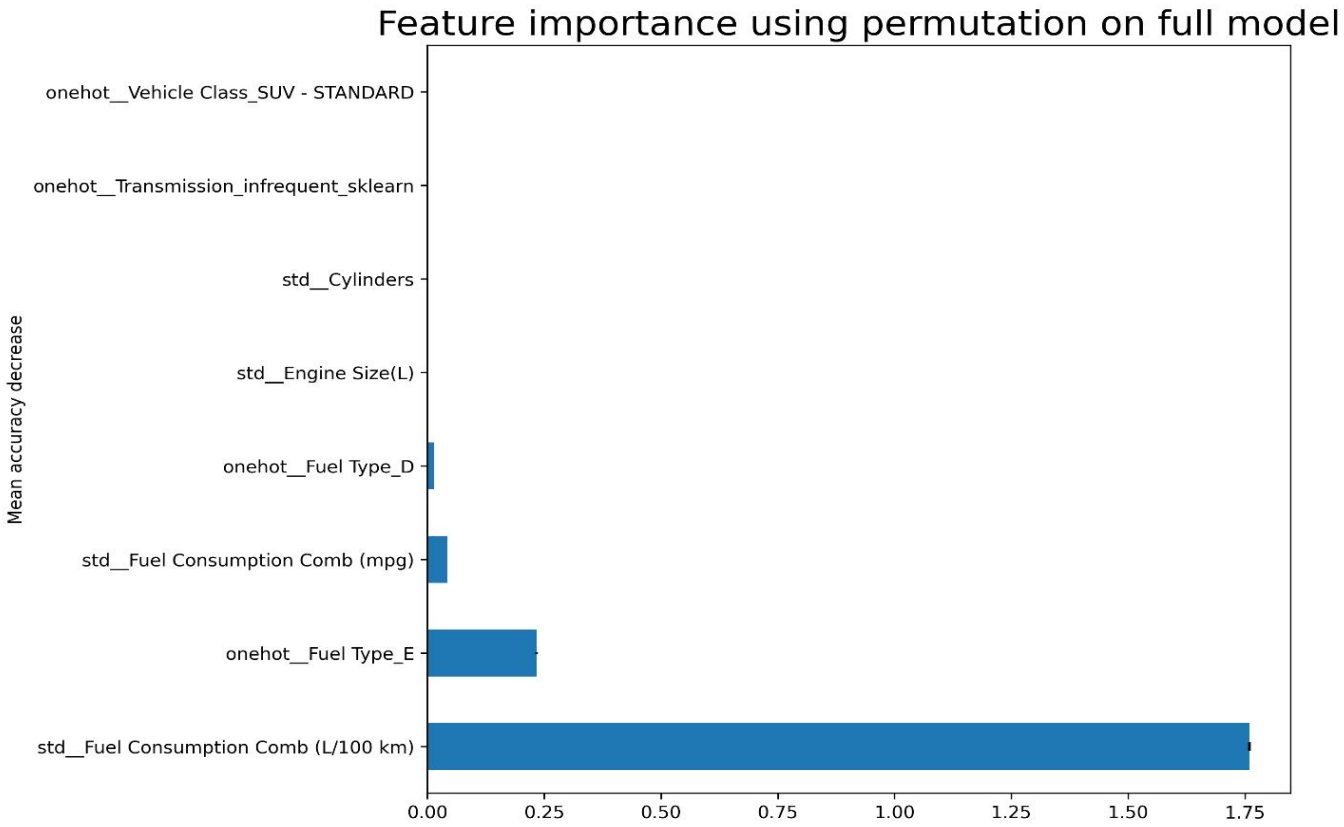
The mean RMSE validation score for random forest is 3.88230, with a standard deviation of 1.47808. As a result, the random forest model is the best algorithm on the dataset. Moreover, all linear regression algorithms, including the lasso (Mean RMSE: 5.85688, STD: 1.39655), ridge (Mean RMSE: 5.67243, STD: 1.18000), and elastic net (Mean RMSE: 5.85659, STD: 1.39651), have comparable average RMSE validation scores. It's possible that the feature variables and the desired outcome have a non-linear relationship, which would make it easier for a non-linear regression algorithm such as the random forest algorithm to provide more accurate predictions. The K Nearest Neighbors algorithms has the highest mean RMSE validation score (Mean RMSE: 6.94699, STD: 1.38358) among all used models, suggesting it does not work well when sample size is small and dimension is high.

Table 2. The machine learning algorithms' performance on test data. Based on below table you can know which model is the best one fits future steps.

Model	Mean RMSE	STD RMSE	STD RMSE under baseline
Baseline model	60.07263	4.26820	
linear regression with l1 regularization	5.85688	1.39655	12.70225

linear regression with l2 regularization	5.67243	1.18000	12.74547
linear regression with an elastic net	5.85659	1.39651	12.70232
Random forest	3.88230	1.47808	13.16488
K Nearest Neighbors	6.94699	1.38358	12.44685

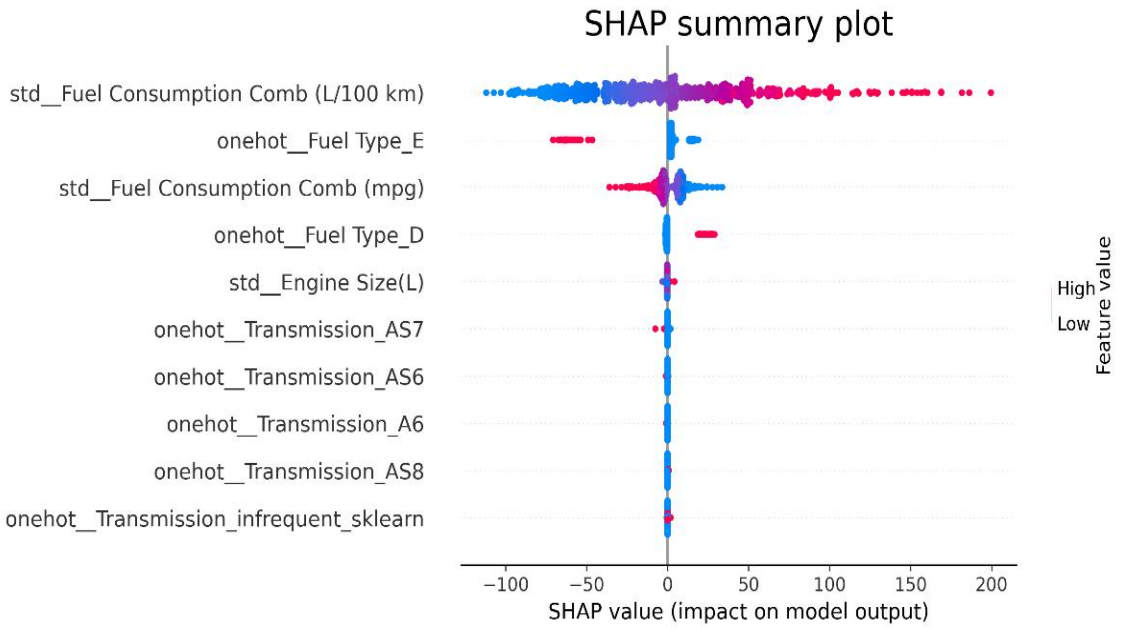
The random forest model is used as the final model because it performed best on the training dataset. On the test dataset, the RMSE is 3.88230. The RMSE is within the 99% confidence interval obtained through the training process, which indicates the cross-validated result is reliable. It suggests that the model is not overfitting and is generalizing well to new data. The R-squared is 0.99582 and the mean absolute error is 2.29316 on the test dataset. These additional evaluation results show the prediction on the test dataset has a high accuracy, so the concern of underfitting is negligible.



This figure shows feature importance using permutation on the model. Fuel consumption (L/100 km) fuel type of Ethanol, fuel consumption (mpg), and fuel type of Diesel and are most important.

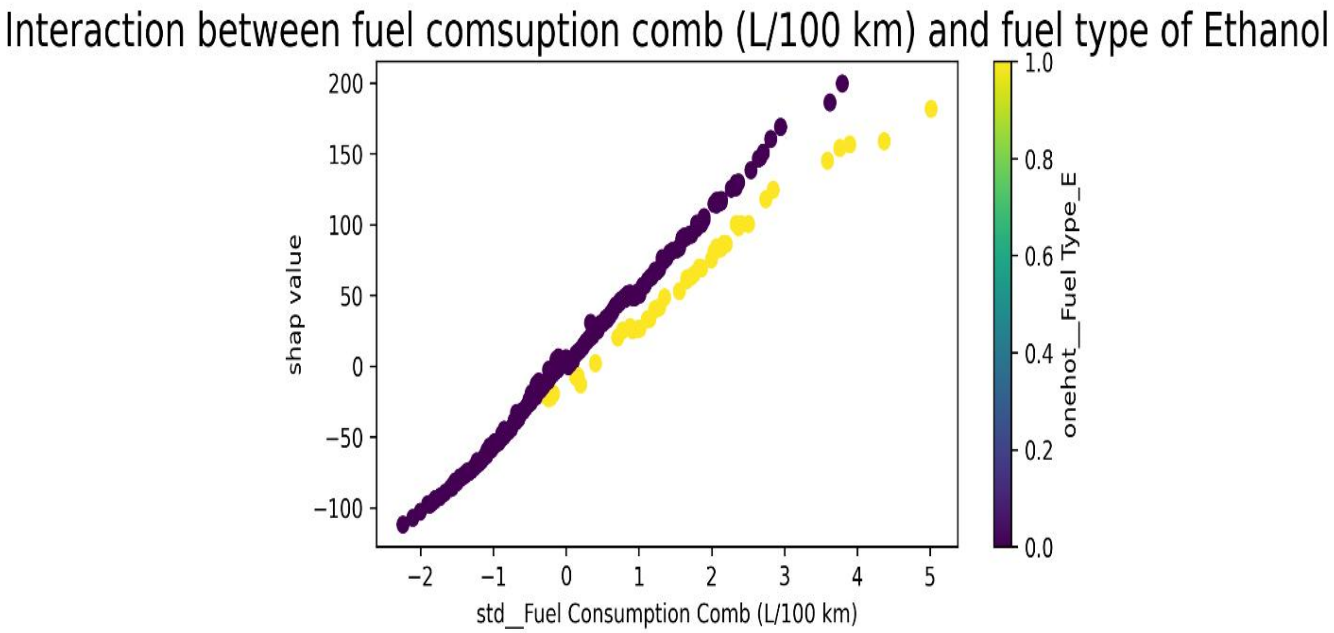
The permutation feature significance gauges how much the model's accuracy to predict outcomes has declined after we randomly permuted the feature's values ten times. In the given plot, the fuel consumption (L/100km) has the highest feature importance, followed by the fuel

type indication (Ethanol), the fuel consumption (mpg), and the fuel consumption (L/100km) (Diesel). The importance of other features is neglectable.



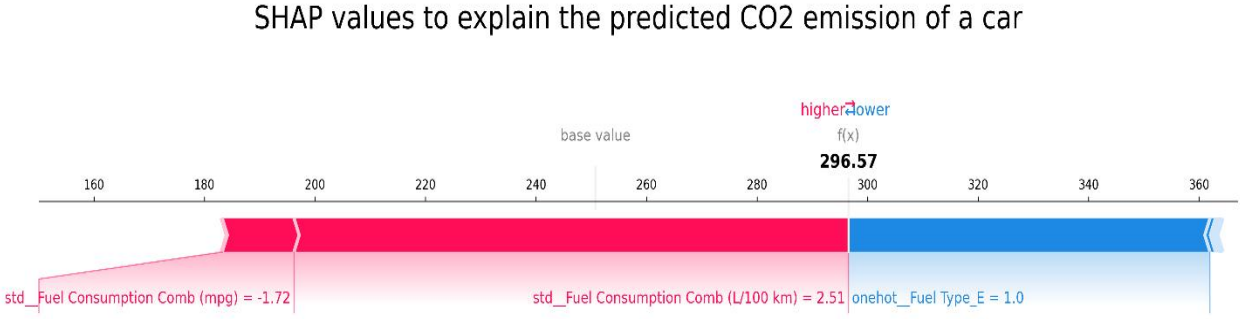
This figure shows a summary of SHAP values by all features,
and the standardized fuel consumption (L/100 km) affects the SHAP values on output the most.

The Shapley value for a feature value shows its contribution to the prediction of a given instance in contrast to the average prediction for the whole dataset. The feature importance of Shapley values is more related to the feature's influence on the model prediction. This plot shows us the impact of features on the prediction of the model, ordered by mean absolute Shapley value of each feature. The fuel consumption (L/100km) influence the prediction of the model the most, followed by indicator of fuel type of Ethanol.



This figure shows interaction between fuel comsuption comb (L/100 km) and fuel type of Ethanol, and cars using Ethanol reduce the SHAP values when holding fuel consumption constant.

There is a clear interaction between fuel consumption (L/100 km) and fuel type (ethanol). The fuel consumption shows two different distributions of Shapley values, which almost completely depend on the fuel type of ethanol. Fuel consumption has a strong linear relationship with predicted CO2 emissions and using ethanol as a fuel reduces CO2 emissions.



This figure shows SHAP values of the 100th car, fuel consumption (L/100 km) and fuel consumption (mpg) increase the output while fuel cars using Ethanol reduce the output.

The SHAP value can explain an individual car model's predicted CO2 emissions. The average predicted CO2 emission is around 250. Standardized fuel consumption in mpg of -1.72 and standardized fuel consumption (L/100 km) of 2.51 increase the car's predicted CO2 emission, while using the fuel type ethanol reduces the predicted CO2 emission.

Outlook

This study aims to explore the use of predictive analysis in cases of CO2 emissions. We find the best machine learning model to be the random forest model in this case. The RMSE of the model on the test data set is 3.8825. Furthermore, we used explainable machine learning methods, including the permutation feature importance and the Shapley value, to extract and interpret information from the random forest model.

In order to improve the model's performance, more thorough parameter tuning could be conducted. There are several alternative random forest algorithms focusing on increasing the strength of the individual trees in the forest and decreasing the correlation between trees to improve performance. It might be helpful to explore these alternative implementations of RF models.

One limitation concerning random forest is that it has a "black box" algorithm. Hundreds of decision trees are randomly generated and entered into the model. The interpretation of its decision process is still extremely difficult, although the decision trees can be visualized.

Reference:

1. CO2 Emission by Vehicles, Debajyoti Podder. 2020, October 20. https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles
2. Hypothesis Test, Correlation and Linear Regression, Zilma Bezerra. 2022, September 22. https://www.kaggle.com/code/zilmabezerra/hypothesis-test-correlation-and-linear-regression
3. CO2 emissions EDA/RF hyperparameter optimization, Tom Bache. 2022, May 23. https://www.kaggle.com/code/tombache/co2-emissions-eda-rf-hyperparameter-optimization