Predicting Passenger Vehicles CO2 Emissions and Fuel MPG

Machine Learning For Business Decision

I. Introduction

Climate change has boosted the negative impact of climate-related weather events globally. For example, an estimated 3.4 million people in the US have been displaced in 2022. Caused by a series of major disasters in 2022, 18 extreme weather events had caused at least $1 billion in damages, hitting most in Louisiana and Florida. With the impact of Greenhouse gas (GHG) and other pollutants, climate experts expect more intense weather disasters as global temperatures rise (Arduengo, 2023). With the unprecedented events in the US, the Biden Administration has set new goals together with the US Environmental Protection Agency (EPA) to set a 40 Miles Per Gallon (MPG) Fuel Efficiency Average for 2026. That is 60% higher than the 2020 average of 24.9 mpg. By increasing the fuel efficiency of cars on the road, it results in the reduction of fossil fuels burned by internal combustion engines. To hit this goal, this would require automakers to build almost 2 million more EVs and PHEVs each year by 2026 by EPA analysts. This incentivises automakers to also optimize currently manufactured engines to become more efficient (O'Dell, 2022).

With this, we have generated goals to help car manufacturers to identify what current cars in production are able to meet this EPA target mpg already and identify which car features affect CO2 emissions the most. Lastly, we hope that this modeling and analysis help car manufacturers develop future cars that are able to comply with emissions, mpg and consumer safety targets. Furthermore, governments, related organizations and other parties may be able to use our models in measuring, studying and observing the local population.

II. Related Work

There have been countless other articles and journals in regards to the fuel consumption of different types of vehicles e.g. gasoline and electric. For example, Yang et al. (2022) explores the use of Machine Learning and Multi-dimensional Big Data to predict the fuel consumption of gasoline vehicles as well as its environmental impact. Data being sourced from an app, information of cars such as Brand Name, Engine, Gearbox (Transmission) and more are used for their regression machine learning model. The following models were used to predict Fuel Consumption (Liter per 100 Km): Linear Regression, Naive Bayes, Neural Network, RandomForest and LightGBM. . Furthermore, according to Yin et al. (2015), they were able to classify vehicles into 3 groups: high, medium and low fuel efficiency with the Support Vector Machine model. Similar to our data set, each row of their data contains vehicle information such as fuel type, engine displacement, transmission type and more. Given these two examples, their methodologies are a similar process to that we have proposed and should serve as justification of the feasibility of this project. However,

III. Data Sources

We will be using 2 public data sources for our machine learning models, one serving as our training and test set while the second data source will serve as our validation data set. The first data set is sourced from the EPA fuel economy website. This data set is called “Model Year 2023 Green Vehicle Guide”, it contains 18 attributes that contains details of different car models. Details such as Displacement, No. of cylinders, Transmission type, Fuel type along with measurements of MPG and CO2 and their scoring (EPA, 2023). Our second data source is from the Government of Canada, published by Natural Resources Canada. Similar to the first data set, the data contains details of different car models, CO2 emissions and their fuel economy measurements for 2023. However, they have decided to segmentize car models by type of fuel source e.g. battery-electric, plug-in hybrid and gasoline (Natural Resources Canada, 2013). This means that we have had to consolidate 3 different .csv files into one which will be brought into further detail in the report. Otherwise, the data within all files contain similar attributes to the US sourced data set.

IV. Problem Definition and Solution

We have chosen to conduct 3 different analyses for this project. Feature Selection will be conducted to identify which variable (car feature) impacts fuel efficiency and CO2 emissions of a passenger vehicle e.g. Transmission type, Engine Size, Fuel type or Vehicle class. Furthermore, we hope that this feature selection analysis can help improve the performance of the consequent analysis that we are about to outline. The second analysis is a Classification model. We aim to create a machine learning model that would be able to predict if a car with given parameters would meet the baseline target of 40 mpg (Yes or No). This would require us to generate a new label attribute for our datasets, a boolean/nominal data type. Lastly, we will develop a regression model to predict CO2 emissions generated by newer car models based on given parameters. These two models can help car manufacturers estimate the potential CO2 emissions of cars lined up for release in the upcoming years as well as whether it can meet the target mpg value. Other than this proposed use, the models can be potentially used in data mining practices for governments and countries, assessing the population's cars, what cars on the road are contributing the most to CO2 emissions, and having poor efficiency. This can help governments create initiatives to promote cars with better MPG ratings as well as anti-pollutant initiatives. These analyses will be conducted with the use of RapidMiner and its functionalities

V. Feature Selection

The feature selection process was integrated with both classification and regression processes. The process itself is within a cross-validation for the respective classification models that we will discuss in the next section. However, before the models are used, backward elimination for feature selection. Backward elimination begins with the full set of attributes of the data set and in each iteration, an attribute is removed. For each removed attribute, the performance is measured. Whichever attribute gave the least decrease of performance is removed from the whole set before a new iteration is started with the modified set. This process ends when it meets one of the following situations:-

- The iteration runs as long as there is any increase in performance.
- The iteration runs as long as the decrease is less than the specified threshold, either relative or absolute. (This can be specified in the operator).

- The iteration stops as soon as the decrease is significant to the level specified by the *alpha* parameter.

We have chosen the first situation. With this, the feature selection process resulted in keeping Displacement, Transmission type, Drive type, Fuel type, Comb CO2 and Vehicle Class. This implies that, in determining which car features would affect the fuel efficiency of a car, its engine displacement, transmission type, drive type, fuel type and CO2 emissions are shown to be impactful. These findings were not that surprising as these attributes are what make up how a vehicle and its engine run, how much fuel is going to be consumed and the main internal factors that affect the performance of an engine, therefore its efficiency.

VI. Experimental Evaluation

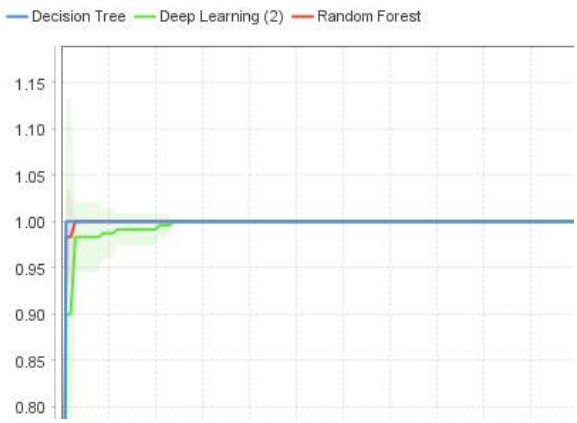
A. Classification Model

For the classification model, the data set was cleaned by the following processes:

- Removing low quality columns such as Model, Make, City MPG
- Parse clc and disp; to replace non-numerical values into NA
- Replacing Missing Values
- Generating a new attribute “Target_yn”, if Cmb MPG >= 40, TRUE, FALSE
- Remove Correlated Attributes

By using the TurboPrep function in Rapidminer, most attributes were removed for low quality. Other attributes were removed for redundancy e.g. City MPG and Hwy MPG as they are highly correlated to Cmb MPG. We also parse the Cmb MPG, cycl and disp attributes to replace non-numerical values into NA then those missing values are then replaced to zero. (This is because the non-numerical values in disp and cyl are the entries for electric vehicles that do not have displacement and cylinder values for engines. We generated an attribute of whether each entry meets the EPA criteria of 40 mpg using an IF function and set this as our label. We then split the data into a 80:20 training to training set with an automatic sampling type.

The training set is then sent to 3 different cross-validation operators at 10 folds and automatic sampling. These operators have our models of RandomForest, Deep Learning and Decision Tree. These models are chosen for their capability of handling missing values to allow the model to be adaptable to different datasets. Furthermore, as previously mentioned, the cross-validation process also contains backwards elimination to conduct feature selection before it is applied to the models we previously mentioned. The training set is also used to produce ROC charts which is the True Positive Rate over False Positive Rate.



From the ROC curves, Decision Tree shows to be a "perfect" classifier, being the most accurate while Deep Learning being the least accurate and RandomForest being in the middle. We then use the "Apply Model" operator to use the trained model on our training set. Furthermore, the trained model is stored to be used on the validation set which will be discussed shortly. Lastly, we output the performance of the model with the test set. Performance is measured with the use of the Confusion Matrix, outputting True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). We will be using Accuracy, Precision and Recall (Sensitivity) as our evaluation metrics for our classification model. We will set the premise that False Positives would be of higher concern than false negatives as we do not want passenger vehicles to be manufactured not being able to meet the EPA target. The following tables are the confusion matrices of our 3 models on both training and test set (where TRUE = Car meets EPA Target).

RandomForest			
Training Set			
classification_error: 0.16% +/- 0.35%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	1687	2	99.88%
Predicted True	1	235	99.58%
Class Recall	99.94%	99.16%	
Test Set			
accuracy: 99.79%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	421	0	100.00%
Predicted True	1	59	98.33%
Class Recall	99.76%	100.00%	

The results of our RandomForest show that during training, the model returned a classification of 0.16% +/- 0.35% or a micro average of 0.16%. Precision of predicting false at 99.88% and predicting true at 99.58% whereas. Furthermore, the sensitivity of the model shows that a 99.94% at predicting TN while predicting TP is at a 99.16%. The test set reflects this performance, giving a 99.79% accuracy overall. However, it produced 1 False Positive which can be a detriment to using the model in real life situations.

Deep Learning			
Training Set			
classification_error: 0.31% +/- 0.44%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	1686	4	99.76%
Predicted True	2	233	99.15%
Class Recall	99.88%	98.31%	
Test Set			
accuracy: 98.96%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	420	3	99.29%
Predicted True	2	56	96.55%
Class Recall	99.53%	94.92%	

The results of our Deep Learning model show that during training, the model returned a classification of 0.31% +/- 0.44% or a micro average of 0.16%. Precision of predicting false at 99.88% and predicting true at 99.58% whereas. Furthermore, the sensitivity of the model shows that a 99.94% at predicting TN while predicting TP is at a 99.16%. In comparison, the test set performance is relatively the same as the training performance. However, the sensitivity of predicting Actual False observations correctly has dipped by 3%. This means that with unseen data, it may be more likely to produce False Positives.

Decision Tree			
Training Set			
classification_error: 0.36% +/- 0.81%)	**Actual False**	**Actual True**	**Class Precision**
Predicted False	1687	6	99.65%
Predicted True	1	231	99.57%
Class Recall	99.94%	97.47%	
Test Set			
accuracy: 99.79%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	421	0	100.00%
Predicted True	1	59	98.33%
Class Recall	99.76%	100.00%	

Lastly, we have the Decision Tree. For the training set, the model had 6 False Positives and 1 True Negative. This is the poorest performance in regards to the training set, however in comparison to the Test set, it has performed almost perfectly. Similar to RandomForest where it outputted 1 False Positive, the model has performed extremely well on the test set.

To further test our models, we are going to conduct a validation test using the Canada Cars data set. However, to allow the model to understand the new dataset, we need to map certain attributes to be similar to the data it was trained on. The following changes have been made:

- Generate new attribute: "Target_yn", if Cmb MPG >= 40, TRUE, FALSE
- Rename attributes to Fuel, Displ, Comb MPG and Comb CO2
- Values in Fuel and Veh Class are mapped to values in the US dataset
- Generate an empty attribute of Drive as the dataset does not contain it

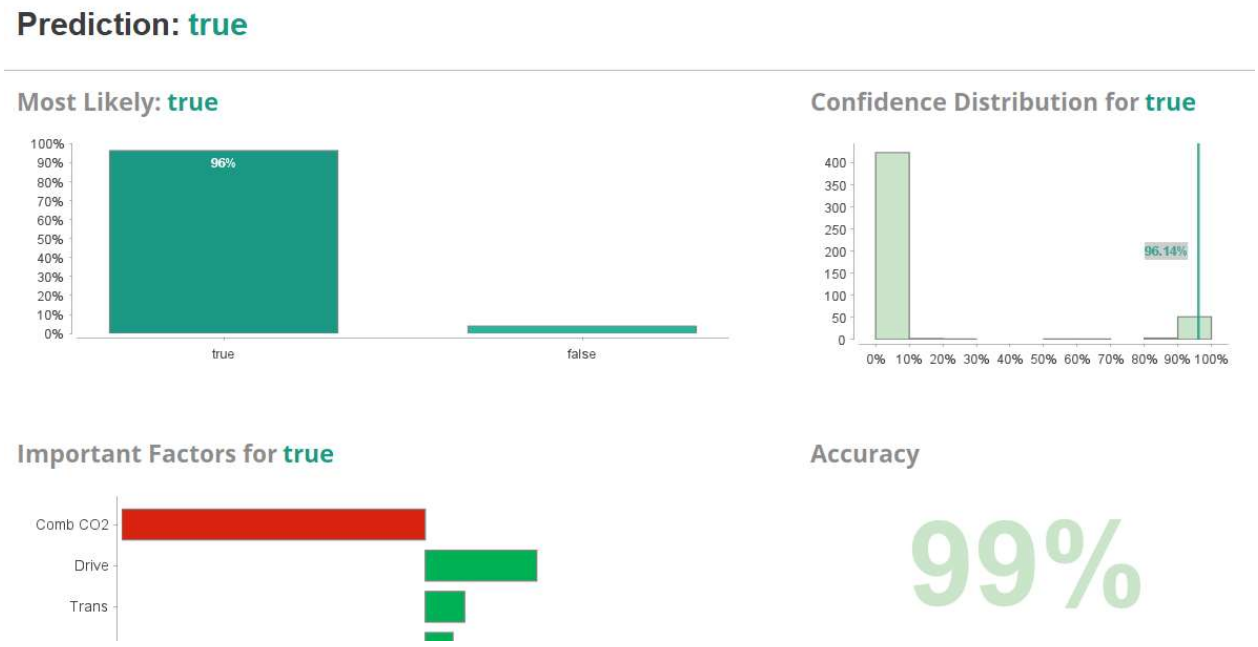
To preface, the purpose of the validation set is to see how generalisable our models are to using data from other countries within the limitations of using the same measurement system as well as mapping attributes to the same values as closely as possible. For example, the Canada dataset's trans attribute is hard to discern and complicated to map to the US one. Furthermore, the Drive attribute is also not present in the Canada dataset. With this, the following tables are our results:

RandomForest			
Validation Set			
accuracy: 54.85%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	509	239	68.05%
Predicted True	213	40	15.81%
Class Recall	70.50%	14.34%	

Deep Learning			
Validation Set			
accuracy: 91.51%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	722	85	89.47%
Predicted True	0	194	100.00%
Class Recall	100.00%	69.53%	

Decision Tree			
Training Set			
accuracy: 71.93%	**Actual False**	**Actual True**	**Class Precision**
Predicted False	509	68	88.21%
Predicted True	213	211	49.76%
Class Recall	70.50%	75.63%	

Based on the measurements above, the accuracy for all models have decreased, RandomForest being the most significantly affected. Furthermore, the number of False Positives has extremely increased for Decision Tree and RandomForest, having a class recall of 70.5% for both. On the other hand, Deep Learning is still performing extremely well, with an accuracy of 91.51% as well as being able to predict 100% of all cars in the dataset that did not meet the Fuel Efficiency target. However, it has a tendency to generate False Negatives which inhibits the accuracy from increasing any further. To conclude, the Deep Learning model performs the best with unseen data therefore, we can choose the Deep Learning model as our final model for implementation and deployment.



We can also point out more interesting findings by using the model simulator as seen above (this uses the Deep Learning model on the training set):

- If the fuel type is "Electricity", it is guaranteed to be meeting the fuel efficiency target
- 4WD (4 wheel drive) decreases the fuel efficiency of a gasoline vehicle, significantly influencing whether a vehicle meets the target
- For each different combination of car features, the measurement of CO2 emissions by the vehicle remains the most important factor for True or False

B. Regression Model

For the regression model, the pre-processed data after feature selection is used with some further cleaning steps as follow:

- Set CO2 emissions as the target role
- Convert engine displacement, transmission type, drive type, fuel type and vehicle class from nominal to numerical data type

After pre-processing and cleaning, the data was split into 80:20 training to test with automatic sampling type. The training set is then sent to 3 cross-validation operators at 10 folds and automatic sampling. The 3 regression models that were tested are Liner Regression, k-Nearest Neighbours and Deep Learning. These models are chosen to be suitable to perform regression for our dataset that have numerical data. In the cross-validation operators, backward elimination is performed to further filter and select attributes that are highly significant to the models. The backward elimination operator is a nested process that does cross-validation within itself to perform feature selection.

After cross-validation, the dataset is multiplied to be sent to the 'Apply Model' operator to test the dataset based on the trained model. The trained model is also stored to be used in the validation process later that will test Canada's dataset. Finally, the performance of the models are outputted to compare and decide on the best regression model for our dataset. Performance is measured and evaluated based on the metrics such as root mean squared error, absolute error, squared correlation and prediction average. The model with best performance metric is chosen based on guidelines as below:

Performance metric	Description
Root Mean Square Error (RMSE)	Choose the model with the minimum value of the Root Mean Square Error if average values are preferred. Average values give more weight to outliers, as explained above.
Average Absolute Error	Choose the model with the minimum value of the Average Absolute Error if median values are preferred. Median values give less weight to outliers, as explained above.
Squared Correlation (R2)	Look for a high value of R2 (close to 1), indicating a high correlation between predicted values and actual values.

The following tables are the performance matrices of our 3 models on both training and test set.

k-NN	
Training Set	
Root Mean Squared Error	42.673 +/- 5.888
Absolute Error	29.286 +/- 3.285
Squared Correlation	0.875 +/- 0.023
Prediction Average	398.415 +/- 10.407
Test Set	
Root Mean Squared Error	38.654 +/- 0.000
Absolute Error	26.614 +/- 28.034

Squared Correlation	0.877
Prediction Average	405.320 +/- 109.899

The results of our 5-Nearest Neighbour model shows that during training, the model returned a RMSE of 42.68 and absolute error of 29.29. Furthermore, the model has squared correlation of 0.875 and prediction average of 398. Meanwhile, the test set reflects a better performance compared to the training model in terms of lower RMSE and absolute error and higher squared correlation.

Linear Regression	
Training Set	
Root Mean Squared Error	70.849 +/- 5.209
Absolute Error	50.707 +/- 3.471
Squared Correlation	0.656 +/- 0.045
Prediction Average	398.413 +/- 9.079
Test Set	
Root Mean Squared Error	68.241 +/- 0.000
Absolute Error	48.691 +/- 47.812
Squared Correlation	0.617
Prediction Average	405.320 +/- 109.899

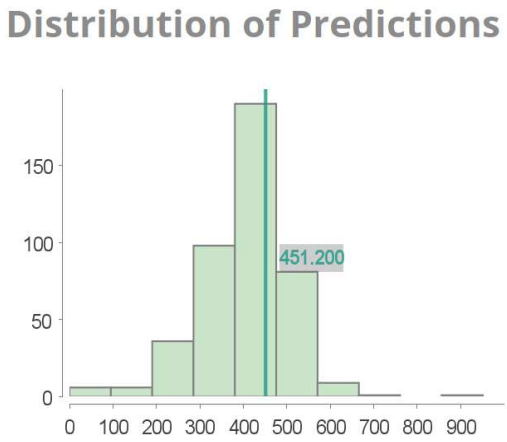
Next, Linear Regression model has a output of 70.85 RMSE and absolute error of 50.71. Furthermore, the model has squared correlation of 0.656 and prediction average of 398. Comparatively, the test set has lower RMSE, absolute error as well as lower squared correlation.

Deep Learning	
Training Set	
Root Mean Squared Error	71.504 +/- 10.301
Absolute Error	55.222 +/- 8.481
Squared Correlation	0.717 +/- 0.058
Prediction Average	398.442 +/- 13.233

Test Set	
Root Mean Squared Error	64.897 +/- 0.000
Absolute Error	49.194 +/- 42.326
Squared Correlation	0.670
Prediction Average	05.320 +/- 109.899

Our third regression model, Deep Learning, returned a RMSE of 71.51 and absolute error of 55.22. Furthermore, the model has a squared correlation of 0.717 and prediction average of 398. Similar to Linear Regression, the test set has lower RMSE, absolute error as well as lower squared correlation.

As mentioned before, the 3 regression models are compared based on RMSE, absolute error and squared correlation. Among the 3 models, k-NN has the lowest RMSE and absolute error, and highest squared correlation in both train and test models. Also, it has to be noted that the test set of the k-NN model has performed better than the train model which implies this regression technique to be the best among the 3 models. The histogram below shows the prediction distribution of the k-NN model that was used to test the dataset. It shows a clear normal distribution of prediction of CO2 emissions based on the test dataset.



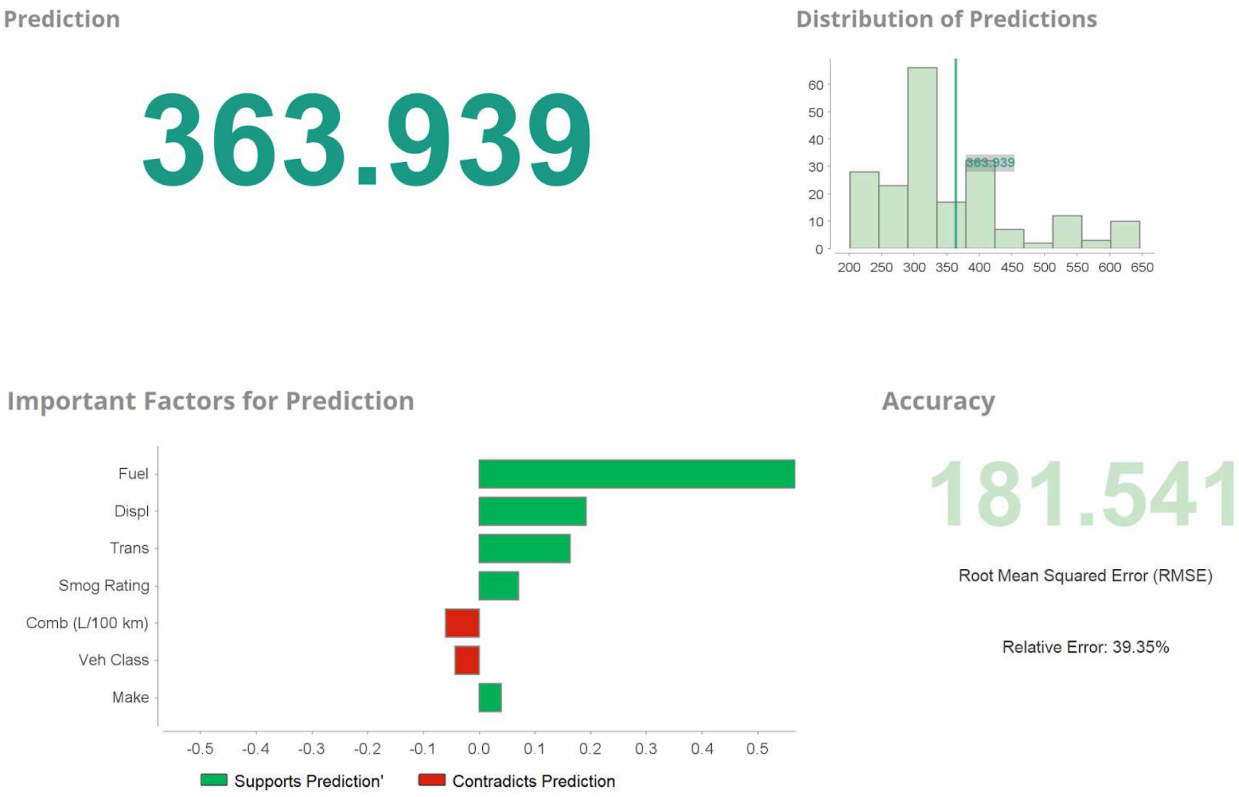
To further test the models, a validation test was conducted using Canada Cars data set after doing some mapping of attributes as mentioned in the Classification Model above.

k-NN	
Root Mean Squared Error	162.471 +/- 0.000
Absolute Error	143.050 +/- 77.029
Squared Correlation	0.440
Prediction Average	203.099 +/- 115.542

Deep Learning	
Root Mean Squared Error	183.296 +/- 0.000
Absolute Error	144.451 +/- 112.832
Squared Correlation	0.150
Prediction Average	203.099 +/- 115.542

Linear Regression	
Root Mean Squared Error	unknown
Absolute Error	unknown
Squared Correlation	0.000
Prediction Average	unknown

Based on the output above of validation data, the performance matrices of all models show a significant reduction compared to the training model. The most affected model is Linear Regression which has failed to perform due to the empty attribute generated, Drive. It shows limitations of this technique, hence it is not best for datasets with missing values/attributes. When comparing k-NN and Deep Learning, k-NN has slightly lower RMSE and absolute error. In terms of squared correlation, k-NN has performed better again with higher correlation compared to Deep Learning. In conclusion, similar to the training model, the validation model of k-NN has proved to be the best regression model for our dataset.



Some interesting findings from the model simulator of k-NN model are as follows:

- The biggest support for this decision is coming from **Fuel**.
- The histogram distribution of CO2 emissions prediction shows positively-skewed where many of the values are near the lower end of the range, and higher values are infrequent.

VII. Conclusion

A. Key Findings

Based on our above analysis, deep Learning was found to be most generalizable to the population for classification, usable for data that may be missing certain attributes the model has trained on, therefore can be used in different cases. It also generated zero false positives which is a very good trait as we want to minimize the amount of cars incorrectly identified as meeting the target.

Second to that, K-NN can be classified as the best model for Regression, as it has lower error and higher squared correlation in comparison to the other models.

Finally, it was identified that Displacement, Transmission type, Drive type, Fuel type and Veh Class were most impactful to the fuel efficiency and Co2 emissions of a vehicle, through the feature selection process, after eliminating the variables that would not be useful to meeting our initial goals.

B. Future Improvements and Limitations

To conclude, we believe there are future improvements that can be taken into consideration along with the limitations that may arise with our current data set:

- Updating the model each year to keep up with current technologies
    - With current innovations in hybrid-electric cars, the model may not be able to accurately depict cars 1-2 years from now, given the rapid change and improvements in technology.
- Integrating driver habits and behaviour in the future.
    - There are external factors outside of the car that are also key variables that can be taken into consideration in the MPG and CO2 emissions of a vehicle.
    - These factors could range from an individual's driving habits, frequency of use and fuel economy consciousness, which are important to consider aside from the car models on their own.
- It is also pivotal to ensure that the data inputted into the model must meet the same measurements (i.e. metric system, the same value notation for attributes) as the training data .

Reference list

Arduengo, R. (2023). *Natural disasters, boosted by climate change, displaced millions of people in U.S. in 2022*. [online] NBC News. Available at: https://www.nbcnews.com/science/environment/natural-disasters-boosted-climate-change-displaced-millions-americans-rcna69732 [Accessed 13 Feb. 2023].

EPA (2023). *Download Fuel Economy Data*. [online] www.fueleconomy.gov. Available at: https://www.fueleconomy.gov/feg/download.shtml EPA Fuel Economy data set.

Natural Resources Canada (2013). *Fuel consumption ratings - Open Government Portal*. [online] Canada.ca. Available at: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64.

O'Dell, J. (2022). *EPA Sets 40 MPG Fuel Efficiency Average For 2026*. [online] Forbes Wheels. Available at: https://www.forbes.com/wheels/news/new-fuel-efficiency-standards-epa/ [Accessed 13 Feb. 2023].

Yang, Y., Gong, N., Xie, K. and Liu, Q. (2022). Predicting Gasoline Vehicle Fuel Consumption in Energy and Environmental Impact Based on Machine Learning and Multidimensional Big Data. *Energies*, [online] 15(5), p.1602. doi:https://doi.org/10.3390/en15051602.

Yin, X., Li, Z., Shah, S.L., Zhang, L. and Wang, C. (2015). Fuel Efficiency Modeling and Prediction for Automotive Vehicles: A Data-Driven Approach. *2015 IEEE International Conference on Systems, Man, and Cybernetics*. [online] doi:https://doi.org/10.1109/smc.2015.442.