1. Introduction

As hotter summers and warmer winters become a common phenomenon, environmental protection has gained increasing global importance. Road transportation-related contamination, particularly from vehicles, is a main contributor to environmental deterioration. To reduce its impacts, characteristics of light-duty vehicles are identified to build a CO2 emissions prediction model.

The audience is assumed to have interest in statistics and machine learning, an application of artificial intelligence enabling systems to learn from experience without explicit programming. No prior knowledge is required.

The dataset consists of model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada. Vehicle manufacturers use controlled experimental procedures to generate fuel consumption data. Afterwards, Environment and Climate Change Canada and Natural Resources Canada collates information to publish the Fuel Consumption Guide for audiences including car purchasers, manufacturers, and government bodies.

The aim of this paper is to shortlist a machine learning algorithm with the greatest predictive power of carbon dioxide emissions based on critical features, such that consumers can compare and purchase vehicles with longer lifetime and lower maintenance costs, saving money in the long run.

2. Data

The data used spans 7 years of compiled features and CO2 emissions of various light-duty vehicles. There are 7385 rows of observations and 12 columns in total. The columns include 11 features: Make, Model, Vehicle Class, Engine Size, Cylinders, Transmission, Fuel Type, Fuel Consumption City, Fuel Consumption Hwy, Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), and 1 target variable: CO2 emissions. With a large sample size to features ratio, dimension reduction is not necessary. There are 5 non-numeric features which require further analysis to see if it can be manipulated or may otherwise be dropped. There are no null or missing values.

Data Visualization

As part of the data exploration, distributions and interactions of discrete and continuous variables and a correlation matrix is plotted and shown in the Appendix. Some notable insights:

- The top company is Ford
- Suv-small is the most common vehicle class
- Majority of vehicles have 8-16L/100km fuel consumption
- All features have a strong positive correlation with CO2 Emissions except for Fuel Consumption Combined which has a strong negative correlation with it
- Median engine size is 3L, while the mode is 2L. Usually, a larger engine is more powerful, accelerates faster, and has a higher speed, but it may burn more fuel and emit more CO2
- Data is centralized on 3 to 8 cylinders. The more cylinders, the less the horsepower and carbon emissions
- Regular (49%) and premium (43%) gasoline make up overwhelming majority of the data
- Strong positive relationship between Fuel Consumption Comb and CO2 Emission, though the data is split by two different slopes due to fuel type. Vehicles using ethanol emit less CO2

Data Preprocessing

Since linear regression and XGBoost require numerical values, one hot encoding is used to replace 5 categorical variables with 93 dummy variables. The “Model” column is dropped since there are 2053 unique values out of 7385. It would not provide much information besides the values in “Make” and would result in an overly complex model. Additionally, normalization is applied to avoid distortion of differences in range and to model the data accurately for linear regression and KNN. The new dataframe is shown in the Appendix.

3. Methodology

Models

Six algorithms are employed and compared to select one in terms of predictive power.

- Linear, Lasso, Ridge Regression

Linear regression is a simple and popular baseline model to evaluate the performances of more complex models. A linear relationship is assumed between input variables *x* and an output variable *y*. However, the model is not penalized for choice of weights. In order to prevent overfitting, two modifications, namely lasso, which penalizes the model for the sum of absolute values of the weights, and ridge, which penalizes for the sum of squared value of the weights, are used.

- K-Nearest Neighbors Algorithm

KNN is a supervised nonlinear learning algorithm. It provides an alternative if the true relationship is nonlinear and is also easy to interpret and scale to big data.The prediction of a new data point is the average of the *k* nearest data points' response variable. "Nearest" is defined by Euclidean distance, which is the length of the difference of two vectors.

- Extreme Gradient Boosting

XGBoost, or Extreme Gradient Boosting, is an ensemble method that aims to aggregate weak learning models to form a stronger and more robust estimator (Zhang et al., 2018). At each iteration, the residual of the previous estimator is used to learn and optimize the loss function. A binary decision tree is selected as a basic learner, and regularization added to avoid overfitting. Extreme refers to pushing computational limits to achieve gains in accuracy and speed.

- Random Forest

Random forest is powerful and accurate on a variety of problems, including features with nonlinear relationships. It is constructed of numerous individual decision trees that run independently to represent a feature class prediction. Each tree gives a predicted value and the forest averages all predictions as output as an ensemble.

All hyperparameters (except in linear regression) are currently following the default setting.

Performance Metrics

Three metrics are used to consistently evaluate and compare model performance.

Root-mean-square error = $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\hat{x}_i)^2}{n}}$

Mean absolute percentage error = $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right|$

Mean absolute error = $\frac{\sum_{i=1}^{n}\left|\hat{x}_i-x_i\right|}{n}$

10-fold cross validation is used to evaluate and compare model results. For each fold, 70% is used as a training set while the remaining 30% is used as a test set.

4. Analysis

Comparison of Algorithms

	Linear Regression	Lasso	Ridge	KNN	XGBoost	Random Forest
RMSE	5.3224	9.3545	5.302839	5.9344	3.5362	3.9224
MAPE (%)	2.9943	5.9765	3.011799	1.3808	0.9388	0.8339
MAE	1.2191	2.346942	1.221484	3.5064	2.3333	2.0493

Random forest has the best performance in terms of MAPE and MAE, whereas XGBoost performs best by RMSE. Hence, we will proceed with random forest as our prediction model.

Parameter Tuning

Hyperparameters are tuned to improve performance. GridSearchCV finds optimal parameters within a given range. Max_depth refers to maximum tree depth and balances local learning and overfitting. Max_features is the maximum number of features to consider at each split. N_estimators is the number of trees in the forest to average over. A higher number will improve but slow down performance. The best result is achieved by max_depth = 19, max_estimators = 9, n_estimators = 125.

5. Results

Random forest was the most accurate among six models. As shown in the Appendix, after parameter tuning of max_depth, max_estimators, and n_estimators, the performance has deteriorated. Hence, it is suggested to use the default hyperparameters or test other parameter sets for tuning.

In terms of feature importance, combined fuel consumption (55% city + 45% highway) is the most relevant variable to accurately predicting CO2 emissions, and to a lesser extent engine size. This is expected as the more fuel required to travel a given distance, the more gasses would be emitted.

6. Conclusion

An effective model was proposed to predict carbon dioxide emissions of light-duty vehicles in this paper. Exploratory data analysis was performed for a thorough understanding of the dataset, as well as data preprocessing to improve accuracy. Four algorithms were used for model development including linear regression, KNN, XGBoost, and random forest. After assessing the root-mean-square error, mean absolute percentage error, and mean absolute error of each algorithm, random forest conclusively produces the best results. Combined fuel consumption and engine size are critical prediction features.

7. Limitations and Suggestions

Besides the three selected hyperparameters that were tuned for in random forest, others such as min_samples_split and min_samples_leaf can be explored to improve model performance.

Periodic model update is recommended as annual publications are provided by governments on the fuel consumption data. Revalidation ensures accurate forecasting and provides indication of whether model redevelopment would be appropriate.

Large engine size is correlated with higher CO2 emissions and are usually less efficient than small engines. One solution is to replace large engines with small turbocharged engines, generating more power and torque but limiting emissions.

Fuel consumptions and carbon emissions data are presented annually by vehicle manufacturers along with government departments. CO2 is the most important greenhouse gas, and accurate predictions and future research allow for significant mitigation and long-term planning through environmental purchases, policy setting, and fuel-efficient technological innovation.

8. References

Bakshi, C. (2022, April 14). *Random Forest regression*. Medium. Retrieved October 12, 2022, from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84#:~:text=Random%20Forest%20Regression%20is%20a%20supervised%20learning%20algorithm%20that%20uses,prediction%20than%20a%20single%20model.

Canada, N. R. (2022, August 22). *Government of Canada*. Natural Resources Canada. Retrieved October 12, 2022, from https://www.nrcan.gc.ca/energy-efficiency/transportation-alternative-fuels/fuel-consumption-guide/21002

Car ownership guides | webuyanycar.com. (n.d.). Retrieved October 11, 2022, from https://www.webuyanycar.com/guides/car-ownership/

Deepanshi. (2021, May 25). *Linear regression: Introduction to linear regression for data science*. Analytics Vidhya. Retrieved October 12, 2022, from https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/

Garba, M. D., Usman, M., Khan, S., Shehzad, F., Galadima, A., Ehsan, M. F., Ghanem, A. S., & Humayun, M. (2020, November 14). *CO2 towards fuels: A review of catalytic conversion of carbon dioxide to hydrocarbons*. Journal of Environmental Chemical Engineering. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S2213343720311052

Medium. (n.d.). Retrieved October 18, 2022, from https://towardsdatascience.com/linear-regression-models-4a3d14b8d368

Real Python. (2022, September 1). *The K-nearest neighbors (knn) algorithm in Python*. Real Python. Retrieved October 12, 2022, from https://realpython.com/knn-python/#a-step-by-step-knn-from-scratch-in-python

RMSE: Root mean square error. Statistics How To. (2021, May 31). Retrieved October 12, 2022, from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Selig, J. (2022, July 4). *What is machine learning? A definition.* Expert.ai. Retrieved October 12, 2022, from https://www.expert.ai/blog/machine-learning-definition/

Sklearn.ensemble.randomforestregressor. scikit. (n.d.). Retrieved October 12, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

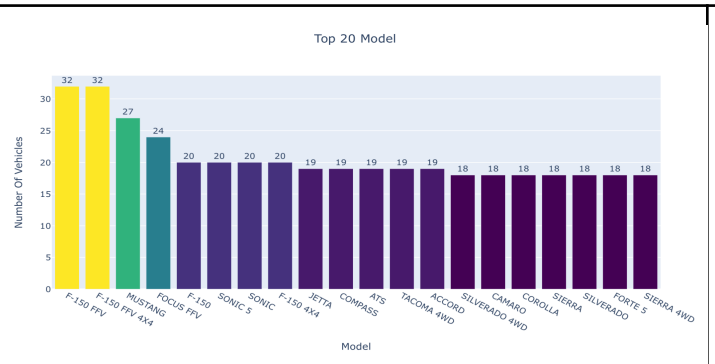
Stephanie. (2022, May 27). *Mean absolute percentage error (MAPE)*. Statistics How To. Retrieved October 12, 2022, from https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

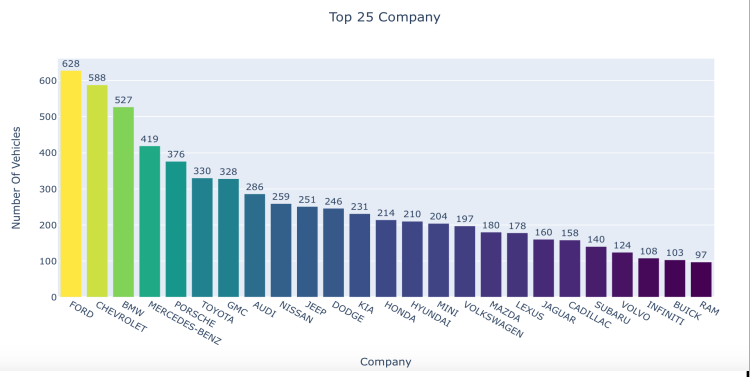
Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and xgboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/access.2018.2818678

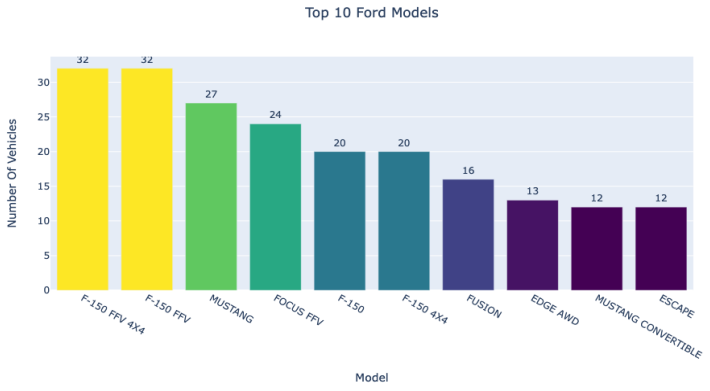
9. Appendix

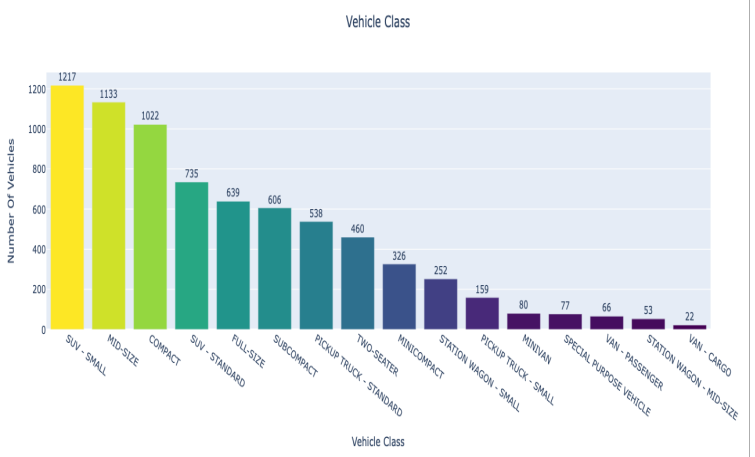
Data Visualization

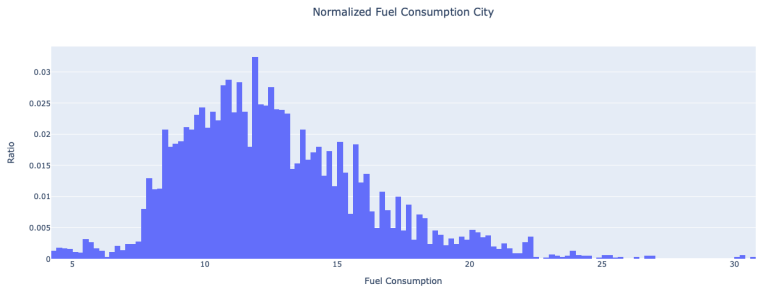
Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	6.7	8.5	33	196
ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6	5.8	5.9	48	136
ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10	28	230
ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	8.1	10.1	28	232
ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9	11.1	25	255
ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24	267
ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	7.5	9.2	31	212
ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2	8.1	9.8	29	225
ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1	8.3	10.4	27	239
ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	Z	9.7	6.9	8.4	34	193
ASTON MARTIN	DB9	MINICOMPACT	5.9	12	A6	Z	18	12.6	15.6	18	359
ASTON MARTIN	RAPIDE	SUBCOMPACT	5.9	12	A6	Z	18	12.6	15.6	18	359
ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	AM7	Z	17.4	11.3	14.7	19	338
ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	M6	Z	18.1	12.2	15.4	18	354
ASTON MARTIN	V8 VANTAGE S	TWO-SEATER	4.7	8	AM7	Z	17.4	11.3	14.7	19	338
ASTON MARTIN	V8 VANTAGE S	TWO-SEATER	4.7	8	M6	Z	18.1	12.2	15.4	18	354

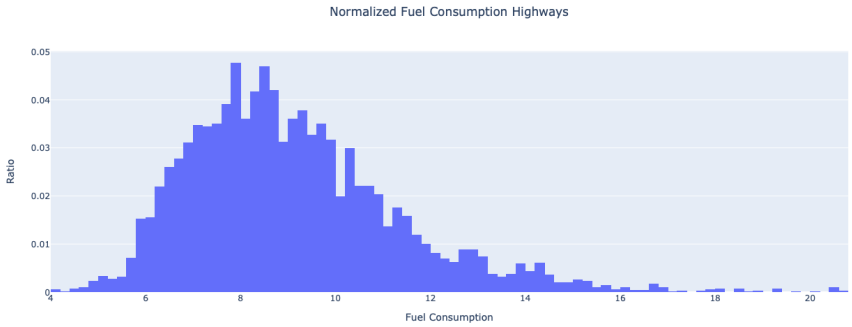


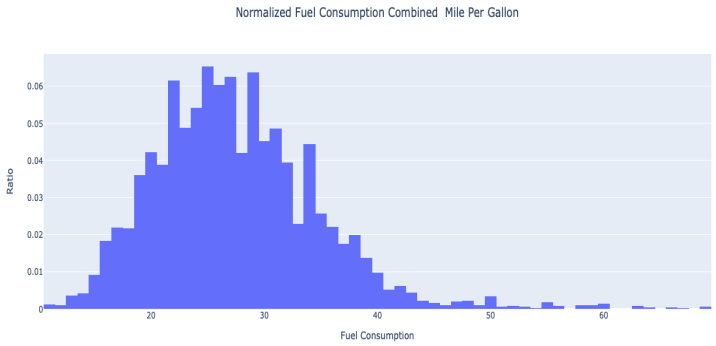


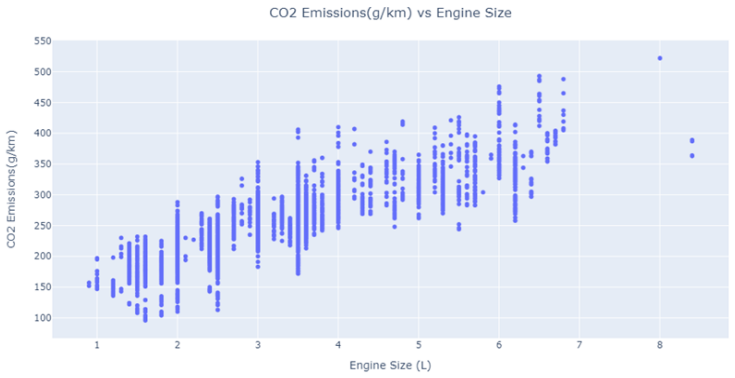


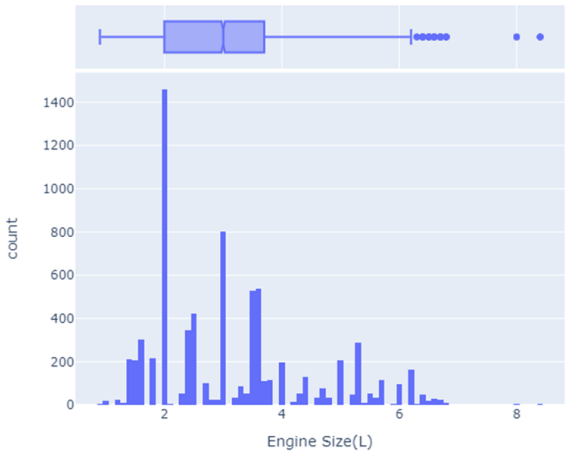


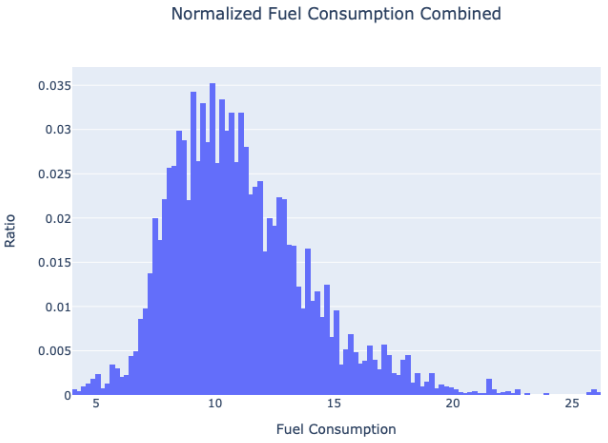

Normalized Fuel Consumption City
Ratio
0.03
0.025
0.02
0.015
0.01
0.005
0
5
10
15
20
25
30
Fuel Consumption

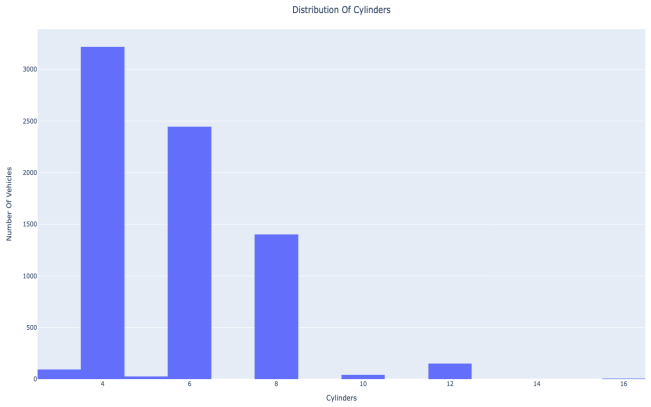

Normalized Fuel Consumption Highways
Ratio
0.05
0.04
0.03
0.02
0.01
0
4
6
8
10
12
14
16
18
20
Fuel Consumption

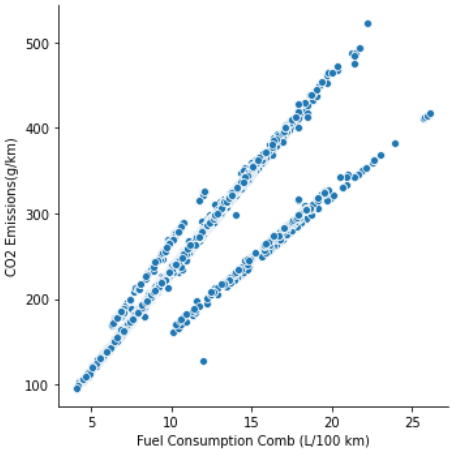

Normalized Fuel Consumption Combined Mile Per Gallon
Ratio
0.06
0.05
0.04
0.03
0.02
0.01
0
20
30
40
50
60
Fuel Consumption

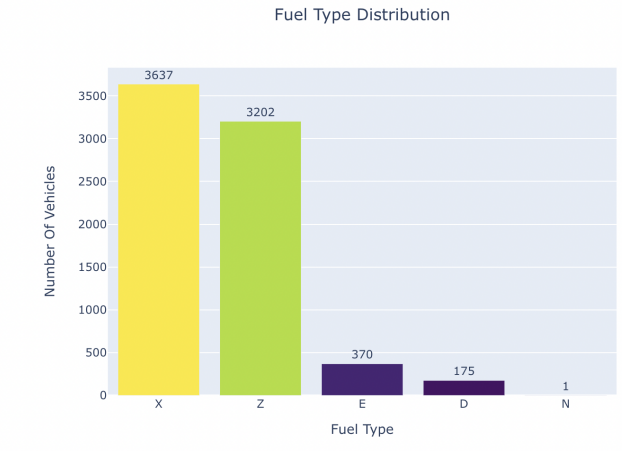

CO2 Emissions(g/km) vs Engine Size
CO2 Emissions(g/km)
550
500
450
400
350
300
250
200
150
100
1
2
3
4
5
6
7
8
Engine Size (L)

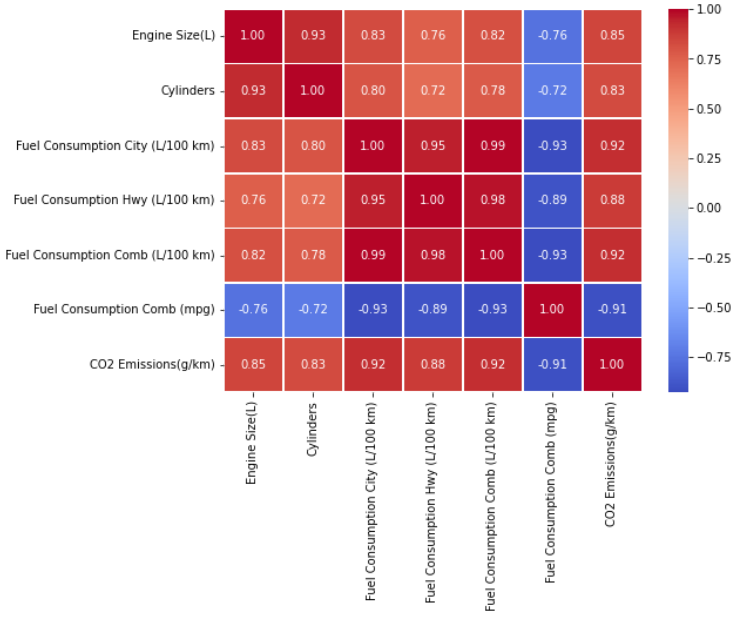

count
Engine Size(L)

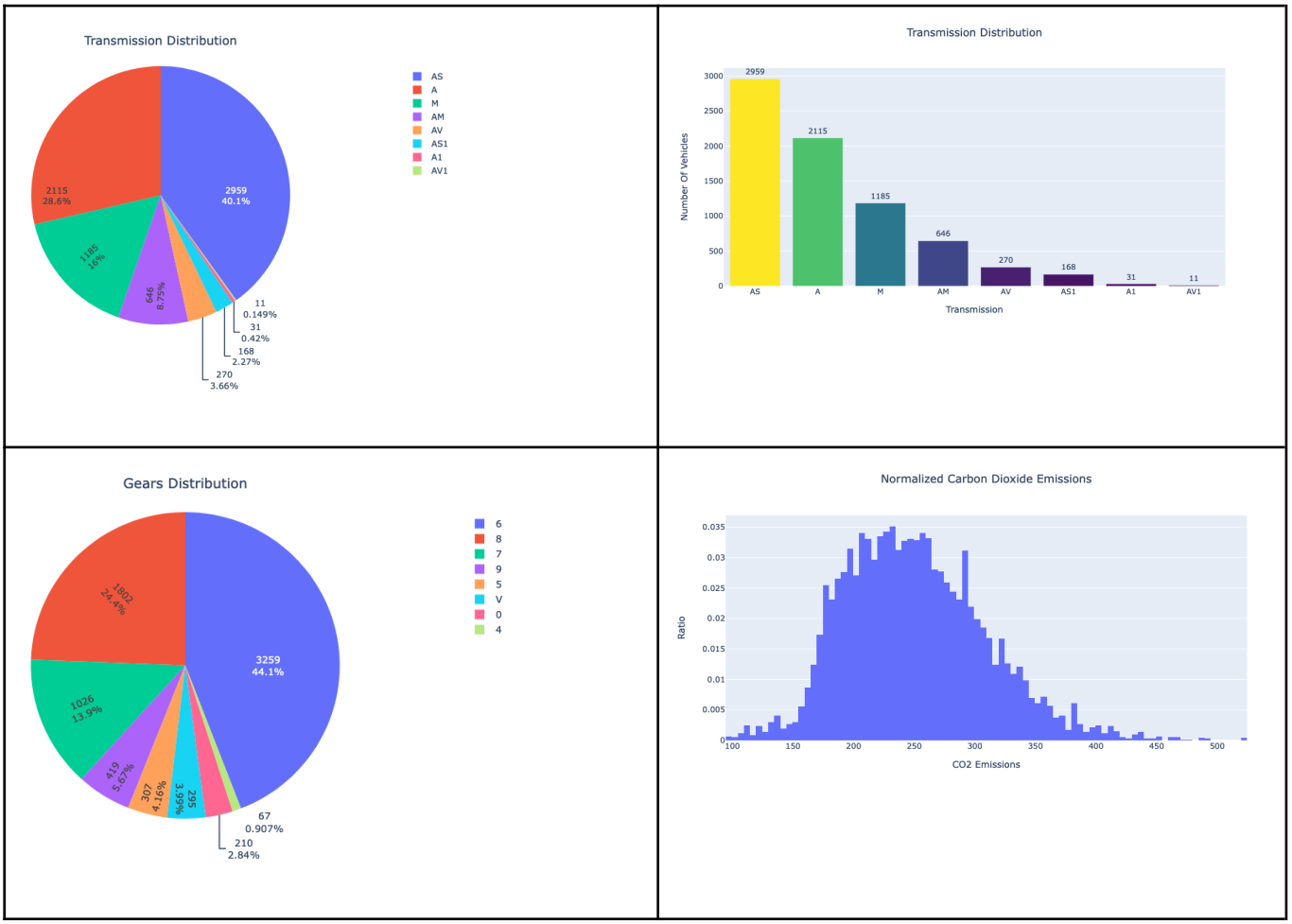

Normalized Fuel Consumption Combined
Ratio
Fuel Consumption


Distribution Of Cylinders
Number Of Vehicles
Cylinders


CO2 Emissions(g/km)
Fuel Consumption Comb (L/100 km)


Fuel Type Distribution
Number Of Vehicles
3637
3202
370
175
1
X
Z
E
D
N
Fuel Type


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)



Data Preprocessing

New Dataset

	Engine Size(L)	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	Gears	Make_ALFA ROMEO	Make_ASTON MARTIN	Make_AUDI	Make_BENTLEY	Make_BMW	Make_BUG
0	2.0	9.9	6.7	8.5	33	5	0	0	0	0	0	
1	2.4	11.2	7.7	9.6	29	6	0	0	0	0	0	
2	1.5	6.0	5.8	5.9	48	7	0	0	0	0	0	
3	3.5	12.7	9.1	11.1	25	6	0	0	0	0	0	
4	3.5	12.1	8.7	10.6	27	6	0	0	0	0	0	

Model Performance

Linear Regression

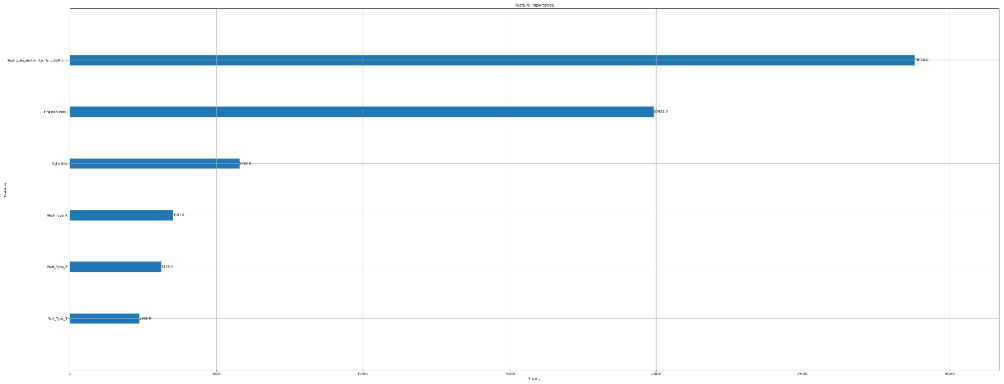
	Linear regression	Ridge regression	Lasso regression
Before tuning	0.9936	0.9753	0.9935

After tuning	NA	0.993477	0.9935

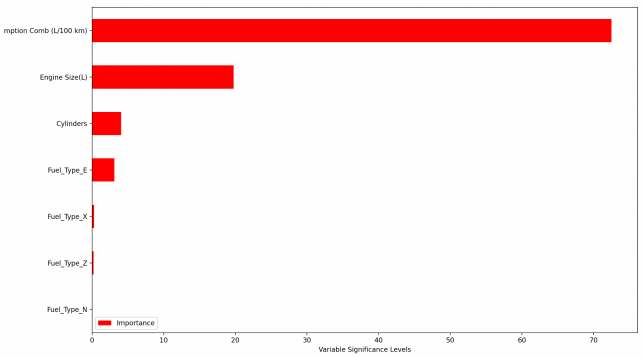
KNN

	Default, k=5	Tuned, k=1	Tuned with bagging(100 estimators)
RMSE	11.421831	7.598125	4.815121
MAPE	7.228700	1.476974	1.199203
MAE	2.992004	3.169585	2.904765

XGBoost



Random Forest



RMSE	MAPE (%)	MAE
14.4933	4.0843	9.4356

Random forest performance after tuning