PREDICTIVE MODELING OF CO2 EMISSIONS FROM VEHICLES

A DATA-DRIVEN ANALYSIS

s15057 - Lakni Kodithuwakku
s15067 - Gayani Pathirana
s15080 - Chamodi Siriwardhana
s15089 - Sanjani Wickramasinghe

Introduction

CO2 emissions are a significant contributor to greenhouse gas emissions, which are known to be a primary driver of global warming and climate change. The transportation industry is a major source of CO2 emissions, with vehicles powered by internal combustion engines being the primary culprits. The amount of CO2 emitted by a vehicle depends on various factors such as the make and model of the vehicle, its fuel efficiency, driving conditions, and technology advancements.

In recent years, advancements in data availability, computational power, and machine learning techniques have significantly enhanced our ability to predict CO2 emissions more accurately. By analyzing large datasets containing information about vehicle attributes, fuel types, driving patterns, and emissions data, researchers and analysts can create models that not only aid in compliance with emissions standards but also inform policies aimed at reducing the carbon footprint of the transportation sector.

The prediction of carbon dioxide (CO2) emissions from vehicles plays a crucial role in understanding and mitigating the environmental impact of the transportation sector. As concerns about climate change and environmental sustainability grow, accurately estimating CO2 emissions is becoming increasingly important for policymakers, researchers, manufacturers, and consumers alike.

In this report, we delve into the methodologies and techniques used to predict CO2 emissions from vehicles. We explore the utilization of machine learning algorithms, statistical analysis, and feature engineering to build predictive models that can assist in estimating emissions levels based on various parameters. By examining a comprehensive dataset, we aim to showcase the potential of predictive modeling in contributing to the development of cleaner and more environmentally friendly transportation solutions.

Description of the Dataset

The dataset "CO_2 Emission by Vehicles" is taken from the Kaggle website, and it contains 7385 rows and twelve columns.

Variable	Description	Type
Make	Company of the vehicle	Qualitative
Model	Car Model	Qualitative
Vehicle Class	Class of vehicle depending on their utility, capacity, and weight	Qualitative
Engine Size (L)	Size of engine used in Liter	Quantitative

Cylinders	Number of cylinders	Quantitative
Transmission	Transmission type with number of gears	Qualitative
Fuel Type	Type of Fuel used	Qualitative
Fuel Consumption City (L/100 km)	Fuel consumption in city roads (L/100 km)	Quantitative
Fuel Consumption Hwy (L/100 km)	Fuel consumption in highways (L/100 km)	Quantitative
Fuel Consumption Comb (L/100 km)	The combined fuel consumption (55% city, 45% highway) is shown in L/100 km	Quantitative
Fuel Consumption Comb (mpg)	The combined fuel consumption in both city and highway is shown in mile per gallon(mpg)	Quantitative
CO2 Emissions(g/km)	The tailpipe emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving	Quantitative

Problem Statement

The problem encompasses the development of a reliable and versatile model capable of estimating CO2 emissions across different vehicles. This model should take into consideration variables such as engine specifications, vehicle weight, fuel efficiency, transmission type etc. The overarching goal is to create a tool that empowers policymakers, manufacturers, and consumers with the ability to make informed decisions regarding vehicle emissions and contribute to the reduction of greenhouse gas emissions from the transportation sector.

Data Pre-processing

Duplicated Entries Handling:
During the initial preprocessing stage, 1103 duplicated entries were identified within the dataset. To ensure data integrity and accuracy, these duplicated records were removed, mitigating any potential biases that could arise from redundant data.

Car Make-Country Mapping:
The 'Make' column, which originally contained 42 unique categories representing the company of the vehicle, was simplified and transformed into the 'make_country' column. This

transformation was achieved by mapping each car to its corresponding country of origin. As a result, the categorical dimension was reduced from 42 to 8, providing a condensed yet informative representation of the vehicle's origin.

Transmission Simplification:
The 'Transmission' column, encompassing 27 unique values representing transmission types with gear information. By extracting the initial character from each entry and mapping it to simplified terms ('M' for 'Manual' and 'A' for 'Automatic'), a new 'transmission_general' column was created. This simplified representation now classifies vehicles broadly as either 'Manual' or 'Automatic,' simplifying the dataset and aiding in more straightforward analyses.

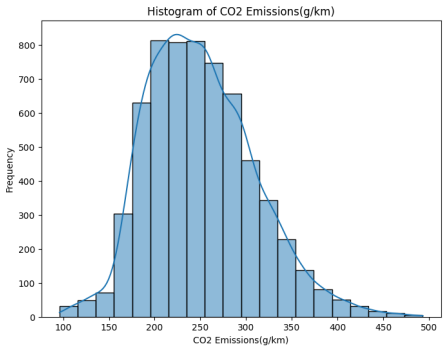
Vehicle Category Mapping:
The vehicle class categories were simplified from 16 original classifications to 6 more concise and representative categories. This reduction in categories is based on shared attributes such as utility, capacity, and weight. The goal is to enhance the dataset's clarity and facilitate analyses by working with a smaller set of more informative vehicle categories.

Data Conversion:
The variables 'Vehicle Class,' 'Transmission,' 'Fuel Type,' 'make_country,' 'transmission_general,' and 'vehicle' were transformed into categorical data type.

Dataset Refinement:
Columns ‘Model,' 'Make,' 'Vehicle Class,' and 'Transmission' were dropped from the dataset, as they were either redundant or had been transformed into more informative features.

One-Hot Encoding:
To facilitate the modeling process, categorical columns such as 'Fuel Type,' 'make_country,' 'transmission_general,' and 'vehicle' were transformed into binary columns through one-hot encoding.

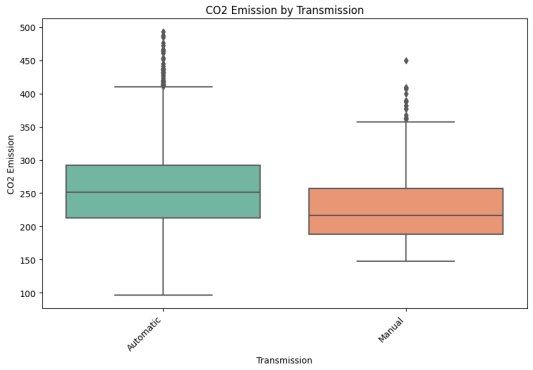
Results of the Descriptive Analysis

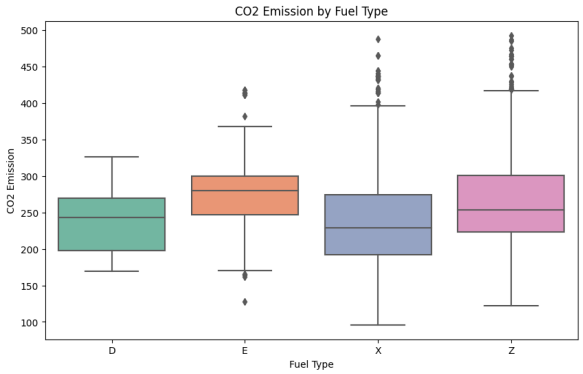


The histogram's shape reveals that a majority of vehicles tend to exhibit relatively lower CO2 emission values. This concentration of lower emission values is indicative of a general trend towards environmentally friendlier vehicles with reduced emissions.

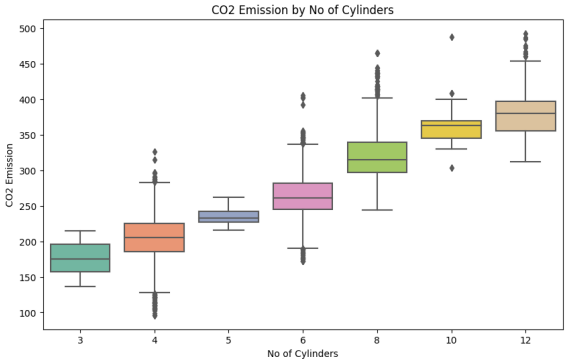
However, the right-skewness of the histogram suggests the presence of a tail towards higher emission values, implying that there are still a notable number of vehicles producing higher CO2 emissions. This could be attributed to various factors such as the vehicle type, fuel efficiency, and technological advancements.

Vehicles with automatic transmissions exhibit a wider spread of CO2 emission range and a higher median, likely owing to their advanced technologies, optimized gear ratios, and suitability for varying driving conditions, particularly in urban settings. The efficiency gains from automatic transmissions' ability to adapt to different situations, coupled with the potential for smoother acceleration and improved fuel consumption, can contribute to their relatively lower emissions in certain scenarios. However, contextual factors such as vehicle weight, type, engine tuning, and driver behavior also play significant roles in influencing emission levels.

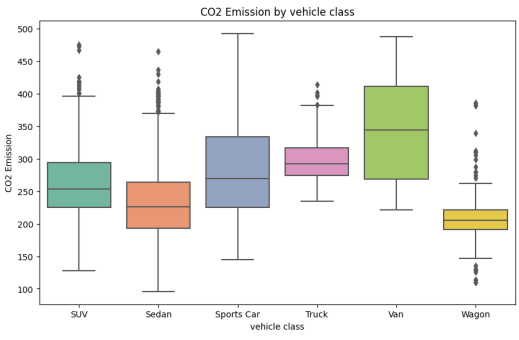




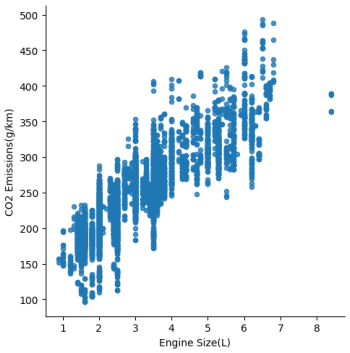
The comparison of median CO2 emissions across different fuel types, particularly noting that type "E" has the highest median CO2 emissions. The higher median emissions associated with the "E" fuel type could stem from a range of factors, including ethanol's lower energy density, distinct combustion characteristics, and potential tuning variations in vehicles designed for ethanol blends.



As the number of cylinders rises, typically in pursuit of greater engine power and performance, several factors contribute to the subsequent increase in CO2 emissions. Larger engines with more cylinders inherently require more fuel to operate, leading to higher fuel consumption and subsequently elevated CO2 emissions.

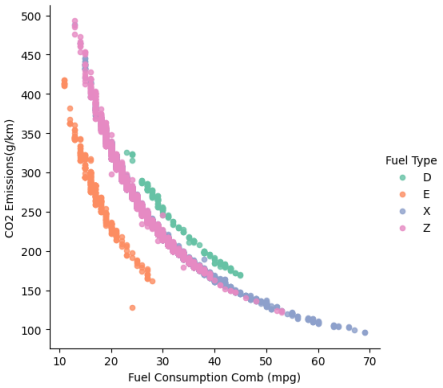


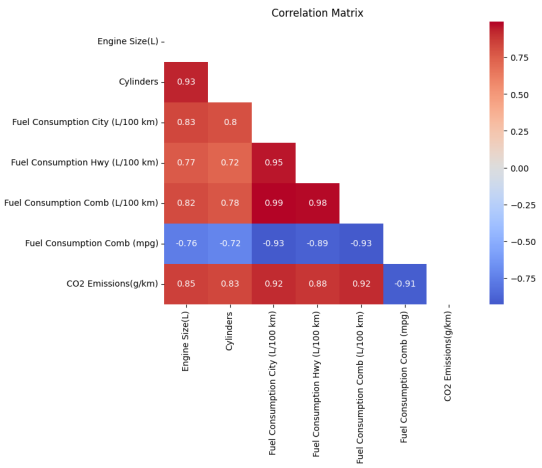
Vans, which exhibit the highest median CO2 emission levels, reflect their larger size and typically higher weight, requiring more energy to propel and resulting in increased fuel consumption. Trucks follow with the second-highest median CO2 emissions, often owing to their heavy-duty nature and utilization for commercial purposes, which might involve carrying substantial loads. Sports cars, known for their performance-oriented design and potentially larger engines, also present a notable median CO2 emission level.



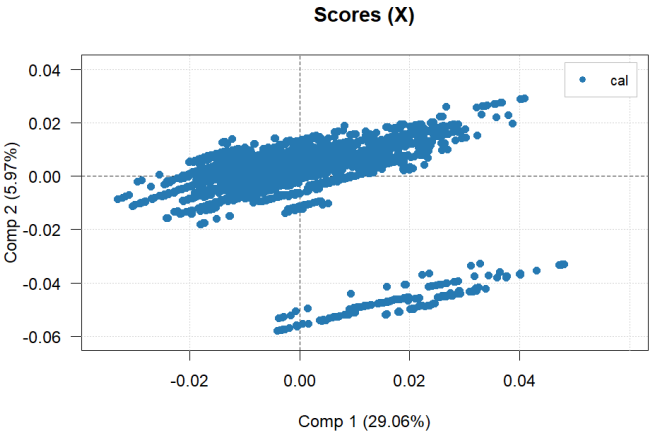
The observed correlation between increased engine size and higher CO2 emissions can be attributed to the intricate relationship between engine capacity, fuel consumption, and power generation. Larger engine sizes inherently accommodate greater volumes of fuel and air, enabling higher power outputs for improved performance. However, this augmented power comes at the cost of increased fuel consumption, as larger engines require more fuel to maintain their enhanced performance capabilities. Consequently, the greater fuel consumption directly translates to elevated CO2 emissions. Additionally, larger engines often drive heavier vehicles, which further amplifies fuel consumption and emissions due to the increased energy required to propel the added mass.

Fuel consumption comb (mpg) represents the number of miles a vehicle can travel on a single gallon of fuel. This pattern highlights the relationship between fuel efficiency and carbon emissions. As fuel consumption decreases and mpg increases, vehicles tend to exhibit lower CO2 emissions, aligning with the principles of energy conservation. A higher miles per gallon value indicates better fuel efficiency, meaning the vehicle can travel farther using less fuel, resulting in lower fuel costs and reduced environmental impact in terms of carbon dioxide (CO2) emissions. The curvature in the pattern suggests initial efficiency improvements yield substantial reductions in CO2 emissions, but as fuel consumption continues to decrease, the rate of emission reduction slows down. This may be due to factors such as the inherent physics of combustion efficiency and the optimal design limits of current technologies.





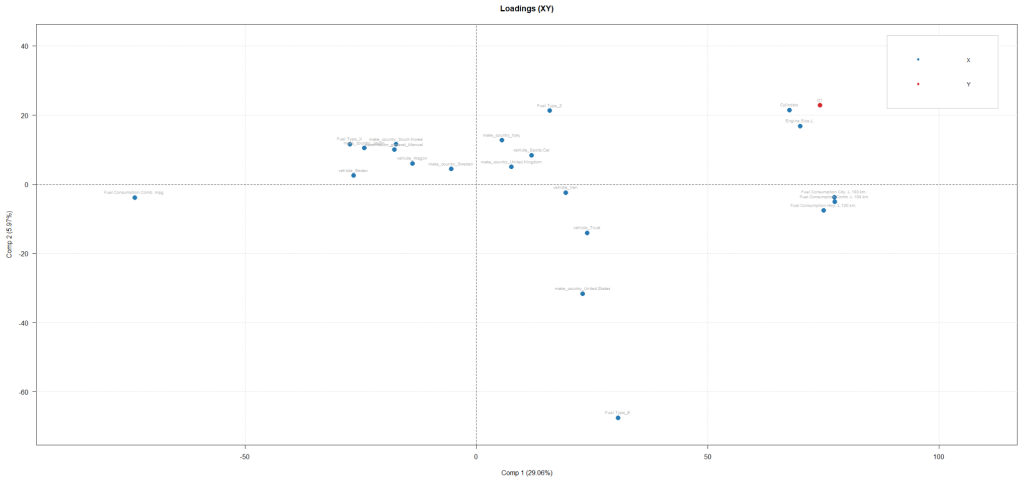
There is a high correlation between Fuel consumption variables and it's reasonable as for each vehicle Fuel consumption in city, Hwy and combined inherent similar patterns. For further analysis, we would only consider the Fuel Consumption Comb (mpg) variable.



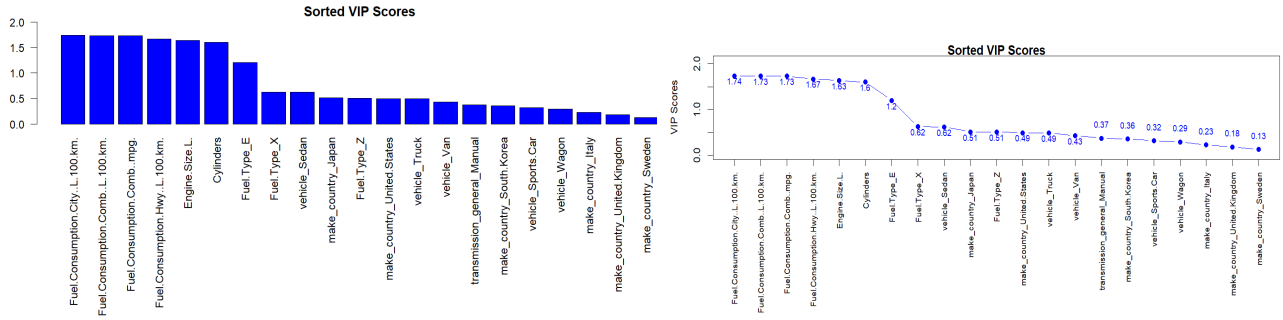
Partial Least Squares Analysis

In the data analysis, partial least squares regression is employed to uncover potential clusters within the observations and identify strong correlations among predictors. Upon careful examination, the observation set did exhibit discernible clusters, as

visually represented in the Scores(X) plot. It is worth noting that the axes of this plot accounted for less than 70% of the total variation in both predictor and response variables. This indicates that the efficacy of PLS in capturing the underlying data relationships is somewhat limited. Nevertheless, the PLS method still managed to elucidate nearly 35% of the total variation.



Turning our attention to the XY loadings plot, it is evident that certain predictors exhibit significant correlations with the response variable (Y). Conversely, some predictors demonstrate orthogonal relationships with the response, indicating a lack of direct correlation and some predictors are close to the origin, indicating lack of importance. Additionally, there are some variable clusters, giving suggestions like to get derived variables or take important variables using feature importance. These findings collectively emphasize the complex interplay between the predictors and the response variable.



VIP scores plot is a bar chart that displays the VIP scores for each input variable. It can help identify which variables have the strongest relationship with the response variable and should be retained in the model. VIP plots can also be used to identify redundant or unimportant variables that can be removed from the model to simplify it and reduce overfitting.

Results of the Advanced Analysis

When selecting the best model for predicting CO2 emissions with higher accuracy, the following evaluation criteria were employed:

1. Mean Squared Error (MSE): MSE quantifies the average squared difference between predicted and actual values. A lower MSE indicates better model performance by minimizing prediction errors.

2. R-squared (R^2): R^2 measures the proportion of variance in the dependent variable explained by the model. A higher R^2 signifies a better fit, indicating that the model explains more variability in the data.

3. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, presenting errors in the same unit as the target variable. A lower RMSE signifies enhanced model accuracy.

4. Mean Absolute Percentage Error (MAPE): MAPE calculates the average percentage difference between predicted and actual values. A lower MAPE indicates greater model prediction accuracy.

To select the finest model, emphasize lower MSE, RMSE, and MAPE values, signifying reduced prediction errors. Furthermore, prioritize higher R^2 values, indicating superior fit to the data. Optimal models strike a harmonious balance between minimized errors and heightened explained variance, resulting in more precise and dependable predictions.

In addition, when comparing train and test accuracies to select the best model, focus on consistency, overfitting avoidance, and generalization. Prefer models that exhibit comparable performance on both sets, indicating equilibrium between accuracy and generalization.

Full Model

In the full model, all the variables were used to predict the CO2 emission.

Model	Train			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
Linear Regression	4.96	1.23%	0.9929	5.29	1.31%	0.9922
Ridge Regression	4.99	1.24%	0.9928	5.26	1.31%	0.9923
Lasso Regression	9.35	2.36%	0.9748	9.33	2.43%	0.9757
Random forest	1.68	0.40%	0.9992	3.68	0.40%	0.9962
XG Boosting	1.75	0.54%	0.9991	3.60	0.88%	0.9964

All models exhibited strong accuracy levels on both the training and testing datasets, indicating their consistent and reliable performance.

Reduced Model

Reducing variables in a model serves the purpose of achieving a parsimonious model, which is characterized by striking a balance between model complexity and performance. A parsimonious model aims to attain a desirable level of goodness of fit while utilizing the fewest possible explanatory variables. This approach promotes simplicity, interpretability, and reduced risk of overfitting. By prioritizing essential variables and excluding unnecessary ones, a parsimonious model not only enhances its efficiency in computation but also facilitates clearer insights into the relationships between variables. In essence, the practice of reducing variables contributes to more efficient, interpretable, and robust models.

The variables 'Fuel Consumption Hwy (L/100 km)' and 'Fuel Consumption City (L/100 km)' redundantly contribute to deriving 'Fuel Consumption Comb (L/100 km)'. Thus, the former two have been removed, retaining only the 'Fuel Consumption Comb (L/100 km)'.

However, since 'Fuel Consumption Comb (L/100 km)' can be converted to mpg, we can also remove it and retain only 'Fuel Consumption Comb (mpg)' which represent all 'Fuel Consumption Hwy (L/100 km)' , ''Fuel Consumption City (L/100 km)' and 'Fuel Consumption Comb (L/100 km)' 3 variables.

Furthermore, in pursuit of a parsimonious model, certain variables with lower importance, as indicated by the Partial Least Squares (PLS) importance plot, have been intentionally removed.

Model	Train			Test		
	RMSE	MAPE	R2	RMSE	MAPE	R2
Linear Regression	15.92	4.30%	0.927	15.89	4.39%	0.9295
Ridge Regression	15.92	4.30%	0.927	15.89	4.39%	0.9297
Lasso Regression	17.20	4.59%	0.9148	17.31	4.76%	0.9163
Random forest	3.73	1.09%	0.996	4.31	1.24%	0.9948
XG Boosting	3.67	1.08%	0.9961	4.89	1.25%	0.9933

Upon fitting the reduced model, the accuracy slightly decreased compared to the full model, yet it remains notably high and the random forest model will be used for further predictions. This outcome underscores the effectiveness of the parsimonious model. Given its high accuracy performance and the added benefits of simplicity and reduced complexity, it is reasonable to continue with the reduced model for practical implementation.

Code: https://colab.research.google.com/drive/1EC1gdvWhOFImnfCcszHiv30dzeOmmQ8Z?usp=sharing