Contents

Data Science Project Report.

Introduction.

We chose to focus on the transportation sector for our project. There has been a significant increase in the measurement of CO2 emissions, which has resulted in environmental changes.
Our project involves creating a Machine Learning model that employs linear regression to estimate the level of pollution that a car engine could emit.
Our focus is on utilising a Machine Learning approach to address the expected CO2 emissions resulting from vehicles.
We collected a large dataset containing diverse varieties of cars.
ML models we will use for prediction: Linear Regression, SVR, Random Forest, and Lasso Regression.

Proof of Concept.

In the context of a machine learning project focused on CO2 emissions by vehicles, a proof of concept (POC) would involve creating a small-scale model that uses machine learning algorithms to predict CO2 emissions based on different vehicle features. The POC would involve collecting and cleaning a small subset of the available data, selecting appropriate features, and training a machine learning model using the data. The purpose of the POC is to demonstrate that the machine learning approach can accurately predict CO2 emissions based on the selected features. The results of the POC will be used to determine whether to invest further resources into developing a full-scale machine learning model for predicting CO2 emissions.

Define the problem

1) Defining the object in terms.

We aim to help car manufacturers in predicting the amount of CO2 their cars will emit by utilising several factors such as engine size, number of cylinders, fuel consumption rates for city, highway, and combined driving, as well as car make, model, and class. This will aid the manufacturers in meeting future CO2 standards set globally, ensuring their cars are environmentally acceptable.

2) How will our solution be used?

When a car manufacturer produces a new vehicle, there are various factors involved that can impact the amount of CO2 the vehicle will emit. For example, a larger engine or one that is less fuel-efficient is likely to emit more CO2 than a smaller or more efficient engine. With our solution, we can create a clear understanding of the expected CO2 emissions for the car manufacturer's product.

3) How will we frame the problem?

Since we have labelled data, this is a supervised learning problem. Our data is considered offline since we are not connected to a database, and the data does not update automatically.

4) How should our performance be measured?

We will measure the performance of our models with RMSE. By using RMSE, the accuracy of the models can be quantified, and improvements can be made to the models to achieve better predictions.

5) Is the performance measure aligned with the business objective?

The performance measurement allows us to assess the accuracy of our predictions for future vehicles' CO2 emissions. This aligns with the business objective of making the best possible predictions for each car.

6) What would be the minimum performance needed to reach the business objective?

We consider the model's performance to be good if the root mean squared error is less than 8 units. The lower the RMSE value, the better the model is at making accurate predictions. The model can be incorrect in up to 5% of its predictions.

7) What are comparable problems? Can you reuse experience or tools?

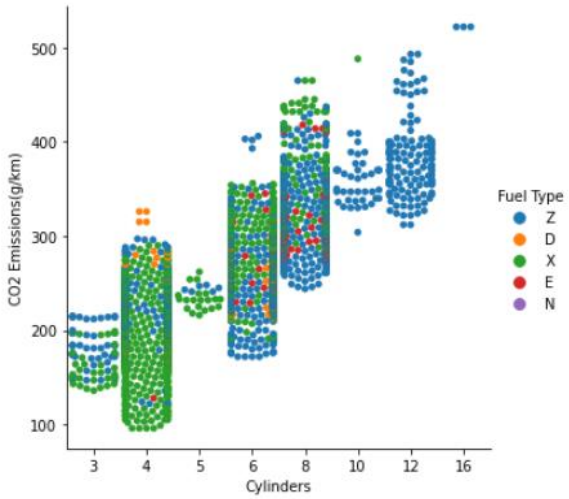
Regression problems with similar dataset sizes can be compared to gain insights into how to approach new problems. For example, predicting housing prices is a common regression problem that can inform how to address other similar problems. Models developed for previous regression problems can also be used as a starting point for new problems.

8) List the assumptions we have made so far.

Engines with large size and more cylinders tend to consume more fuel and produce more CO2 emissions.

9) Verify assumptions if possible.

Based on our visualisation we can see that there is a rise in CO2 emissions as the size of a vehicle's cylinder increases. (Figure 1)



X = Regular gasoline
Z = Premium gasoline
D = Diesel
E = Ethanol (E85)
N = Natural gas

Figure 1

Data set

The dataset records information about how the CO2 emissions of a vehicle can vary based on various features. The data was obtained from the official open data website of the Canadian government and has been compiled into a single dataset. The dataset covers a period of seven years and contains a total of 7385 rows and 12 columns.

* Source link to the dataset: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

Data inspection.

Our data is very clear, with no missing values and significant outliers. Furthermore, we have a variable called “fuel consumption combined” that we could utilize instead of using “fuel consumption highway” and “fuel consumption city” separately, which would help us reduce the number of dimensions.

```
RangeIndex: 7385 entries, 0 to 7384
Data columns (total 12 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Make                              7385 non-null   object
 1   Model                             7385 non-null   object
 2   Vehicle Class                     7385 non-null   object
 3   Engine Size(L)                    7385 non-null   float64
 4   Cylinders                         7385 non-null   int64
 5   Transmission                      7385 non-null   object
 6   Fuel Type                         7385 non-null   object
 7   Fuel Consumption City (L/100 km)  7385 non-null   float64
 8   Fuel Consumption Hwy (L/100 km)   7385 non-null   float64
 9   Fuel Consumption Comb (L/100 km)  7385 non-null   float64
 10  Fuel Consumption Comb (mpg)       7385 non-null   int64
 11  CO2 Emissions(g/km)               7385 non-null   int64
dtypes: float64(4), int64(3), object(5)
```

df.describe()

	Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
count	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000
mean	3.160135	5.615030	12.556534	9.041706	10.975071	27.481652	250.584699
std	1.354125	1.828307	3.500274	2.224456	2.892506	7.231879	58.512679
min	0.900000	3.000000	4.200000	4.000000	4.100000	11.000000	96.000000
25%	2.000000	4.000000	10.100000	7.500000	8.900000	22.000000	208.000000
50%	3.000000	6.000000	12.100000	8.700000	10.600000	27.000000	246.000000
75%	3.700000	6.000000	14.600000	10.200000	12.600000	32.000000	288.000000
max	8.400000	16.000000	30.600000	20.600000	26.100000	69.000000	522.000000

Exploratory Data Analysis.

“Make” column describing the brand.
“Model” column identifies a specific type or version within a product line.
“Transmission” column indicates the type of transmission, such as an automatic, manual, or continuous variable.
“Vehicle Class “ is a categorical column that describes the size of the car and characteristics. Example would be SUV, convertible, compact, etc.
“Fuel” type column specifies the type of fuel required for the vehicle, such as gasoline, diesel, ethanol, or natural gas.

Attributes.

Only “Make”,” Model”, ”Vehicle Class”, “Fuel Type” and “Transmission” are categorical, all other attributes are numerical. Our target is a numerical feature and is Co2 Emission.

```
columns = df.columns
print(columns)
```

```
Index(['Make', 'Model', 'Vehicle Class', 'Engine Size(L)', 'Cylinders',
       'Transmission', 'Fuel Type', 'Fuel Consumption City (L/100 km)',
       'Fuel Consumption Hwy (L/100 km)', 'Fuel Consumption Comb (L/100 km)',
       'Fuel Consumption Comb (mpg)', 'CO2 Emissions(g/km)'],
      dtype='object')
```

Visualizing the data

For more insights we visualized data by using heatmap. It shows us a strong correlation between Fuel Consumption Comb (L/100 km) and Co2 Emissions(g/km).

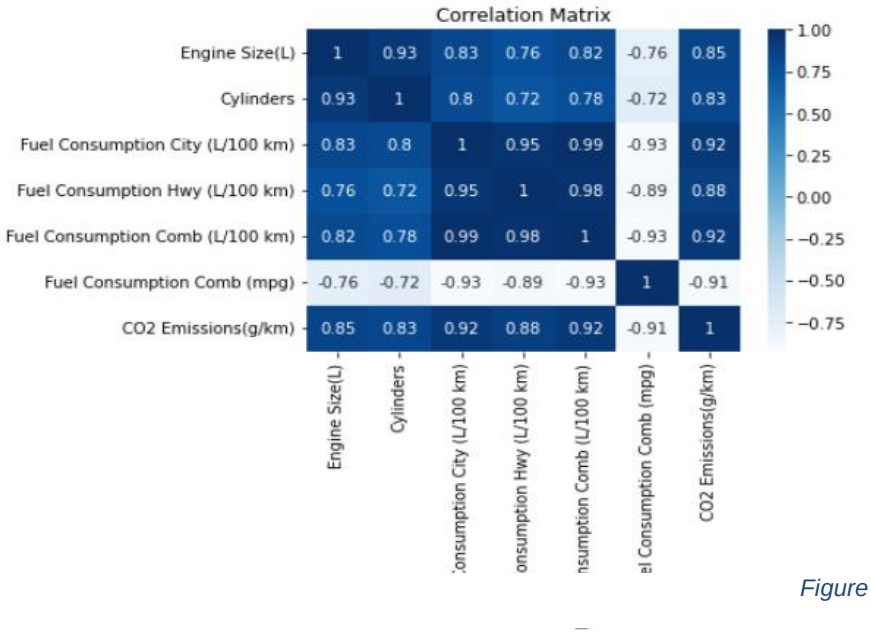


Figure 2

We have 7385 samples in our dataset and you can see the distribution of them.(Figure 3)

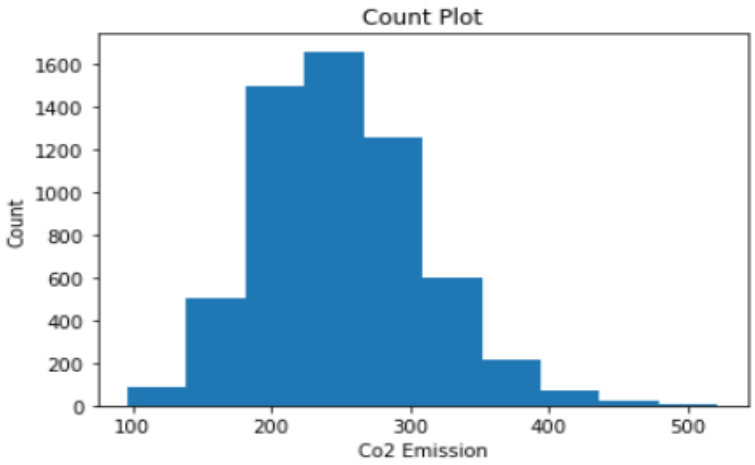


Figure 3

Project structure and layout.

1. Choose the Co2 emissions dataset to predict car emissions.
2. Created Trello board for project management.
3. Attended class and had a meeting to check off the checklist and start working.
4. Received ideas about Clustering and Linear Regression from classmates but struggled with coding.
5. Aimed to produce a POC on the dataset and write it down.
6. Worked mostly on coding and produced the first result from a Random Forest regressor.
7. Cleaned up a code and created the framework for the report/results.
8. Decided to include categorical data in predictions using one-hot encoder.

9. Updated the X_value to fuel_consumption_combined and substituted ‘Make’ with ‘Fuel Type’ for better results.
10. Explored different models such as Lasso, Ridge, SVR.
11. Got help from the teacher to improve the checklist and used cross_val_predict to better understand results.
12. Trained more models to compare results and vizualized the data with EDA.
13. Created a new test-set for predictions and chose the two best models, SVR and Random Forest.
14. Used RandomsearchCV for SVR and GridsearchCV for Random Forest to get the best hyperparameters.
15. Completed coding and planned to start working on Streamlit.
16. Finished with Streamlit application and preparing presentation.

Method and Step-by-step analysis

We have decided to use grid search cross-validation instead of random search because our dataset is not very large, and we believe that grid search will produce better results.
For our project we used five different regression models.

- Linear - Linear regression is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables. The model aims to find the best linear relationship between the input variables and the output variable.
- SVR - Support vector regression is a regression algorithm based on support vector machines that is used to predict continuous outcomes. SVR works by finding a hyperplane in a high dimensional space that has the maximum distance to the closest data points.
- Random Forest-Random Forest is a machine learning algorithm used for classification and regression problems. It is an ensemble learning method that combines multiple decision trees to make predictions. With random forest it builds a large number of decision trees and then takes the average of the predictions to obtain a final prediction.
- Lasso - Lasso is a linear regression algorithm used for predicting continuous outcomes. It is similar to Ridge regression but includes a different regularisation term in the function to penalise large coefficients.
- Ridge - Ridge regression is a linear regression algorithm used for predicting continuous outcomes. In some way it is similar to linear regression, but it includes an additional regularisation term in the function that penalises large coefficients resulting in less overfitting. The regularisation term is controlled by the parameter called alpha.

To be able to use our categorical data in our dataset we use a technique called one-hot encoding. One-hot encoding involves creating a binary vector for each category where the vector is as long as the total number of categories and contains a 1 in the position corresponding to the category and 0s elsewhere.

Results

By deleting 2070 columns, we saw improvements in both performance and speed for certain models. When we reduced the number of columns from 2080 to 10 in some models, there were mixed results, but overall, the score improved. Random Forest, for instance, demonstrated greater accuracy and speed.

Results before/after for our training set.

	2080 columns		10 columns	
Models	RMSE	R2 Score	RMSE	R2 Score
SVR	2.82	0.997	2.55	0.998
Random Forest	1.29	0.999	1.41	0.999
Linear Regression	2.97	0.997	4.92	0.992
Ridge Regression	3.02	0.997	4.92	0.992
Lasso	5.03	0.992	5.03	0.992

After scaling we got the best timing for SVR but results did not change . When we normilized it it get even worse. Best models are SVR and RandomForset. With scaling and normalizing timing significantly got better. From 43.6 seconds to 1.89 seconds.
Scaling turned out to be better. We decrease by 80% in time and RMSE becomes 10% worse.
We can see in our training result for all our models that random forest and support vector regression has lowest root mean squared error. Random forest has 1.41 rmse and support vector regression is 2.55 rmse. With these results we decide we want to use gridsearch on random forest and randomsearch for support vector regression. Now that we have the best hyperparameters for both of our best models we want to use it for the testset.

As RMSE is a mean measure, we possibly miss the distance between our data and the mean by using it.
To compare and check every single piece of actual data, prediction, and error visualization, we created two histograms, one for actual data and the other for our predicted values, and plotted them combined. The outcome is shown in Figure 4.We also utilized a boxplot to demonstrate how closely our predictions matched our actual data and how our model performed when compared to our precedents. You can see the results in Figure 5. Because they are so close together, our model performs well on the test set.

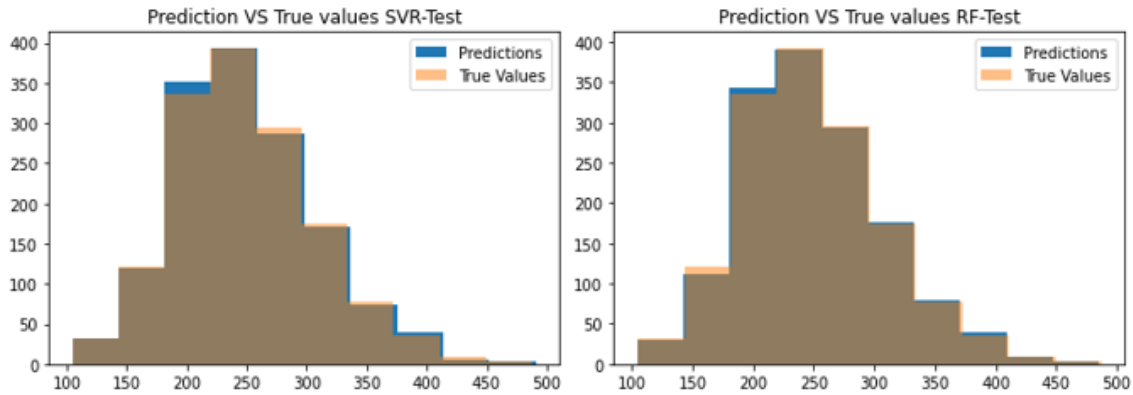


Figure 4

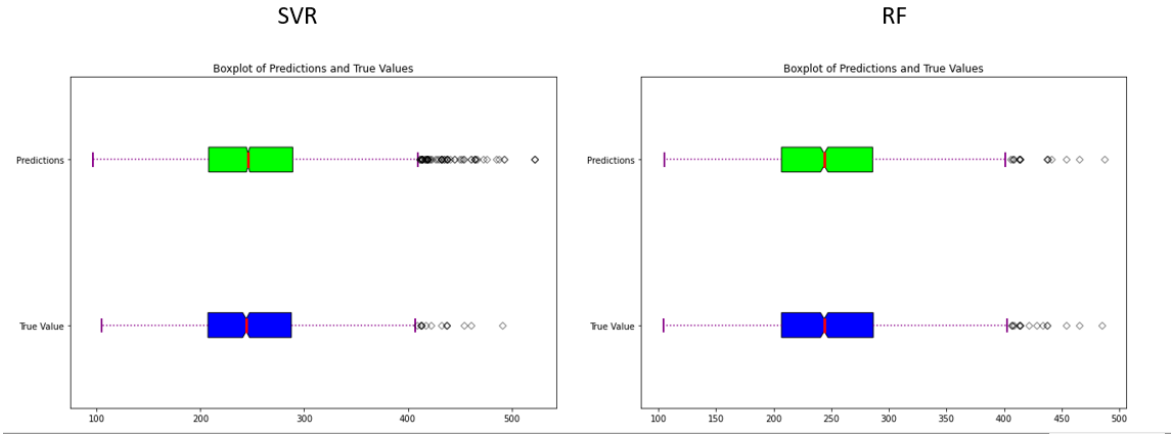


Figure 5

Result with the best hyperparameter for random forest.

- **RMSE: 3.1943621194739924**
- **R^2 : 0.9970334143438864**

Result with the best hyperparameter for support vector regression.

- **RMSE: 3.574434739958213**
- **R^2 : 0.9962854745091765**

.

Summary and further developments

In conclusion, our project on Co2 emissions by vehicle would be a useful tool for predicting a vehicle's carbon footprint based on its characteristics. With the help of accurate prediction, manufacturers can make informed decisions about global regulations and buyers about their purchases.

For further developments we think it would be interesting to cluster brands and get a better understanding on what kind of brand has lower co2 emission with the same fuel consumption. (Figure 6)

We can also incorporate more data sources, for example about traffic and driver behaviors. The analysis would be more comprehensive for factors that affect Co2 emissions.
We can try to use ensemble models for further improving the rmse.

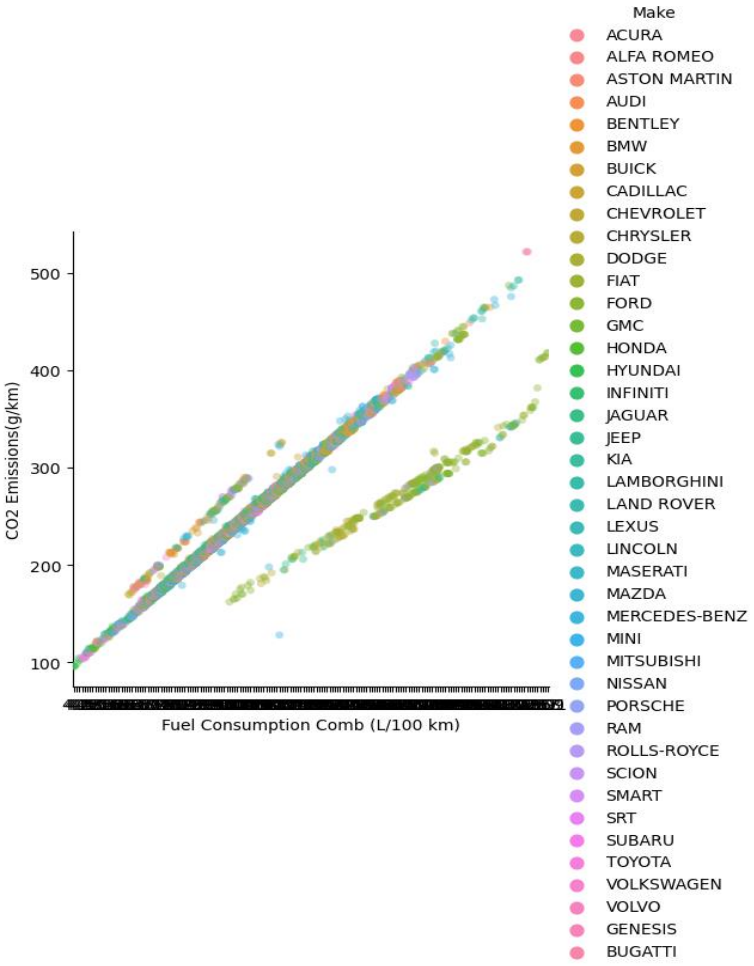


Figure 6