Automobile CO2 Emissions Prediction

Noah Shimizu
Sai Bhargav Tetali
Sreekar Lanka
Vishwak Venkatesh

Problem Statement:

We chose to predict the Carbon dioxide (CO2) emissions of automobiles and understand which factors affect emissions. In the dataset, we have specifications like number of cylinders, engine size, fuel type, transmission type, etc. for ~26k models of automobiles sold in Canada from 1995 - 2022.

Why we chose this problem:

CO2 emissions are a major problem as they are greenhouse gasses that drive global warming. 27% of CO2 emissions in the US is caused by transportation with passenger cars being a major contributor. The EPA enforces strict regulations through standardized lab tests to ensure standards are met. However, with recent emissions scandals (Eg - Volkswagen cheating emission tests), we wanted to explore the possibility of measuring emissions independent of lab tests. Do other vehicle features help in predicting emissions with some accuracy? Can we uncover any interesting patterns and trends? To answer these questions, we selected this problem.

Exploratory Analysis:

This dataset contains the CO2 emissions in g/km of some car models released from 1995 to 2022. We have vehicle specifications like number of cylinders, fuel consumption rates in city/highway, car type, engine size etc.

We started by looking at the distribution of the CO2 emissions (fig. 1) where we see that the median CO2 emissions was 269 g/km, 25^{th} percentile 230 g/km and the lowest 94 g/km. For reference, eco-friendly cars have less than 100g/km emissions. Hence, we can assume that most of the car models in the dataset are not eco-friendly.

To better understand the relationships between variables, we have plotted boxplots & scatterplots for number of cylinders, engine size, fuel type, car brand, etc. with the CO2 emissions as the value of interest (figs. 2-5). We observed that there is a clear positive relationship between the number of cylinders and CO2 emissions. The same has been observed with engine size. Also, we observed that the luxury brands like Bugatti, Ferrari and Bentley have

higher CO2 emissions whereas affordable brands like Honda, Toyota, Hyundai have lower emissions.

Fuel consumption also has a positive relationship with CO2 emissions which makes sense (figs. 6-7). The more fuel a car consumes, the more fuel is burnt leading to higher CO2 emissions. A time series analysis with the median and mean values of CO2 emissions for each year was also done (fig. 8). We can see that starting from around 2005 the car models have reduced CO2 emissions. This can be attributed to the more stringent policies regarding emission control implemented in the past 15 years.

We have also created a correlation table between the numeric columns (fig. 9). As previously observed, we see a highly positive correlation between engine size, cylinders & fuel consumption with CO2 emissions. We also see strong positive correlations between cylinders, engine size & fuel consumption which implies that our features have multicollinearity.

Solutions & Insights:

To better distill the information present in the dataset, we decided to engineer some of the categorical features. For example, Vehicle Class had values such as SUV - Standard, SUV - Small, etc. which might not add much information while introducing noise to our models. Hence, we aggregated these granular vehicle classes together to create a new feature Vehicle Class 1. A similar approach was taken to aggregate different transmissions into a column called Transmission Type.

Using a combination of the numeric & categorical features (Table 1), we set out to predict CO2 emissions through these four models:

- Linear Regression
- k-Nearest Neighbors
- Random Forest
- Boosting

Linear Regression & PLS Regression:
We attempted various linear regressions in our model. Details regarding each of these models it outlined in Table 2. We see that Ordinary Least Squares performed poorly using only numerical data, and improved greatly from using categorical variables.

As discussed however in EDA, our predictors exhibit large amounts of multicollinearity. We thus choose the dimension reduction method of PLS Regression. We measure RMSE of the model in a 5-fold CV on the number of

PLS components used in modeling, the results of which can be seen in fig 10. We see it performs just as well as OLS, yet we consider it a better model, as we believe it to be theoretically more robust.

Interpreting this model, we refer to table 3. Table 3 lists two sets of coefficients. The first is the coefficients of the variables, assuming each variable is scaled to have standard deviation of 1. The second is the coefficients for the unscaled variables. We rank variables based upon their coefficients of scaled predictors, as it is standardized for the spread of our x variables. Looking at our scaled predictions, we see that fuel consumption tends to be greatly influential, with greater fuel consumption increasing CO2 emissions. Also greatly influential we see the fuel types being greatly influential, with again ceteris paribus ethanol based cars having much less CO2 emissions, and diesel cars having much more. Interpreting this model is again difficult, as model features are structurally correlated.

K - nearest neighbors:
In the K-nearest neighbors model we have split the data into training and test sets in the ratio of 7:3 respectively. We have 5 numerical columns after scaling them and 3 categorical columns as features. For categorical columns dummy variables are created. The numerical columns have been scaled to the range 0 to 1.

We went through several iterations of feature selection to arrive at the model which gave the least test RMSE (figs. 12-14). The details for the different models fitted are given in Table 3.

Random Forest:
Random forest performs better than both linear regression and KNN, but here as well we see a similar trend like in KNN and PLS - when we consider all features we see better Test RMSE than when we consider only numeric features. This reinstates the fact that categorical variables in the data are significant.

When we optimize the random forest model using the number of trees and depth, we see even better results (figs. 15-20). From the feature importance plot (fig. 21) we see that the most important features are fuel consumption and the fuel type. We can say that for a car manufacturer, it would make sense to keep the mileage of the vehicle in check to reduce the CO2 emissions.

Boosting:
We fit the data using two boosting techniques - Gradient Boosting & XGBoost. Our motivation behind using boosting was to try and outperform our Random

Forest model through an ensemble of weak learners. The results for both models are presented in Table 5.

We see that both models perform poorly when only numeric features are used. This indicates that although most numeric features are highly correlated with CO2 emissions, it isn't enough to explain all the variability in the data. Hence, we can deduce that there is still some useful information to extract from the categorical variables.

Using all categorical features resulted in reductions in RMSE, however, the variable count had increased twenty-fold. This raised the possibility that this large number of features may be causing more noise than adding to the signal. Hence, to reduce any noise that may have crept in, we removed a few categorical variables (Eg - Make). As we can see from Table 2, this improved model performance.

Subsequent model tuning yielded further improvements (figs. 22-25). Finally, the best boosting model we obtained was an XGBoost Model which was trained on all features except Make, Transmission & Vehicle Class 1. Its optimized # of estimators was 500 with maximum depth set as 4. This yielded a very low test RMSE of 1.29.

This model considered Fuel Consumption along with Fuel Type 'E' (Ethanol) as the most important predictors (fig 26). The former makes sense as we have already seen the strong correlation between fuel consumption and CO2 emissions. We have seen earlier that ethanol fueled cars have lesser CO2 emissions on average compared to diesel cars. We can see that the other fuel types, number of cylinders & engine size are also useful in improving predictions. Surprisingly, the model does not consider the transmission type as an important predictor.

Conclusion:

We picked this problem in order to identify if there was a way to predict a ballpark estimate of CO2 emissions using car features. Through the different models that we have fitted on the data, we have been able to develop an XGBoost model that estimates CO2 emissions quite accurately. This model could then be used to predict emissions of a new car model released in the future.

Some ways that we could improve our model in the future would be to get more vehicle data. We could also better aggregate the Make feature thereby adding some information to the model.

APPENDIX

Please find figures, plots and tables here

Sources

Dataset - https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64
Description of dataset - https://www.nrcan.gc.ca/sites/nrcan/files/oee/pdf/transportation/fuel-efficient-technologies/2022%20Fuel%20Consumption%20Guide.pdf

Tables

Table 1 - List of Numerical columns and Categorical columns in the dataset

Numerical Columns	Categorical Columns
Model Year, Engine Size(L), Cylinders, Fuel Consumption (City (L/100 km), Fuel Consumption(Hwy (L/100 km))	Make, Fuel Type, Vehicle Class 1, Transmission Type

Table 2 - Linear & PLS Regression model results

Model	Feature Count	Test RMSE
OLS Regression Num Only	5	21.26
OLS Regression Num + Fuel Type	10	5.33
OLS All Predictors but Name & Make	27	5.28
15 Dim. PLS All Predictors but Name & Make	27	5.28

Table 3 - k-Nearest Neighbors model results

Features used	Feature Count	Test RMSE	Reference plot

Numerical Only (5)	5	8.19	Fig 12
All but Car Name & Make (8)	25	6.84	Fig 13
All but Car Name, Make & Vehicle Class 1 (7)	14	4.99	Fig 14

Table 4 - Random Forest model results

Features used in model (#)	**Test RMSE**
Only Num (5)	6.77
All features (108)	3.25
Only Num with tuning	6.82
All features and tuning	2.79
All Features excl. Make, Transmission & Vehicle Class Grouped (13)	3.17
All Features excl. Make, Transmission & Vehicle Class Grouped and tuning (13)	2.73

Table 5 - Boosting model results

Features used in model (#)	**Gradient Boosting - Test RMSE**	**XGBoost - Test RMSE**
Only Num (5)	14.66	6.5
All Features (108)	5.75	1.83
All Features with Tuning	1.71	1.59
All Features excl. Make & Transmission (25)	5.75	1.66
All Features excl. Make & Transmission with Tuning	1.37	1.54
All Features excl. Make, Transmission & Vehicle Class Grouped (13)	5.77	1.62
All Features excl. Make, Transmission & Vehicle Class Grouped with Tuning	1.38	1.29

Figures

Fig 1 - Histogram of CO2 emissions

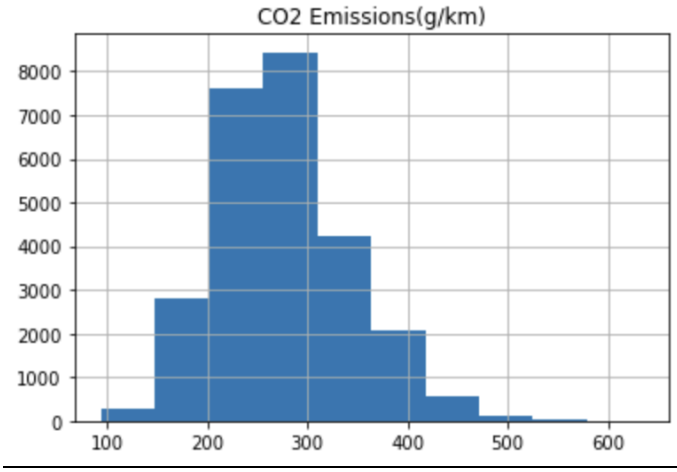

Fig 2 - Boxplot of # of cylinders vs CO2 emissions

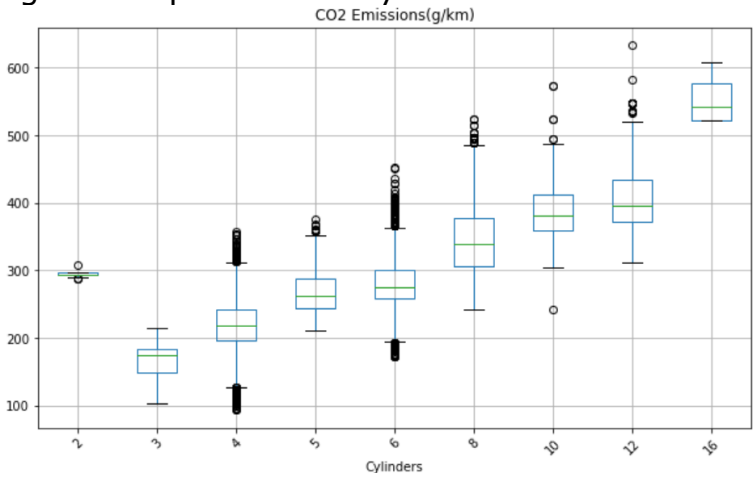

Fig 3 - Scatterplot of engine size vs CO2 emissions

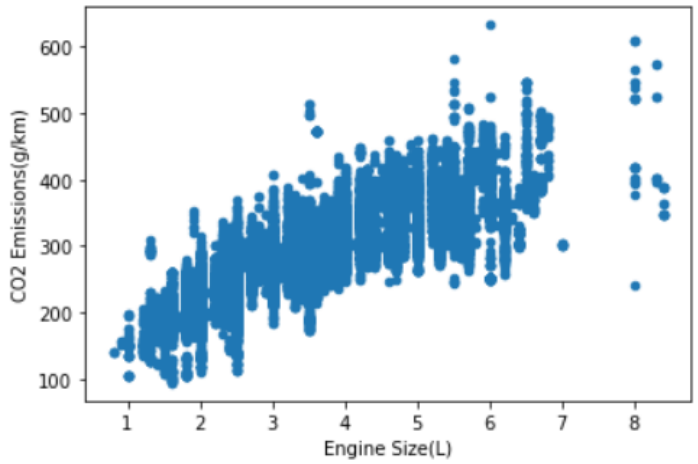

Fig 4 - Boxplot of fuel type vs CO2 emissions

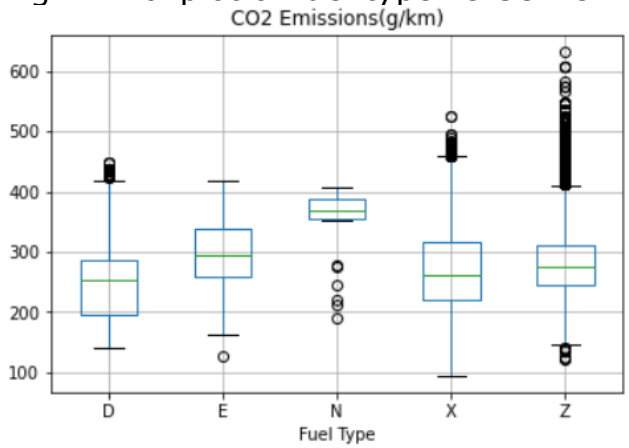

Fig 5 - Boxplot of car brand vs CO2 emissions

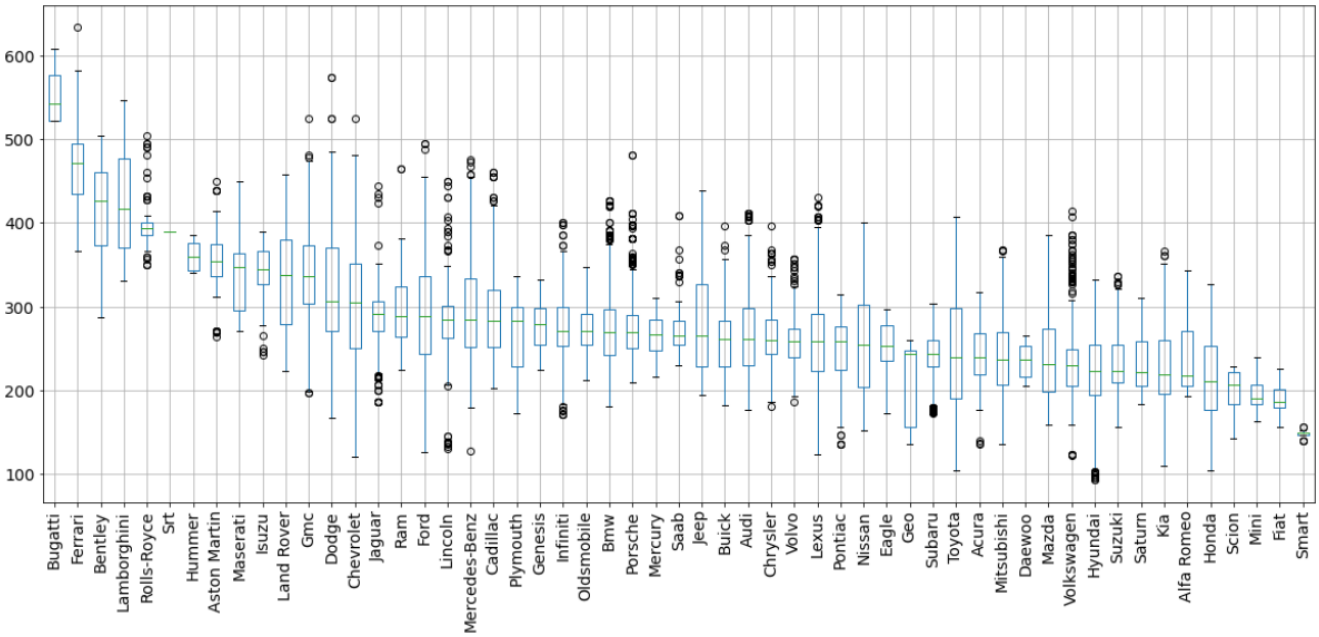

Fig 6 - Scatterplot of fuel consumption (City) vs CO2 Emissions

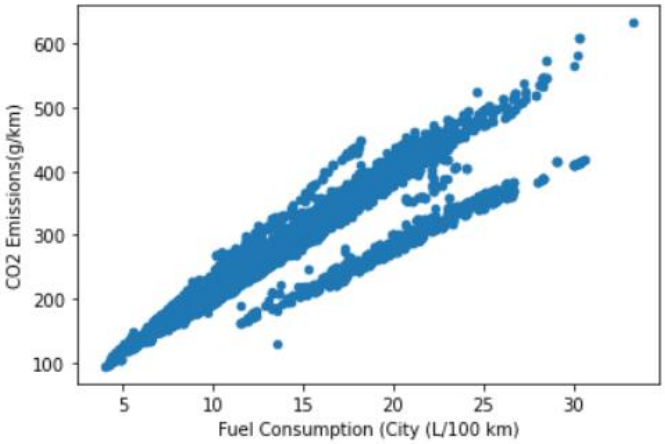

Fig 7 - Scatterplot of fuel consumption (Highway) vs CO2 Emissions

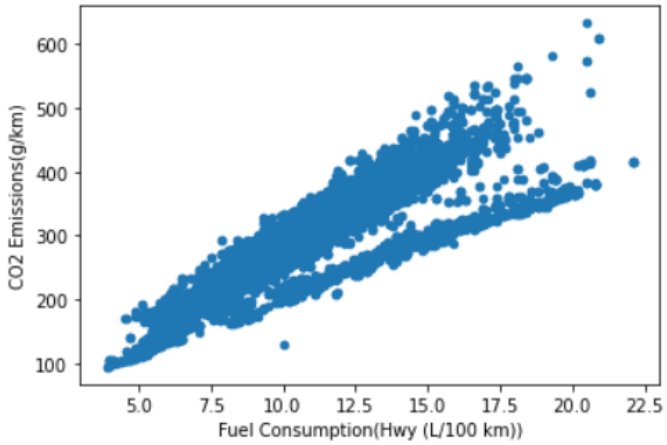

Fig 8 - Time series trend of mean & median CO2 emissions over the years

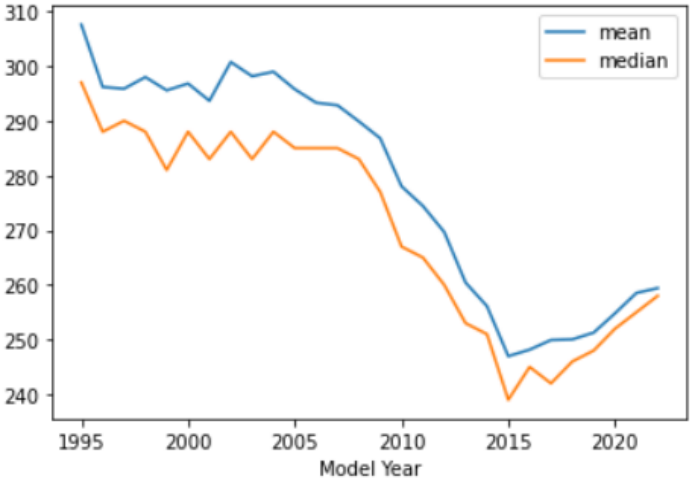

Fig 9 - Correlation matrix between the numeric predictors

	Model Year	Engine Size(L)	Cylinders	Fuel Consumption (City (L/100 km)	Fuel Consumption (Hwy (L/100 km))	CO2 Emissions(g/km)
Model Year	1	-0.05	-0.04	-0.25	-0.25	-0.28
Engine Size(L)	-0.05	1	0.91	0.82	0.76	0.83
Cylinders	-0.04	0.91	1	0.79	0.7	0.79
Fuel Consumption (City (L/100 km)	-0.25	0.82	0.79	1	0.95	0.93
Fuel Consumption (Hwy (L/100 km))	-0.25	0.76	0.7	0.95	1	0.91
CO2 Emissions(g/km)	-0.28	0.83	0.79	0.93	0.91	1

Fig 10 - Errors of PLS vs number of components

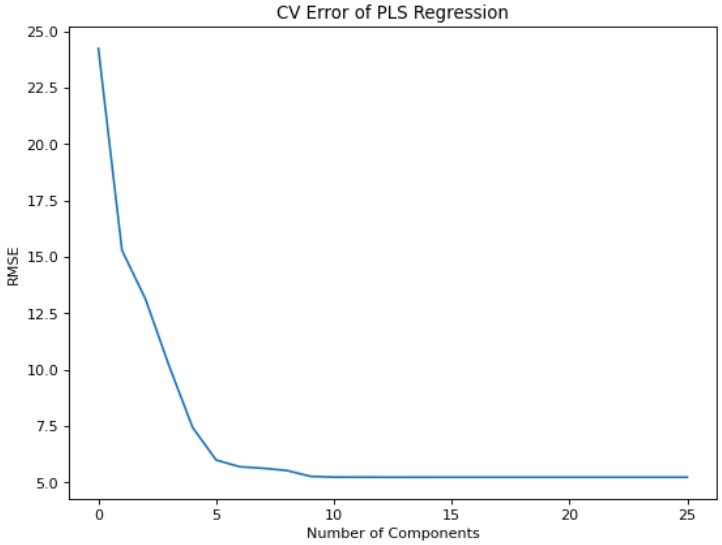

Fig 11 - Coefficients of Final Linear Model

	Fuel Consumption (City (L/100 km)	Fuel Consumption(Hwy (L/100 km))	E	D	Z	X	N	Model Year	Pickup Truck	Engine Size(L)
Coefficients for Unscaled Predictors	12.723875	9.334802	-115.956215	41.657679	8.295716	8.025978	-63.447181	0.122888	1.754425	0.428784
Coefficients for Scaled Predictors	47.300347	24.038138	-23.007961	5.437825	4.032430	3.986521	-2.319931	0.948620	0.584539	0.570606

Fig 12 - Cross-validation plot between mean Test RMSE and number of nearest neighbors with only numeric columns

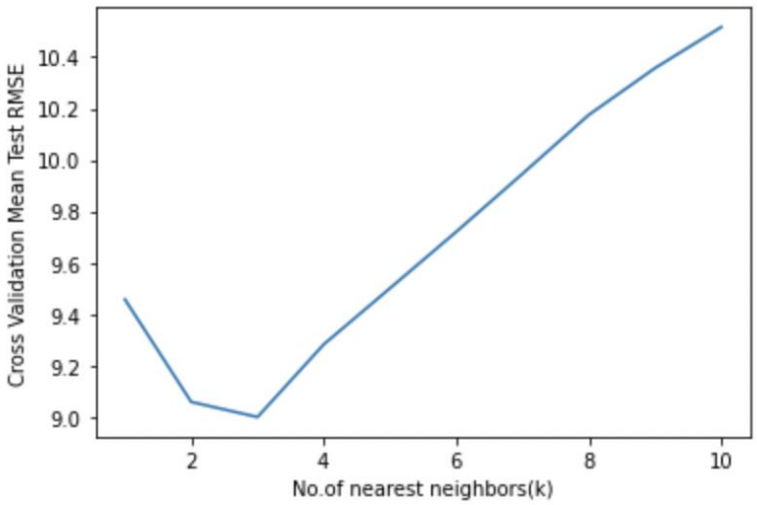

Fig 13 - Cross-validation plot between mean Test RMSE and number of nearest neighbors with numeric columns and 3 Categorical Columns

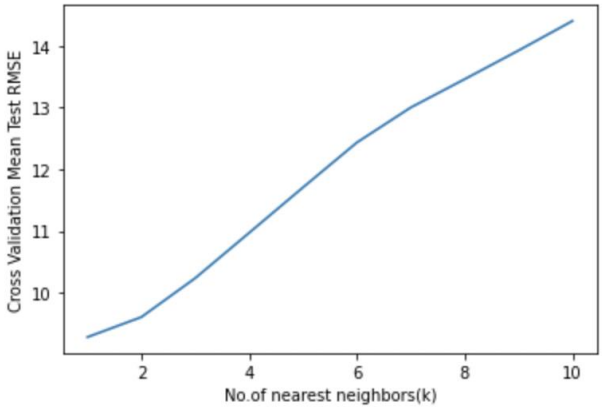

Fig 14 - Cross-validation plot between mean Test RMSE and number of nearest neighbors with 'Vehicle Class 1' excluded

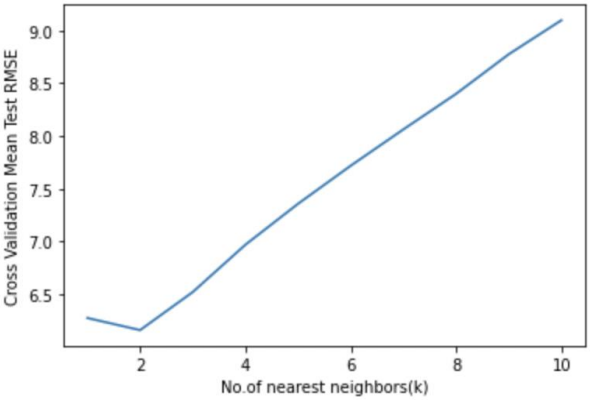

Fig 15 - Random forest model with all Variables : Optimizing for number of trees. We see that the minimum RMSE is at 150 trees.

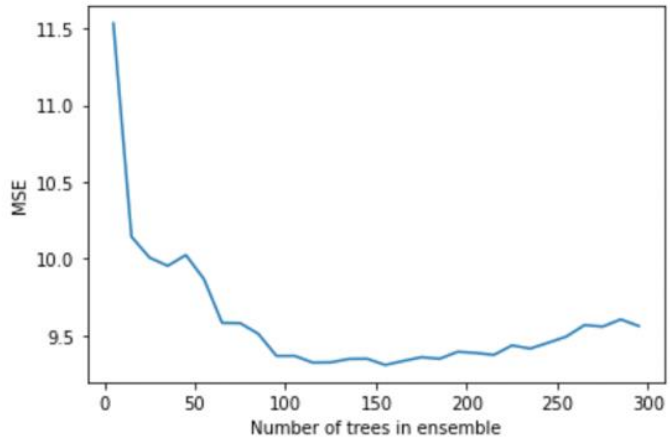

Fig 16 - Random forest model with all Variables : Optimising for depth of trees. We see that minimum RMSE is at 19

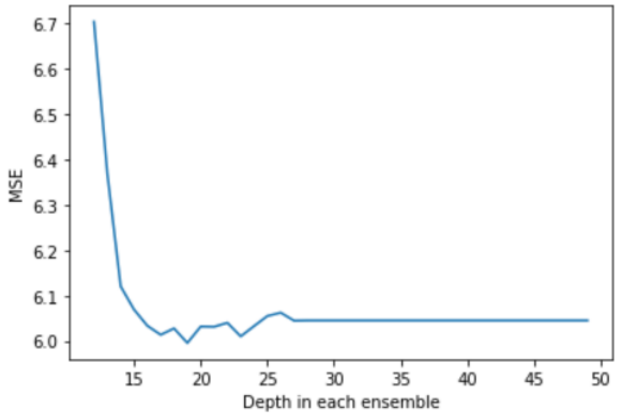

Fig 17 - Random forest model with only Numeric features : Optimising for no. of trees. We see that minimum RMSE is at 150

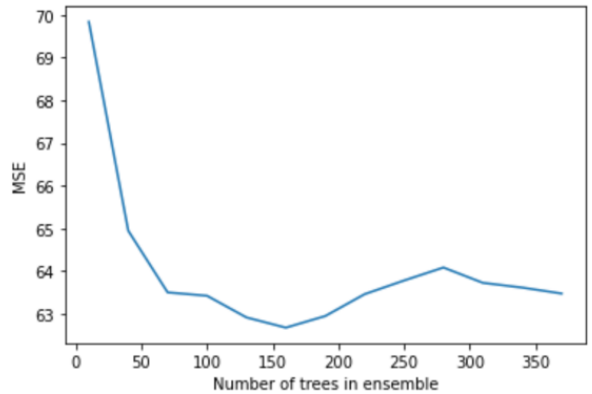

Fig 18 - Random forest model with only Numeric features : Optimising for depth of trees. We see that minimum RMSE is at 19

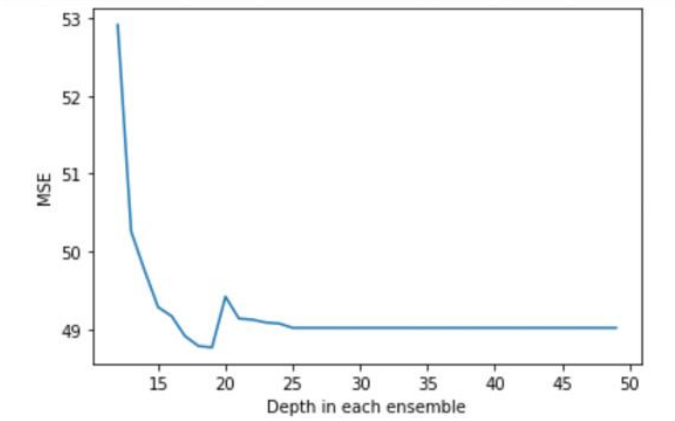

Fig 19 - Random forest model with all features except Make, Transmission type and grouped Vehicle class: Optimizing for no. of trees. We see that minimum RMSE is at 120 trees

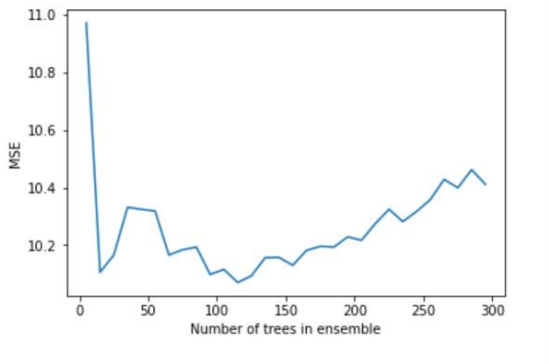

Fig 20 - Random forest model with all features except Make, Transmission type and grouped Vehicle class : Optimizing for depth of trees. We see that minimum RMSE is at 17

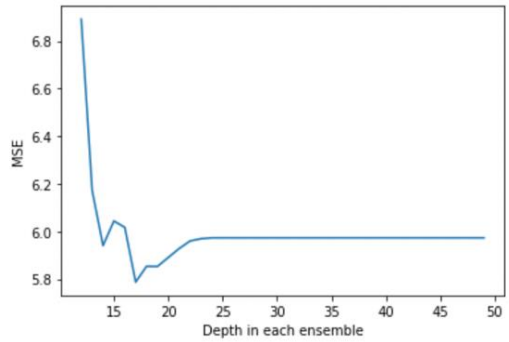

Fig 21 - Random Forest Best Model Feature Importance

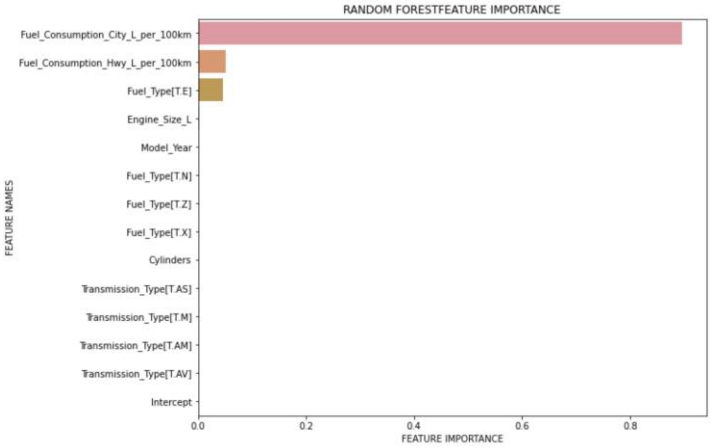

Fig 22 - Optimizing Gradient Boosting on # of trees. Selected value was 300.

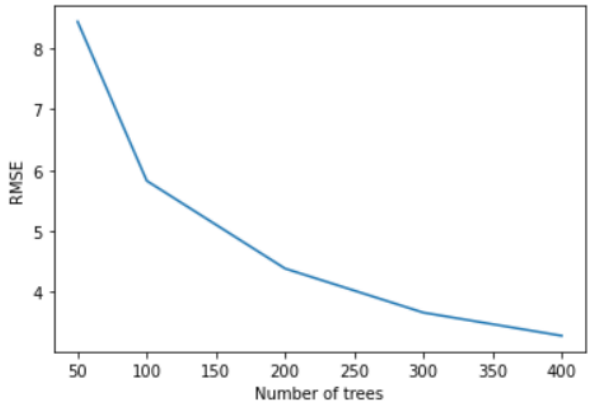

Fig 23 - Optimizing Gradient Boosting on max depth. Selected value was 7.

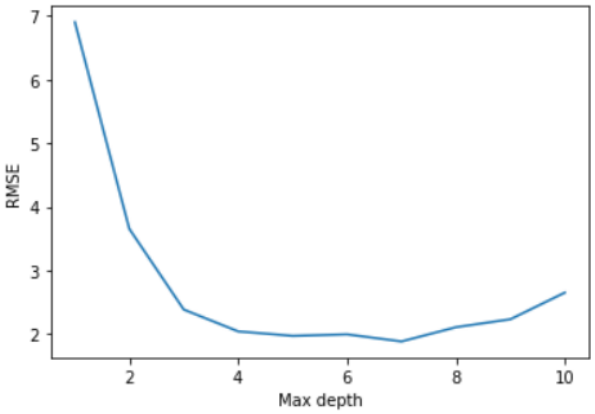

Fig 24 - Optimizing XGBoost on # of trees. Selected value was 500.

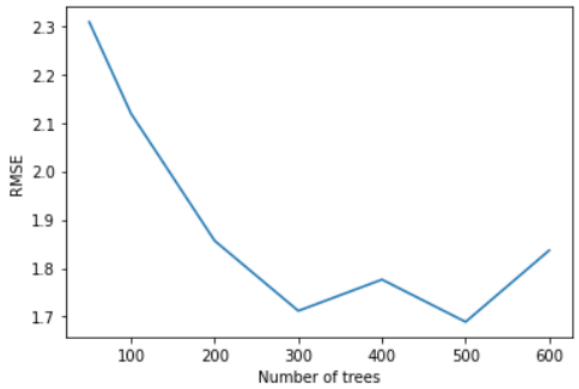

Fig 25 - Optimizing XGBoost on max depth. Selected value was 4.

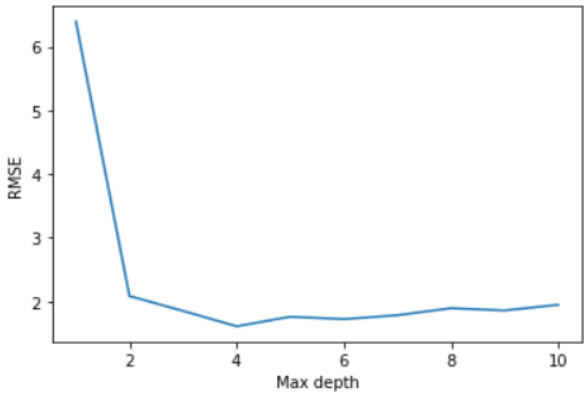

Fig 26 - XGBoost Best Model Variable Importance