Final Project Report

Predicting CO2 Emissions of Motor Vehicles by Engine Specifications and Manufacturer

Contributors
Rubeena Rasheed
Norah Alamri
Connor Mennenoh
Pradyumna Deshpande

Illinois Institute of Technology
CSP571-Data Preparation and Analysis
FALL 2022

TABLE OF CONTENTS

Abstract

With the effects of climate change becoming ever more present both in the daily lives of average people and a topic of discussion in an ever-increasing number of political contexts, the subject of emissions is being discussed now more than ever. Although there are many sources of emissions, the ones closest to most people's minds are the emissions from their cars and other vehicles. Our objective was to examine a data set that included several key features of motor vehicles, along with their CO2 emissions, in order to determine if a predictive model could be used in determining the likely amount of emissions a vehicle would produce, in lieu of more costly and time-consuming methods like direct measurement after a vehicle has already been put into production. To do so, a data set of vehicle specifications, along with the manufacturer, was utilized. We employed feature engineering to convert the largely categorical data into a format we could fit several models, too, in order to determine the feasibility of predicting the final emissions of a vehicle purely from its static features. Ultimately, we were able to produce multiple models that, while not offering immaculate performance, did provide predictive power that was promising and had the potential for future attempts with similar data.

1. Introduction

With so many changes and developments in the world revolving around the question of climate change, hardly a day goes by that the topic of emissions and how to lower them does not come up. While there are many sources of emissions, the one that is most relatable to people and is a primary driver of global emissions is that of motor vehicles. In developed nations, motor vehicles are almost entirely ubiquitous, with each vehicle generating a non-trivial amount of emission with each use and an even less non-trivial amount over the course of its lifetime. However, not all vehicles are created equal; and many groups, such as climate activists, governments, and corporations, have a vested interest in knowing how emissions differ between different vehicles with a greater level of precision than simply assuming larger vehicles emit more.

A fair amount of work has been done to create and refine processes for measuring the emissions of vehicles in a way that transfers to their actual performance in real-world situations. Such investigations are the source of the type of data that this report will utilize. Still, we wanted to investigate if it would be possible to produce a predictive model that could take the specifications of a vehicle, elements that can be defined early on in the design process, such as fuel type, engine size, number of cylinders, etc. and use these features to determine the final emissions level the vehicle will produce with any level of reliable precision. We also sought to make inferences about the relationship between the different features and emissions, as well as search for any broader trends among the various categories, such as the maker of the vehicles themselves. Once we have determined if such a model is feasible, we intend to determine what variety of models performs best, discuss possible extensions to this model, and finally examine how this type of analysis could be extended or refined in future work.

2. Objectives

We had several objectives we wished to accomplish in this project. Most were determined from the outset, but as we explored the data more thoroughly, other questions arose. The following objectives are listed in descending order of priority.

1. Create a model that can predict the CO2 emissions of a motor vehicle using its specifications with some modicum of accuracy.
2. Compare multiple models and determine which performs the best and provide the most inferential value.
3. Identify avenues of future research.

3. Data Source and Pre-Processing

We found the data set we chose to use on *kaggle.com*. One of the reasons we chose to work with this particular data set was the fact that this data was essentially entirely clean to begin with. It had disadvantages over other emissions data sets, but it was in very good shape to begin the exploratory data analysis with only a small amount of cleanup. The raw data consisted of 7385 rows and 12 features. The features were primarily categorical, with only engine size and fuel consumption being the only continuous variables. Handling these multi-valued categorical variables would present our first major challenge in the initial processing of this data. A table describing the initial features follows.

TABLE 3.1. Description of Initial Features

Feature Name	Description	Data Type
Make	The company that manufactured the vehicle	String
Model	The particular production model of the company	String
Vehicle Class	A category that the vehicle belongs to which generically encompasses size, weight, and capacity	String
Engine Size	Size of the engine in liters	Numeric
Cylinders	Number of Cylinders	Numeric
Transmission	Transmission type with the number of gears	String
Fuel Type	Type of fuel used such as regular or premium gasoline, or diesel	Character, corresponding to a String
Fuel Consumption City	Fuel consumption in typical city driving in Liters per 100 Kilometers	Numeric
Fuel Consumption Highway	Fuel consumption in typical highway driving in Liters per 100 Kilometers	Numeric
Fuel Consumption Combined L/100 km	Combined fuel consumption (55% city and 45% highway) in Liters per 100 Kilometers	Numeric
Fuel Consumption Combined mpg	Combined fuel consumption (55% city and 45% highway) in Miles per Gallon	Numeric
CO2 Emissions	CO2 emissions measured from the vehicle tailpipe,	Numeric

	measured in Grams per Kilometer	

Initially, we attempted to simply create dummy variables for all of the variables using the fastDummies package in R. Still, this naive approach yielded a new set of over one thousand features. The primary offender causing so many new features to be required was the Model variable, as every manufacturer had its own suite of models they produced, meaning each row essentially had a unique model. This feature was subsequently removed as it added no new information not captured in the other features already. Additionally, the data set had three different features measuring fuel consumption. One measuring highway fuel consumption, one measuring city fuel consumption, and two weighted averages of the two, one in miles per gallon and the other in liters per one hundred kilometers. We removed all but the combined fuel consumption in liters per one hundred kilometers, as we felt this captured the most general information about the fuel consumption using the units most familiar worldwide, thus improving potential interpretability later on.

These two alterations had a significant impact on the final number of features once we introduced dummy variables for the categorical features that remained, reducing the final count from over one thousand to just over one hundred. We likely could have proceeded much in the same way as we did from here, but we wanted to explore the remaining features a bit more to see if any further reductions could be made. To do so, we looked at the vehicle class feature and noticed many of the categories were, in fact subcategories of more general classes such as different sizes of SUVs and trucks, and compact cars. As shown in the code below, we condensed the categories down to these more general size classifications, allowing for another significant reduction in the final feature count.

```
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "SUV - SMALL"] <- "SUV"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "SUV - STANDARD"] <- "SUV"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "PICKUP TRUCK - STANDARD"] <- "PICKUP TRUCK"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "PICKUP TRUCK - SMALL"] <- "PICKUP TRUCK"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "VAN - CARGO"] <- "VAN"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "VAN - PASSENGER"] <- "VAN"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "MINIVAN"] <- "VAN"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "STATION WAGON - MID-SIZE"] <- "STATION WAGON"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "STATION WAGON - SMALL"] <- "STATION WAGON"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "MINICOMPACT"] <- "COMPACT"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "SUBCOMPACT"] <- "COMPACT"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "MID-SIZE"] <- "SEDAN"
raw_data$Vehicle.Class[raw_data$Vehicle.Class == "FULL-SIZE"] <- "SEDAN"
```

The final feature we considered reducing was transmission/gears. It is important to note that this variable did not simply house a numeric value as one may expect but included both the number of gears available and the type of transmission the vehicle used. This presented the question of how we might be able to reduce this category without losing information about the transmission type. To answer this question, we would need to begin exploring the features in more depth to determine what patterns existed both in the transmission gear feature, as well as the others.

4. Exploratory Data Analysis

In order to better understand the data we were working with, we generated box plots for the continuous features we chose to keep, namely engine size, number of cylinders, and combined fuel consumption.

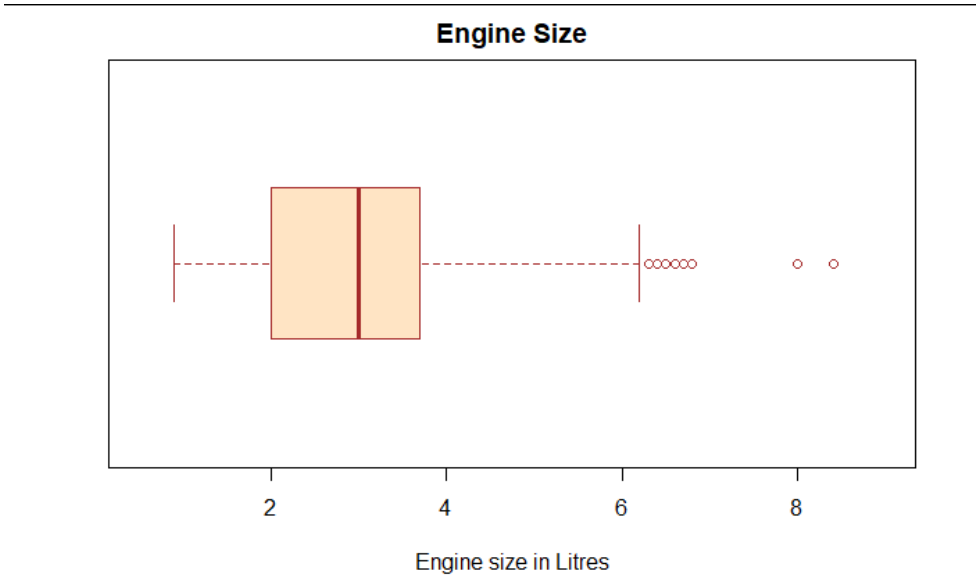

Figure 4.1: Engine Size in Litres

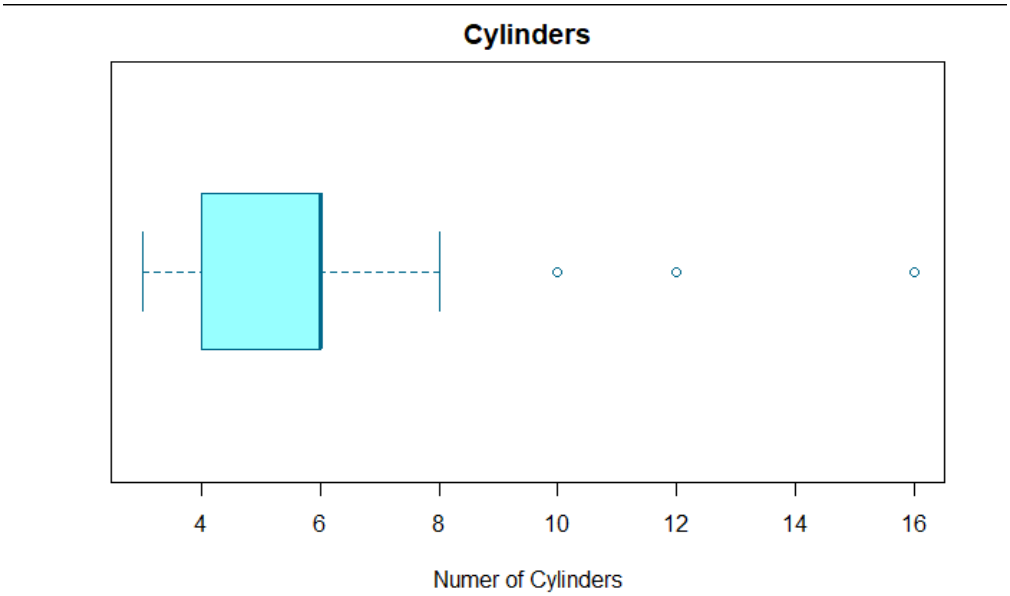

Figure 4.2: Number of Cylinders

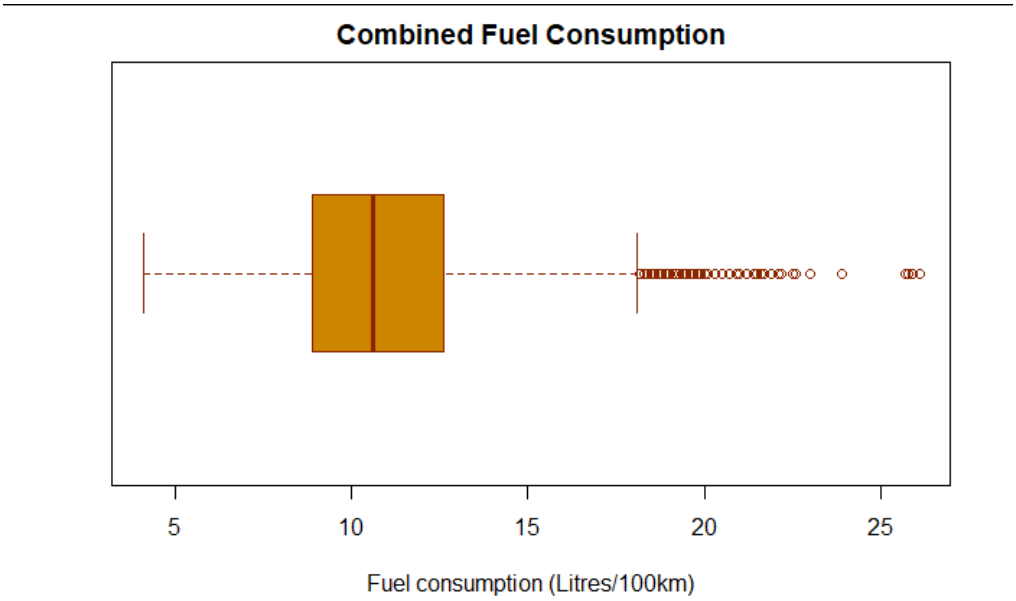

Figure 4.3: Combined Fuel Consumption

As shown above, each category had outliers, but the bulk of the data points was within a common range. For these quantities, we chose to keep the outliers in the data, as they often corresponded to non-standard or special-use vehicles. Eliminating these outliers would have produced a better-performing model overall. Still, it would have reduced its applicability to only common-use vehicles, a reduction in generality we wished to avoid.

Exploring the categorical variables was more limited and primarily consisted of looking at the distribution of the various quantities in the data set. While this did not yield as much information as the box plots, they did help to motivate our earlier choice of combining the Vehicle Class categories into a more condensed set, as many of the delineated subcategories were small on their own, and by combining them would capture a better picture of how vehicles in that approximate size and chassis design behaved as a whole, rather than separating them.

Finally, to begin understanding the potential relationship between features and our target response: CO2 emissions, plots were generated showing box plots or averages of CO2 emissions for various values of our features.

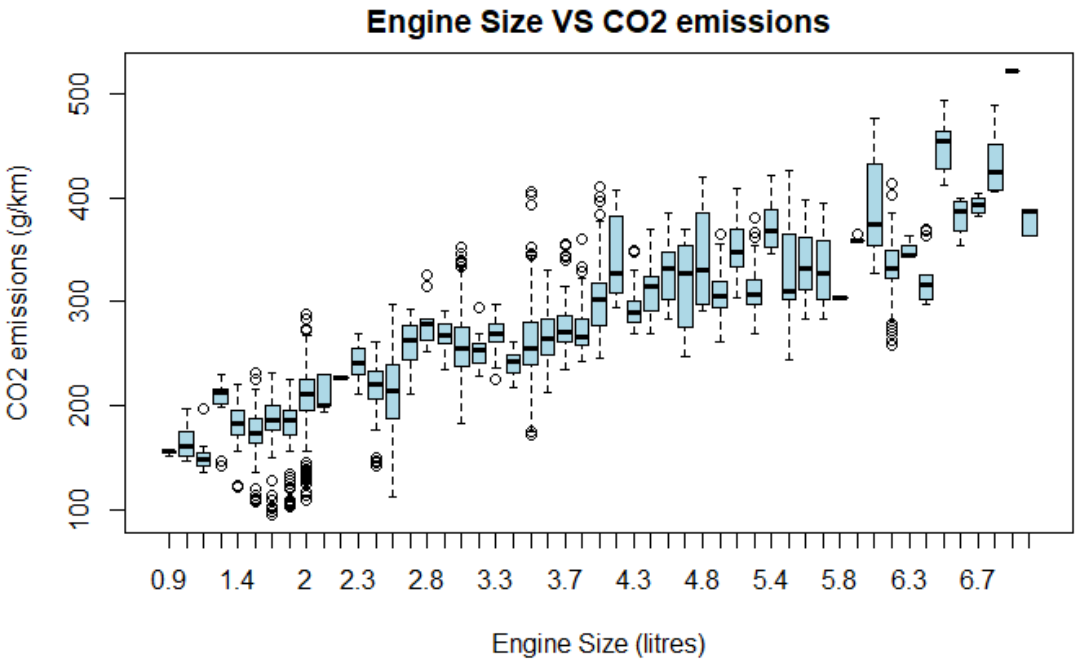

Figure 4.4: Engine Size VS C02 emissions

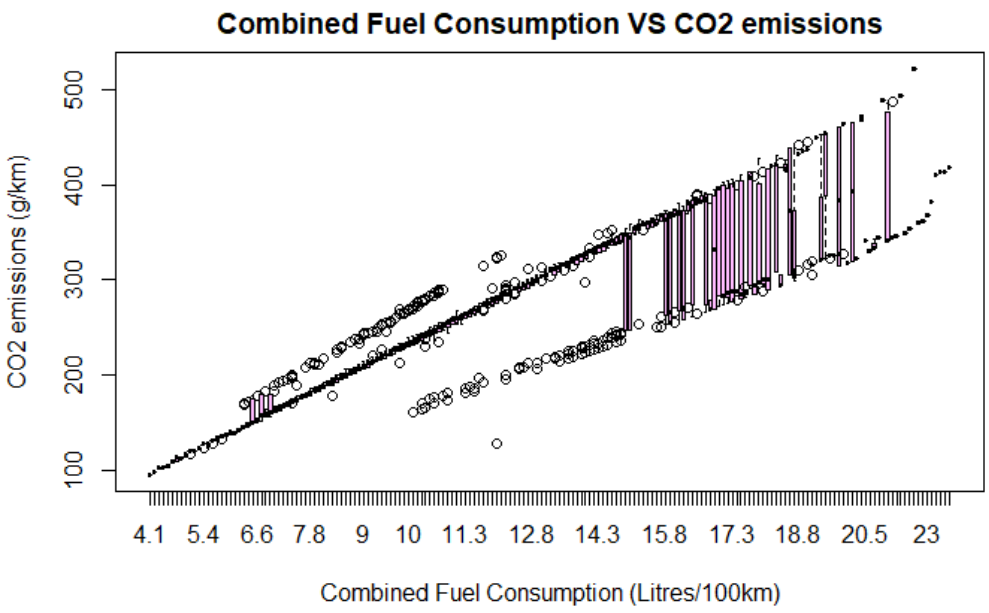

Figure 4.5: Combined Fuel Consumption VS CO2 emissions

In Figure 4.6, we can see the relationship between CO2 emissions and cylinder size. The higher the size of the cylinder, the higher its CO2 emissions of it. This showcases the high correlation between these two variables and the direct effect of the engine cylinder size on its emission value.

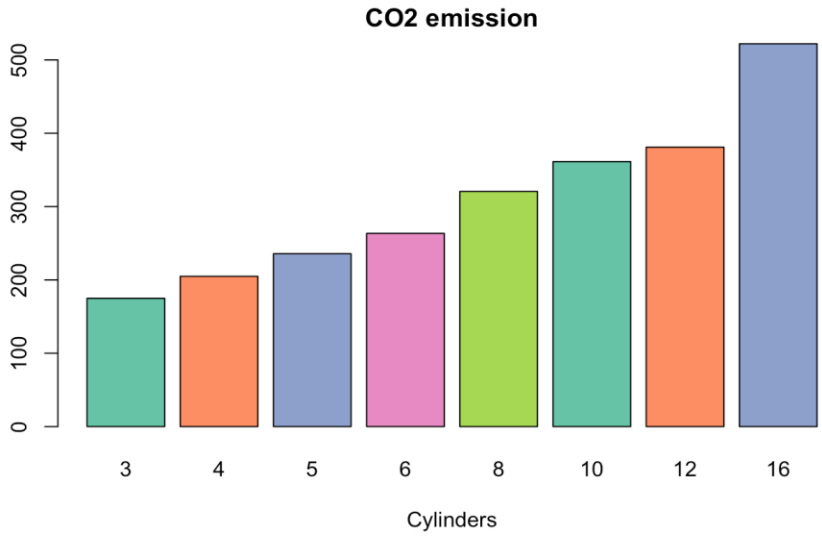

Figure 4.6: Engine Cylinder size VS CO2 emissions

Figure 4.7 shows the type of fuel with its relation to CO2 emission. The lowest type of fuel in CO2 emissions is N, which is natural gas. The highest type of fuel in CO2 emissions is E and Z, respectively, which are Ethanol (E85) and premium gasoline. X is the regular gasoline, and it is surprisingly lower than the premium gasoline with respect to CO2 emissions.

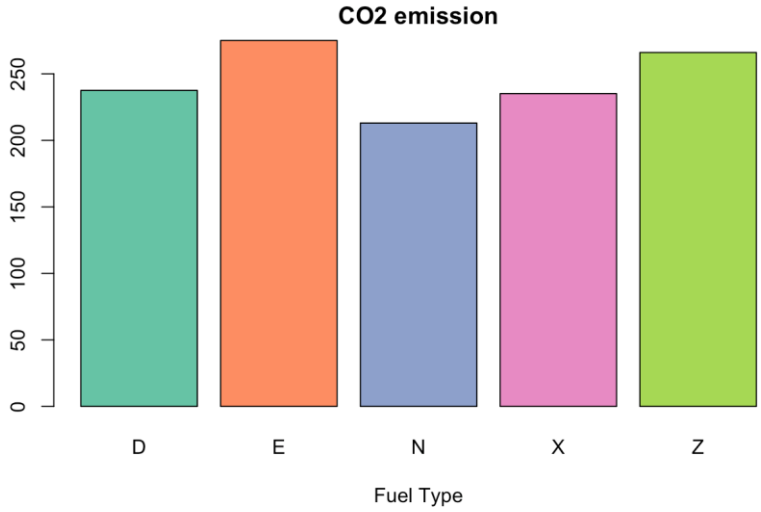

Figure 4.7: Fuel Type VS CO2 emissions

Figure 4.8 shows the relationship between CO2 emissions and the car transmission type. We can not see a clear pattern here to describe the relationship. We faced this issue when we tried to reduce the number of categories of Transmission types. We could not find a link between the categories in order to combine them together. Therefore we decided to keep it as is.

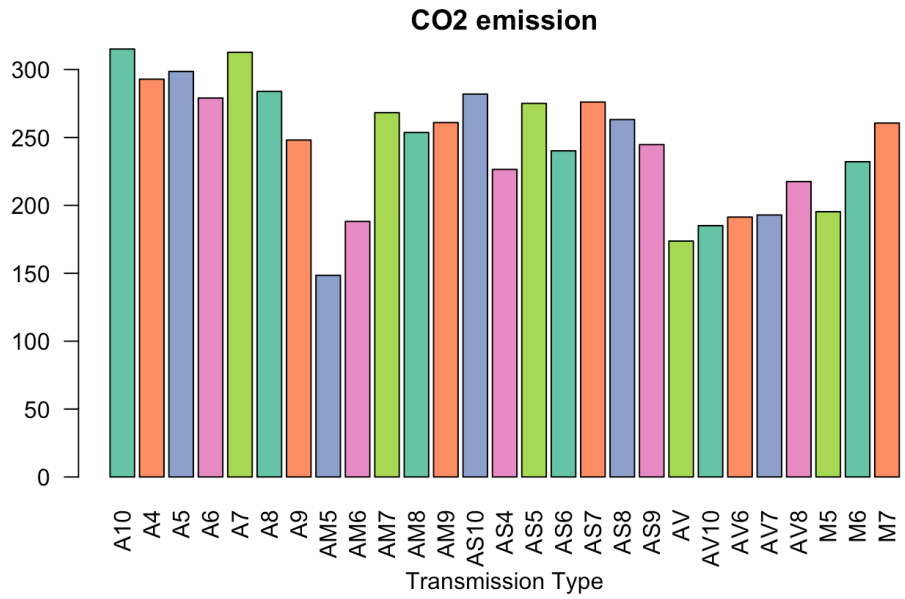

Figure 4.8: Car Transmission Type VS CO2 emissions

Figure 4.9 is important as it visualizes the relationship between the vehicle type and CO2 emission. The vehicles with the highest CO2 emissions are vans and pickup trucks. This makes sense because the size of the engines in these two vehicle types is always large, and in figure 4.6, we have seen that the larger the engine size, the higher the CO2 emission.

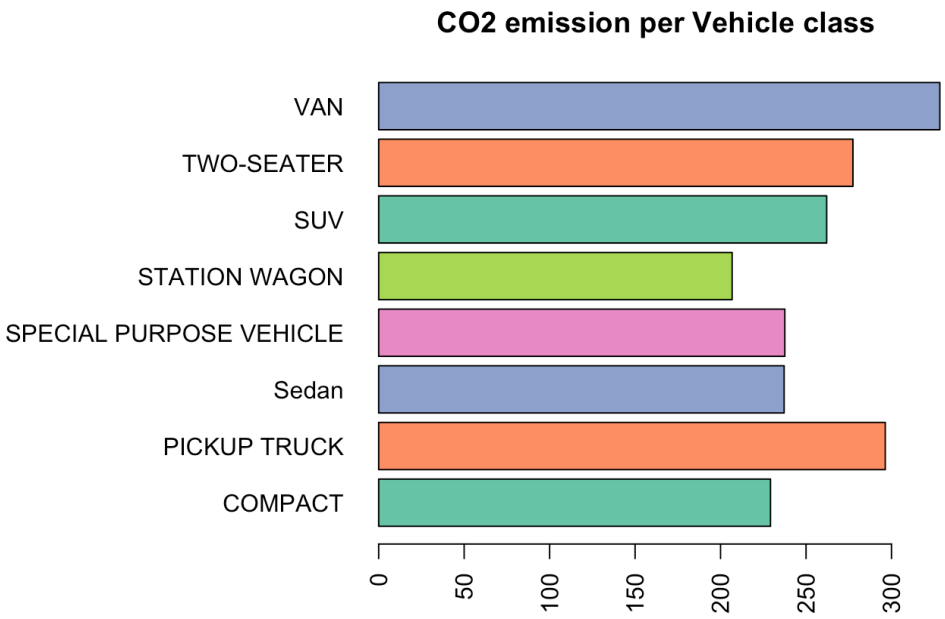

Figure 4.9: CO2 emission per Vehicle class

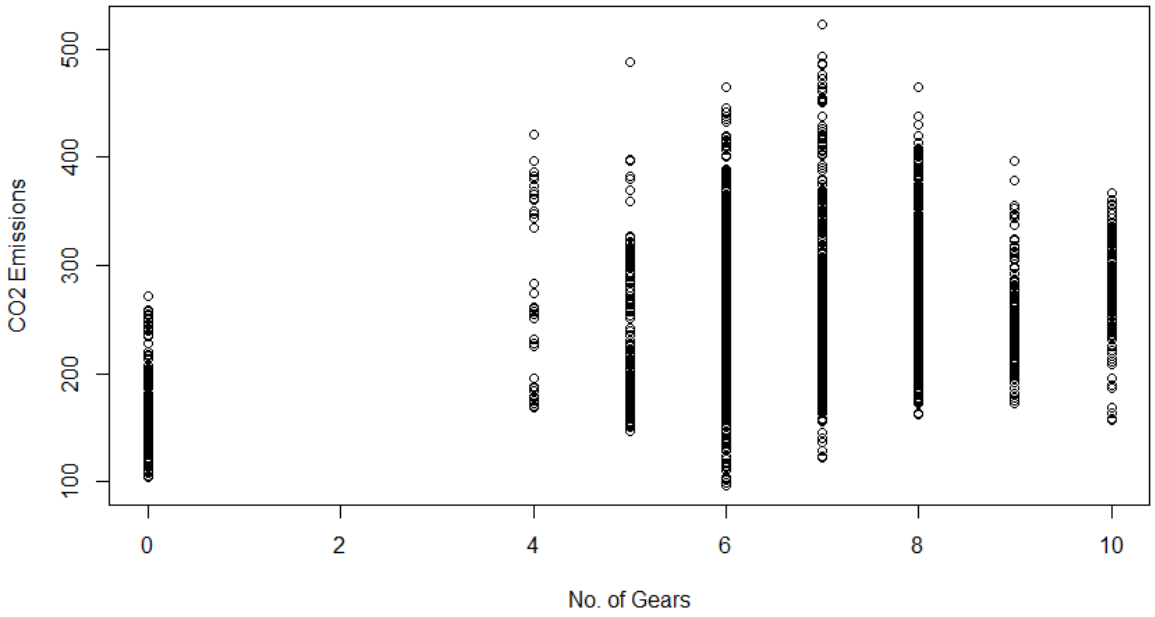

Figure 4.10: CO2 emission VS No. of Gears

lt was at this stage that the intended outcome of the analysis began to look feasible in practice rather than just in concept. Comparing records by certain features and their corresponding CO2

emissions made two important conclusions immediately clear. The first was that no quantity had a completely perfect correlation with CO2 emissions. Some quantities came very close, such as the engine size, whose correlation with CO2 emissions is one of the strongest shown in the correlation matrix generated later in this report. However, many quantities did have clearly visible trends that aligned with CO2 emissions, showing a positive correlation in the plots we generated. This pattern of relationship, especially one that appears somewhat linear, increased our confidence that our primary objective of simply making any model that could produce usable results was possible. Less promisingly, when looking at the transmission CO2 emission values by the number of gears, removing the transmission type, no clear pattern was evident. Our intention had been to group transmissions by the number of gears if we could find some sort of grouping across gear numbers. Still, the final plot shown above indicates that there did not seem to be a strong enough relationship between the number of gears and CO2 to justify segmenting them. For this reason, the Transmission feature remained unaltered and kept its large number of dummy variables in the final data as a result.

Once our initial exploration of the data was complete and dummy variables had been made for each remaining categorical variable, our data set was reduced to 7385 observations with a total of 85 features. This feature count was larger than we had initially hoped for but is low enough that the approaches we intended to take would not be horribly overwhelmed by the curse of dimensionality. To seek out any remaining information not captured in the above plots and exploration, a correlation matrix was generated to find any remaining notable connections between the features. Initially, we attempted to generate a correlation matrix for all 85 features, but the result was essentially unreadable and consisted of primarily non-correlated cells. To fix this, we made use of a function written by Catherine Williams. We posted for use on the Towards Data Science forum to produce a correlation matrix only consisting of quantities with correlations above a certain significance threshold, in this case, 0.4.

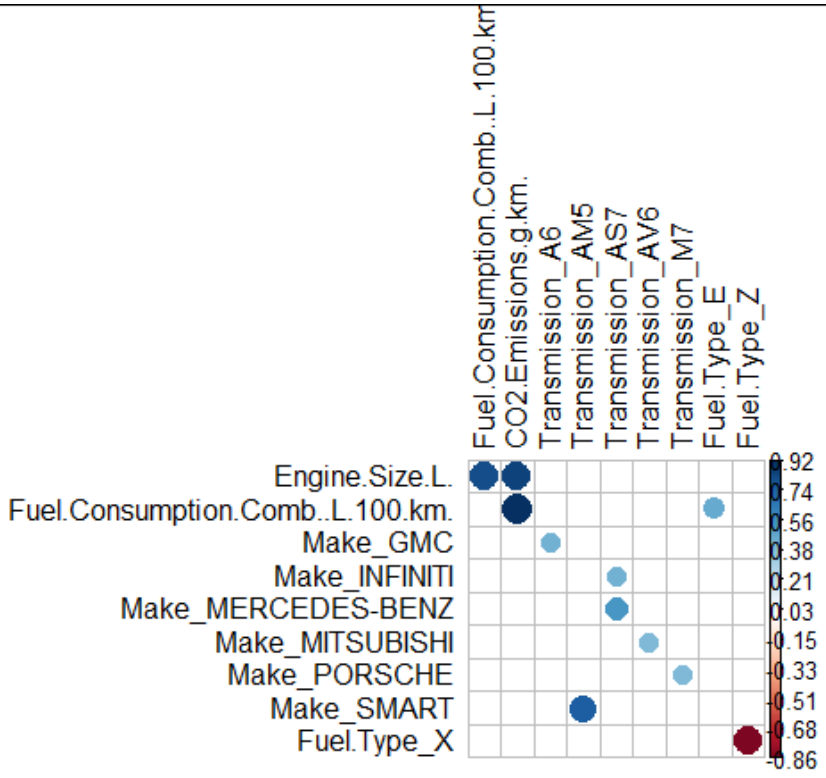

Figure 4.11: Correlation Matrix

The result is primarily as expected, demonstrating the stronger correlation between Engine Size and emissions, as well as combined fuel consumption with emissions. The other notable correlations are between manufacturers and particular transmission types, which indicates a tendency for certain manufacturers to produce certain types of transmissions. With our data cleaned, our features reduced and engineered, dummy variables made, and the initial exploration of our data complete, we moved on to fitting a variety of models to the data to determine in earnest if our hopes for predicting emissions were viable.

5. Modeling and Analysis

We chose to fit four models in total. Three of the models would be based on a core linear model, one being an ordinary least squares regression model, and the other two using the ridge regression and lasso regression techniques to improve upon the standard least squares potentially. The last model would be a decision tree model, namely a random forest ensemble model. Our focus on primary linear models arose from the observation of somewhat linear correlations between CO2 emissions and the most highly correlated features. By adding the random forest as well, we were able to introduce a model with lower bias that still had the potential for good performance and a fairly low variance from being an ensemble model. For the last model, we have chosen Gradient boosting, which is a famous ensemble learning technique. Gradient Boosting aims to reduce the high variance of learners by averaging a number of models - in our case, 10,000 trees- these trees were fitted on bootstrapped datasets sampled with replacement from the training data.

To evaluate the performance of the models, an 80/20 train/test split was utilized, with each model being trained on eighty percent of the original data set and then tested on the remaining twenty percent that had been held out. The generated models were tested by predicting the CO2 emissions for the test data, then comparing it to the known response values for the test data by calculating the mean squared error. By comparing these values, shown in the table below, the overall prediction accuracy on out-of-sample data could be compared for each model.

TABLE 5.1. Known response values of each model

#	Model	R2	MSE	RMSE	MAE
1	Linear Regression	-	6775.53	-	-
2	Lasso - Linear Regression	0.99	31.36	5.60	3.03
3	Ridge - Linear Regression	0.972	103.71	10.18	7.58
4	Random Forest	0.994	19.25	4.39	2.32
5	Gradient Boosting	0.996	14.25	3.77	2.04

Before the results above are discussed further, it should be noted that there is inherent variability in using a single train/test split. For ease of comparison, we chose to use this method of evaluation but also investigated a handful of other splits to see if these types of performance were consistent with other potential splits.

The model performance was massively variable. The ordinary least squares model represents a worst-case scenario, as we were able to produce a model, but when tested out of sample, its results are functionally useless. The ridge regression was able to improve upon the OLS model massively but still produced a model with a large error, making its predictive power limited at best. Lasso and Random Forest are where the models actually began to produce predictions that were usable. In the case shown above, Random Forest outperformed Lasso by a relatively small margin, but in other splits, the converse was true. In each case, we observed that Lasso and Random Forest consistently outperformed the other two models and performed within a relatively close margin of each other. The Gradient Boosting Model has shown the best performance of all models, with slightly better results than Random Forest. The Gradient Boosting Model resulted in an MSE of 14.25 and an RMSE of 3.77. To build the Gradient Boosting Model, we used 10,000 trees and averaged their results.

6. Conclusion

To conclude, our goal for this project was to create a model that predicts the value of CO2 emissions based on the attributes available in our dataset. First, we started with the data investigation and EDA, then with the data preprocessing and cleaning. For the EDA, we split the tasks into two, one for categorical variables and the other for numerical variables, the EDA resulted in a number of graphs that describe the dataset in a more accurate way. From the EDA, we have learned that the engine size, cylinder number, and fuel consumption of the vehicle contribute directly to the amount of CO2 emissions. As for data pre-processing and cleaning, the data set was initially clean and ready for use. However, we wanted to enhance the quality of the data by reducing the number of categories of some variables. We tried to reduce the number of categories in the Transmission type variable and Vehicle type. We only succeeded in doing so in the latter one as we couldn't find a relationship between categories of the transmission types. Hence we could not combine them. After that, we created dummy variables for the categorical data in order to prepare it for the training phase. For the modeling phase, we started by splitting the dataset into training and testing datasets, then started building 5 different models: Linear Model, Lasso Linear Model, Ridge Linear Model, Random Forest Model, and Gradient Boosting Model. After training the models, we measured their performances using 4 different measuring metrics: R2, MSE, RMSE, and MAE.

The model with the best performance was the Gradient Boosting Model, with an MSE of 14.25 and an RMSE of 3.77. For the Gradient Boosting Model, we have trained 10,000 trees and averaged their results, this explains the outstanding performance of this model in comparison to the other models. We have learned a lot from this project in terms of teamwork, solving new technical problems, and working with a tight schedule. In the future, we plan to explore other machine-learning models and apply cross-validation techniques to improve the quality of training and model performance.

References

Kaplan, Jacob. “Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables [R Package fastDummies Version 1.6.3].” The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN), November 29, 2020. https://cran.r-project.org/web/packages/fastDummies/index.html.

Podder, Debajyoti. “CO2 Emission by Vehicles.” Kaggle, August 5, 2020. https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles.

Williams, Catherine. “How to Create a Correlation Matrix with Too Many Variables in R.” Medium. Towards Data Science, March 1, 2020. https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57.

Related Work

Buberger, Johannes, Anton Kersten, Manuel Kuder, Richard Eckerle, Thomas Weyh, and Torbjörn Thiringer. “Total CO2-Equivalent Life-Cycle Emissions from Commercially Available Passenger Cars.” *Renewable and Sustainable Energy Reviews* 159 (2022): 112158. https://doi.org/10.1016/j.rser.2022.112158.

Hooftman, Nils, Maarten Messagie, Joeri Van Mierlo, and Thierry Coosemans. “A Review of the European Passenger Car Regulations – Real Driving Emissions vs Local Air Quality.” *Renewable and Sustainable Energy Reviews* 86 (2018): 1–21. https://doi.org/10.1016/j.rser.2018.01.012.

Martinet, Simon, Yao Liu, Cédric Louis, Patrick Tassel, Pascal Perret, Agnès Chaumond, and Michel André. “Euro 6 Unregulated Pollutant Characterization and Statistical Analysis of after-Treatment Device and Driving-Condition Impact on Recent Passenger-Car Emissions.” *Environmental Science & Technology* 51, no. 10 (2017): 5847–55. https://doi.org/10.1021/acs.est.7b00481.

Wang, Haikun, Lixin Fu, and Jun Bi. “CO2 And Pollutant Emissions from Passenger Cars in China.” *Energy Policy* 39, no. 5 (2011): 3005–11. https://doi.org/10.1016/j.enpol.2011.03.013.

Weiss, Martin, Lukas Irrgang, Andreas T. Kiefer, Josefine R. Roth, and Eckard Helmers. “Mass- and Power-Related Efficiency Trade-Offs and CO2 Emissions of Compact Passenger Cars.” *Journal of Cleaner Production* 243 (2020): 118326. https://doi.org/10.1016/j.jclepro.2019.118326.