

Faculty of Engineering

Industrial Engineering Department

Ferdowsi University of Mashhad

Authors: Mohammad Jamalzehi, Kosar Ghazali

Contact: mohammadjamalzehi.aca@gmail.com

kosar.ghazalii@gmail.com

Title: Predicting Carbon Dioxide Release Using Regression Analysis

Table of Contents

Introduction.....	3
Literature Review.....	3
Data Description	4
Exploratory Data Analysis.....	5
Preprocessing	10
New suggested method	10
Normalization	11
Data Splitting	11
Regression Analysis.....	11
Linear regression.....	11
K-nearest neighbors (KNN).....	11
Decision tree regression.....	11
Random forest.....	12
Lasso regression.....	12
Ridge regression.....	12
Elastic net regression	12
Gradient boosting regression	12
Evaluation Metrics	13
Conclusion	13
Bibliography	13

Introduction

The importance of CO₂ emissions, especially those caused by fuel consumption from vehicles, cannot be overstated. Carbon dioxide is a major contributor to global warming and climate change; reducing the amount of CO₂ emitted into the atmosphere can help reduce these negative effects. Analyzing fuel consumption and types can provide solutions to minimize carbon dioxide emissions in both short-term and long-term ways.

One effective approach for minimizing CO₂ emission is through statistical analysis such as regression analysis or other machine learning methods that predict how much carbon dioxide will be released based on certain factors like vehicle type and fuel type.

This data allows researchers to better understand which strategies are most successful at reducing overall pollution levels while maintaining efficiency in transportation systems.

Additionally, predictive analytics also allow governments and businesses alike to identify areas where further investment should go towards improving air quality standards across their respective jurisdictions or industries respectively

Finally, implementing regulations that require more efficient technologies when it comes to combustible engines would also help reduce overall emissions rates significantly over time as well as incentivizing drivers with rebates for using cleaner fuels such as electric cars rather than gasoline powered ones could have a considerable impact on lowering our collective carbon footprint worldwide if done correctly.

Literature Review

Several studies have identified various factors that influence CO₂ release, including economic growth, energy consumption, population, industrial activities, and transportation. Regression analysis allows for the identification and quantification of these factors' impact on CO₂ emissions.

Multiple regression analysis is commonly employed to predict CO₂ release. Researchers often include independent variables such as GDP, energy consumption, fossil fuel consumption, and population size. Some studies have also utilized time series analysis to account for temporal variations in CO₂ emissions.

The accuracy of CO₂ release predictions using regression analysis varies across studies. Factors such as the quality and availability of data, model specifications, and inclusion of relevant variables influence the accuracy. However, most studies have reported reasonably accurate predictions, with regression models explaining a significant proportion of the variation in CO₂ emissions.

Data Description

This dataset captures the details of how CO2 emissions by a vehicle can vary with different features. It has been taken from Canada Government's official open data website and contains data over a period of 7 years, with 7385 rows and 12 columns. Abbreviations are used to describe the features such as:

Model:

4WD/4X4 = Four-wheel drive

AWD = All-wheel drive

FFV = Flexible-fuel vehicle

SWB = Short wheelbase

LWB = Long wheelbase

Transmission:

A = Automatic

AM = Automated manual

AS = Automatic with select shift

AV = Continuously variable

M = Manual

Fuel type:

X = Regular gasoline

Z = Premium gasoline

D = Diesel

E = Ethanol (E85)

N = Natural gas

Make: Company of the vehicle

Model: Car model

Vehicle Class: Class of vehicle depending on their utility, capacity and weight

Engine Size: Size of engine used in liter

Cylinders: Number of cylinders

Transmission: Transmission type with number of gears

Fuel type: Type of Fuel used

Fuel Consumption City: Fuel consumption in city roads (L/100 km)

Fuel Consumption Hwy: Fuel consumption in Hwy roads (L/100 km)

Fuel Consumption Comb: The combined fuel consumption is shown in L/100 km

Fuel Consumption Comb mpg: The combined fuel consumption is shown in mile per gallon

The response variable in this data set is CO2 Emissions (g/km). This variable can be used to gain insight into the effect of various independent variables on fuel consumption. The independent variables are Make, Model, Vehicle class, Engine Size, Cylinders, Transmission and Fuel type. These factors have a direct impact on the amount of fuel consumed by a vehicle as well as its emissions output.








Exploratory Data Analysis (EDA) is an important tool that can help us understand how these factors influence fuel consumption and emissions levels. Through EDA we can generate summary statistics such as means and medians for each factor which will give us an indication of their effects on our response variable CO2 Emissions (g/km).

Exploratory Data Analysis

In *Table 1* and *Table 2* you can see a summary of both categorical (named as factor in table) and numeric variables:

Skim variable	N missing	Complete rate	N Unique Character
Make	0	1	42
Model	0	1	2048
Vehicle class	0	1	16
Transmission	0	1	27
Fuel Type	0	1	5

Table 1- Factor variable summary

Skim variable	N missing	Complete rate	Mean	Sd	P0	P25	P50	P75	P100	hist
Engine size	0	1	3.16	1.354	0.9	2.0	3.0	3.7	8.4	
Cylinders	0	1	5.615	1.823	3.0	4.0	6.0	6.0	16.0	
Fuel consumption city	0	1	12.5534	3.5	4.2	10.1	12.1	14.6	30.6	
Fuel consumption highway	0	1	9.04	2.224	4.0	7.5	8.7	10.2	20.6	
Fuel consumption comb	0	1	10.975	2.892	4.1	8.9	10.6	12.6	26.1	
Fuel consumption comb (mpg)	0	1	27.481	7.231	11.0	22.0	27.0	32.0	69.0	
CO2 Emissions	0	1	250.5846	58.512	96.0	208.0	246.0	288.0	522.0	

If we take a look at fuel consumption in both city and highways which are shown by *Figure 1* we can observe that fuel consumption is highways is higher than in cities. The cause of this is fairly obvious. Since vehicles are driven faster in highways than in city.

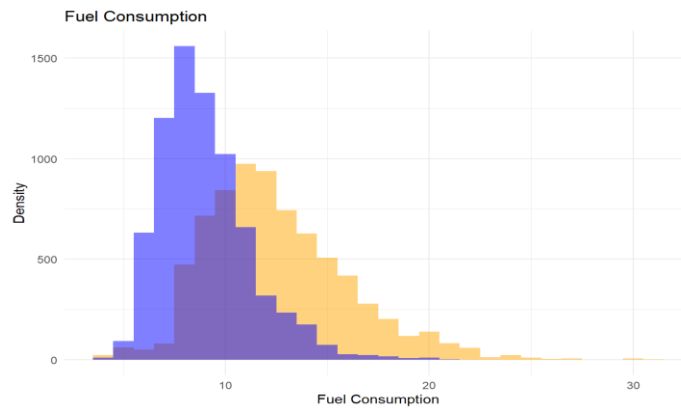


Figure 1- City and highway Fuel Comparison

There are two important factors when it comes to CO2 emissions and they are *Fuel Consumption* and *Fuel Type*. in *Figure 2* we can also observe for both city and highways what kind of Fuel types are being used. Fuel types Z and X are used dominantly both in highways and city which are "Premium gasoline" and "regular gasoline"

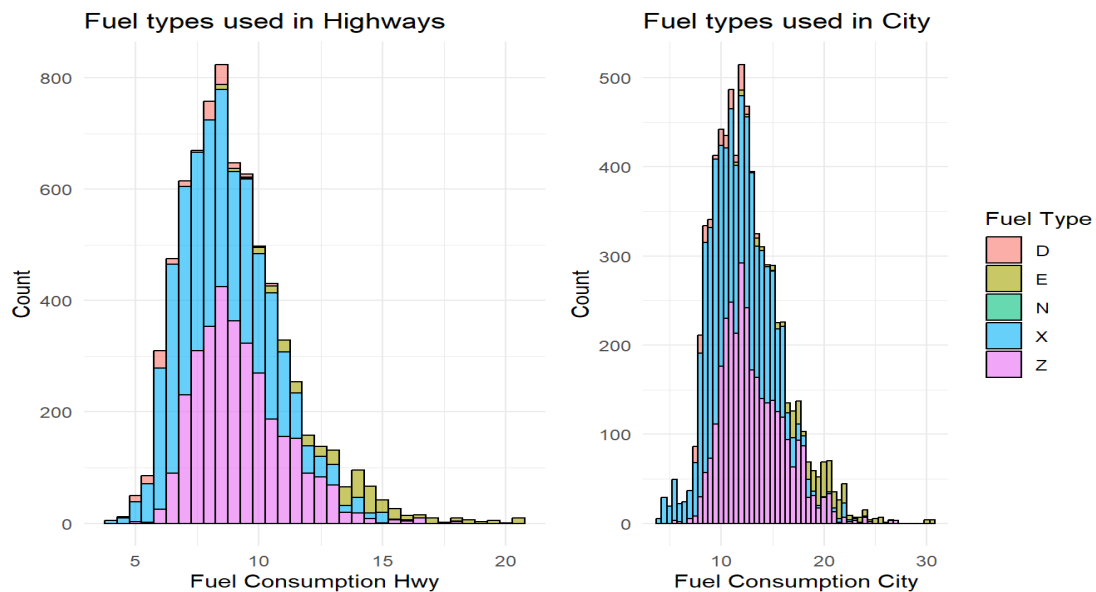


Figure 2 – Fuel types used in highways and in city

Transmissions, transfer power from the engine to the wheels. Figure 3 and Figure 4, we aim to investigate the potential correlation between transmissions and fuel consumption, which ultimately contributes to the emission of CO₂.

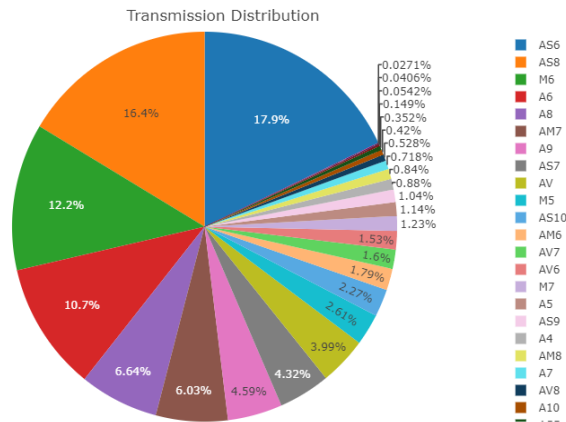


Figure 3 - Transmission Distribution

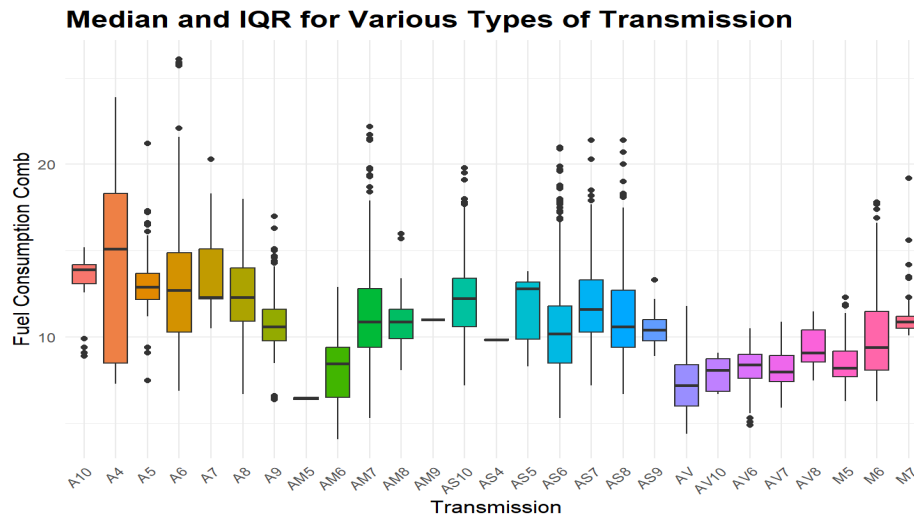


Figure 4 – Median and IQR For Various types of transmission

Based on our observation, it is evident that the majority of transmissions exhibit similar levels of fuel consumption. Therefore, for the purpose of predicting CO₂ emissions, we can safely disregard this particular feature in our further analysis.

The company that made a vehicle is hardly related to the amount of CO₂ emitted by the vehicle; as it can be observed in Figure 5.

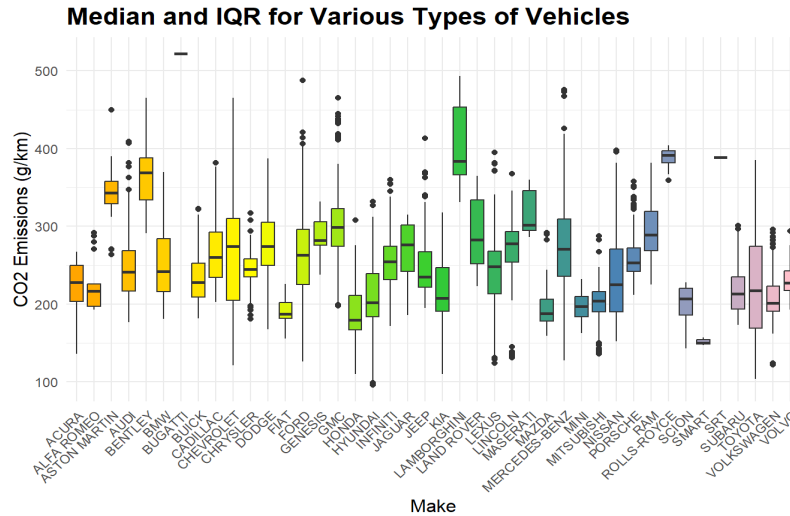


Figure 5 - Median and IQR For Various types of vehicles

Fuel Consumption in city and highways has high correlation with CO2 emissions as it can be observed in Figure 6. There is a strong linear relation between these two variables.

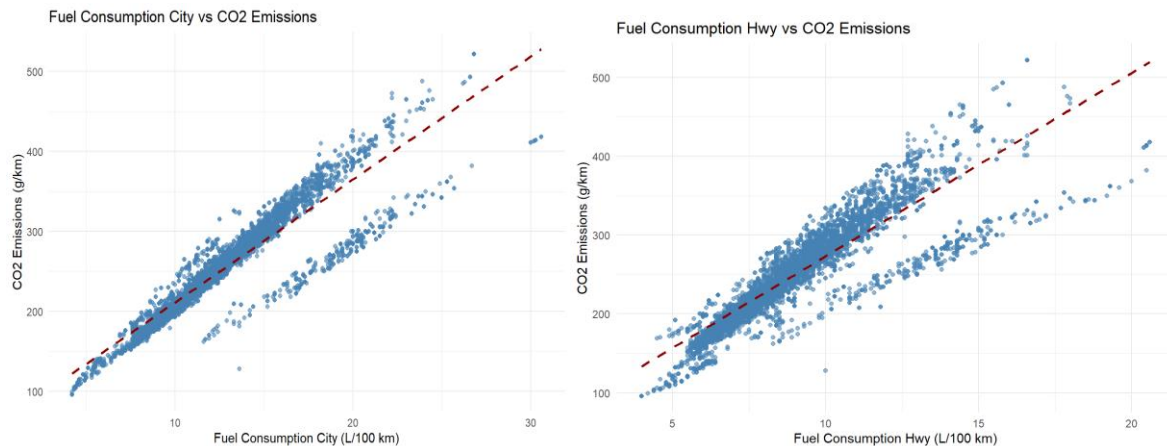


Figure 6- Fuel consumption and CO2 Emissions in City and Highway

However, Figure 6 displays the presence of two distinct distributions. It is imperative to investigate the underlying reasons for this phenomenon in order to appropriately tailor our regression models.

As previously discussed, the fuel type plays a crucial role in determining CO2 emissions. Certain fuel types contribute to lower CO2 emissions compared to others. Figure 7 illustrates that Ethanol Fuels emit approximately 10 grams less CO2 per kilometer compared to other fuels. To ensure the development of an effective model, it is recommended to convert these fuel types into numerical data during the preprocessing stage of regression methods. This conversion will facilitate the construction of a robust model.

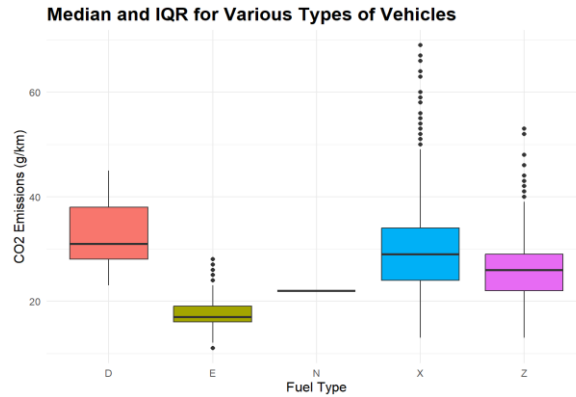


Figure 7 - Median and IQR For Various types of Fuels

In Figure 8, the data illustrates the fuel consumption of different classes of vehicles. The analysis reveals that larger vehicles exhibit higher fuel consumption per kilometer. This information holds significance in the context of establishing models that incorporate the variable of "car size", which will be further elaborated upon in subsequent sections of this paper.

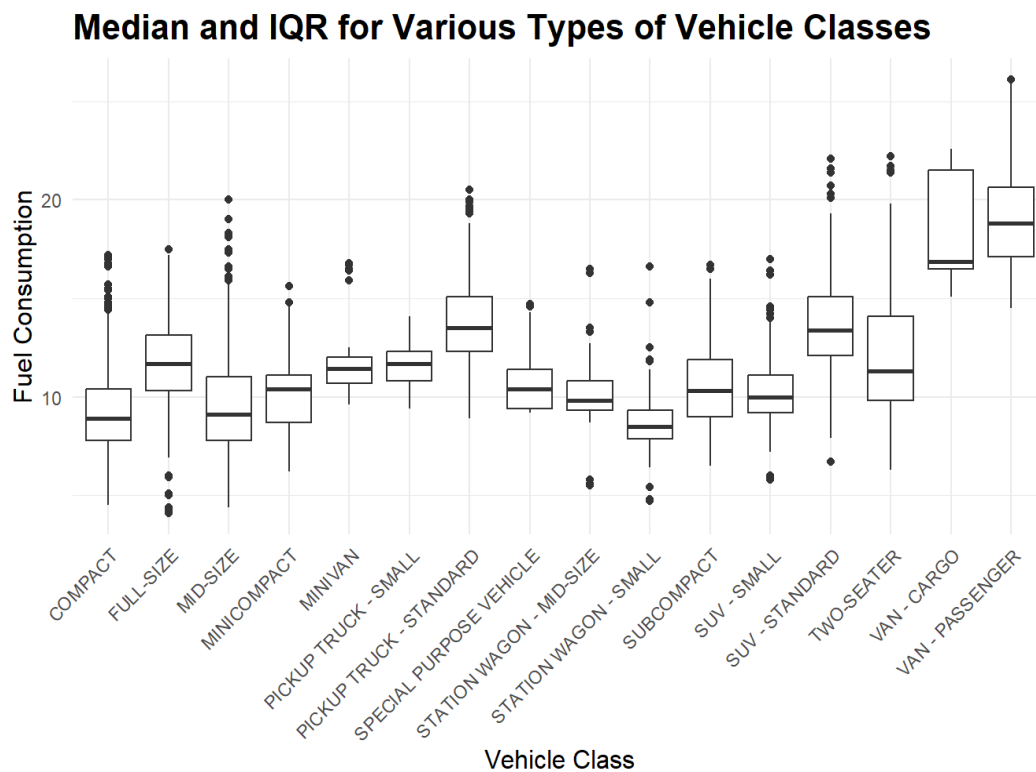


Figure 8 - Median and IQR for Various Types of Vehicle Classes

In the concluding phase of the exploratory data analysis (EDA), it is essential to present a concise graphical summary of the numerical variables and their correlation with CO2 emissions. To achieve this, we can generate a correlation matrix, depicted in Figure 9.

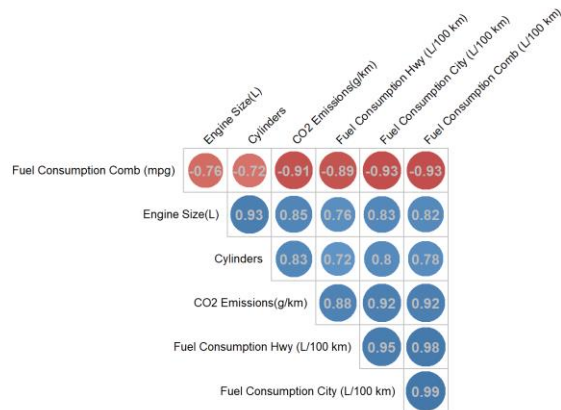


Figure 9 – Correlation matrix

Preprocessing

As evident from the dataset, there is only one vehicle that utilizes the N type of fuel. Considering its negligible correlation with the response variable, it is advisable to exclude this particular vehicle from further analysis.

New suggested method

As mentions before we can add a variable called car size:

$$car\ size = \log \left(engine\ size * cylinders * Fuel\ consumption\ Comb \left(\frac{L}{100km} \right) \right)$$

We used logarithm to omit linear dependency

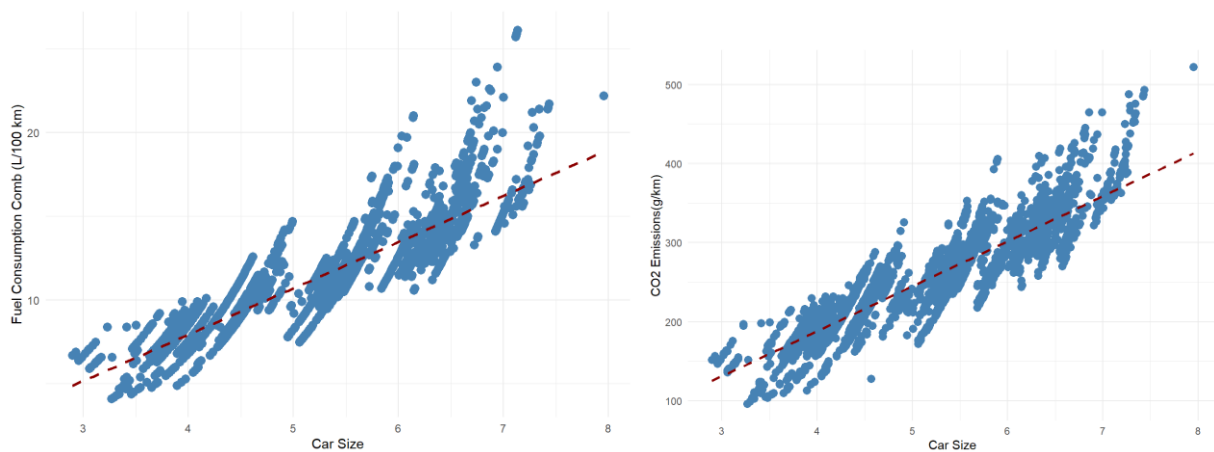


Figure 10- car size and it's correlation between CO2 and Fuel consumption

Normalization

To ensure optimal evaluation metrics, certain regression methods require the use of a normalized dataset.

Data Splitting

However, to ensure that the evaluation metrics of the regression models are optimal, it is necessary to normalize the dataset before training and testing. Normalization is the process of scaling the values of the dataset to a standard range, typically between 0 and 1 or -1 and 1. This ensures that all variables have equal importance in the model and prevents any one variable from dominating the prediction. Therefore, for certain regression methods, normalization of the dataset is a crucial step in achieving accurate and reliable results.

Regression Analysis

In the field of predicting CO₂ emissions, various regression methods have been extensively utilized to understand the intricate relationship between different factors and carbon dioxide output. Eight such regression methods, namely linear regression, K-nearest neighbors (KNN), decision tree, random forest, lasso regression, ridge regression, elastic net, and gradient boosting, have proven to be particularly effective in this domain.

Linear regression

Linear regression is a fundamental method that assumes a linear relationship between the dependent variable (CO₂ emissions) and the independent variables. It seeks to identify the best-fit line that minimizes the sum of squared errors, providing insights into the direct influence of various factors on CO₂ emissions.

K-nearest neighbors (KNN)

K-nearest neighbors (KNN) is a non-parametric method that predicts CO₂ emissions based on the similarity of neighboring data points. It calculates the average of the K nearest data points to estimate the emissions, allowing for flexibility in capturing non-linear relationships.

Decision tree regression

Decision tree regression employs a tree-like model to predict CO₂ emissions by splitting the data based on different variables. It recursively partitions the data into subsets, enabling the identification of complex relationships and interactions between factors.

Random forest

Random forest is an ensemble learning method that combines multiple decision trees to make accurate predictions. By aggregating the results of individual trees, random forest accounts for both the bias and variance of the model, resulting in robust CO₂ emission predictions.

Lasso regression

Lasso regression is a regularization technique that adds a penalty term to the linear regression objective function. By shrinking the coefficients of less important variables to zero, lasso regression selects a subset of relevant variables, enhancing the interpretability of the model.

Ridge regression

Ridge regression, similar to lasso regression, introduces a penalty term to prevent overfitting. However, instead of eliminating variables entirely, ridge regression shrinks their coefficients, allowing for the inclusion of all variables while reducing their impact on the CO₂ emission prediction.

Elastic net regression

Elastic net combines the properties of both lasso and ridge regression. It adds a penalty term that is a mixture of the lasso and ridge penalties, striking a balance between variable selection and coefficient shrinkage.

Gradient boosting regression

Gradient boosting is an ensemble method that combines multiple weak prediction models, typically decision trees, to create a strong predictive model. It sequentially builds new models, focusing on the errors made by previous models and continuously improving the prediction accuracy of CO₂ emissions.

In summary, these eight regression methods provide a diverse set of tools for predicting CO₂ emissions. They offer different approaches to capturing linear and non-linear relationships, handling overfitting, selecting relevant variables, and improving prediction accuracy, enabling researchers and practitioners to make informed decisions and contribute to the mitigation of carbon dioxide emissions.

Evaluation Metrics

R-squared and RMSE are two commonly used metrics for evaluating the performance of regression models. R-squared, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. RMSE, or root mean squared error, measures the average distance between the predicted values and the actual values in the dataset. It is a measure of how well the model can predict new data and is typically expressed in the same units as the dependent variable. A lower RMSE value indicates a better fit of the model to the data. Together, R-squared and RMSE provide a comprehensive evaluation of the accuracy and reliability of regression models.

In table 3, all 8 models that has been described are shown with their RMSE and R^2 values.

Model	RMSE	R2
Linear	2.7530000	0.9980000
KNN	0.0539683	0.9971000
Random Forest	3.2737110	0.9966076
Lasso	5.1116510	0.9918463
Elastic Net	5.1182180	0.9918253
Ridge	12.3900231	0.9520955
Tree	13.9534000	0.9458000
Gradient Boosting	25.5441164	0.7963832

Table 3-

Conclusion

Due to need of normalization the scale of RMSE values aren't suitable for evaluation. However R-squared values are fairly close and scaled. We can observe that the polynomial linear model is the best model and then can be used to train and be tested by train-test datasets.

Bibliography

1. Statistics Using R, Sudha G. Purohit