Linear Regression Model to Predict Car Emissions

1. **Introduction:**

This project aims to provide a linear model that will help predict the emissions a car may produce based on different features (attributes of the car). This project also intends to try different types of combinations within the dataset which include performing a Standard Scaling to the dataset, performing second-degree polynomial transformations to the dataset and no manipulation whatsoever.

Additionally, various data frames will be created utilizing KBest, Variance Threshold and Correlation criteria to select the best possible features for the model and the combinations previously stated will be applied to each of this data frames so that finally the best model may be selected for the prediction of emission values.

1. **Dataset Analysis:**

The data set used in this analysis is called ‘CO2\_emissions’ and its a .csv comprised of information concerning attributes of a car including its emissions. The head of the table can be visualized in the following figure:

Table

Description automatically generated

Figure 1. Data frame .head(5) for ‘CO2\_emissions.csv’

The data in the table had to be cleaned so that a proper linear regression model could be obtained. For this, the function .info() was applied to visualize the column information in the following table and given the result, various data manipulation techniques were used:

RangeIndex: 8197 entries, 0 to 8196

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Make 7858 non-null object

1 Model 7845 non-null object

2 Vehicle Class 7855 non-null object

3 Engine(L) - Cylinder 7854 non-null object

4 Transmission 7873 non-null object

5 Fuel Type 7850 non-null object

6 Fuel Consumption City (L/100 km) 7839 non-null float64

7 Fuel Consumption Hwy (L/100 km) 7854 non-null float64

8 Fuel Consumption Comb (L/100 km) 6246 non-null float64

9 Fuel Consumption Comb (mpg) 4678 non-null float64

10 CO2 Emissions(g/km) 7872 non-null float64

dtypes: float64(5), object(6)

The measures implemented include:

* Changing column names to more accessible names
* Checking for nulls and removing them
* Filling the ‘fuel\_comb’ column with data from ‘fuel\_mpg’ given that it was the same data in different units
* Applying lower case to every value of columns which had object type values and replacing whitespaces to underscores
* Reducing unique value types of columns: ‘type’, ‘make’, ‘transmission’ and ‘fuel’ columns

Finally, after extensive cleaning the final data set that was obtained (figure 2) is presented in the following figure which is comprised of 6444 rows and 11 columns instead of 8197 rows and 11 columns as previously stated.

Graphical user interface, table

Description automatically generated

Figure 2. Final data frame ‘Lab01\_ERu63902\_cleaned.csv’

1. **EDA**

The correlation between features is shown in the following heat map plot (figure 3). ‘fuel\_city’, ‘fuel\_hwy’ and ‘fuel\_comb’ seem to have high correlation and there seems to be collinearity between the attributes. This is confirmed with a Variance Inflation Factor (VIF) on the standard scaled dataset as shown in the following table and the decision was made that since the attributes refer to the same physical thing only on fuel feature would be kept which is in this case the one with the lowest VIF ‘fuel\_hwy’.

A picture containing table

Description automatically generated

fuel\_city 493.966581

fuel\_hwy 148.316080

fuel\_comb 1108.034622

emissions 7.545789

engine(l) 9.275746

cylinder 8.096618

dtype: float64

Figure 3. VIF scores and correlation heat map.

The following plots (figure 4) will show the distribution of some features in the dataset. Aside from Cylinders, the features shown follow in general a normal distribution. Cylinders data includes mostly 3 main values: 4, 6 and 8 cylinders; which appear to be the most popular.

Chart, histogram

Description automatically generated

Figure 4. EDA plots to show correlation among features in dataset

1. **Feature Observation and Hypothesis**

The following plots (figure 5) give us a little more insight into the relationship of emissions and ‘cylinder’ and ‘fuel\_hwy’ features.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 5. Scatter plots for ‘Emissions’ vs ‘Highway Fuel’ and ‘Cylinder’

Scatter plots appear to follow a positive direct relationship; however, ‘cylinder’ data seems to be centered around 3 main values of cylinders as previously stated. Overall, this data makes for the hypothesis that there is a very high chance that the greater the ‘cylinder’ and ‘fuel\_hwy’ datapoint is the more emissions the car will produce for the most part.

There is also a possibility that since data for ‘cylinder’ is clustered around 3 main values: 4, 6 and 8 the model will be better at predicting the ‘emissons’ for those popular cylinders values which might make the model unstable when outliers are presented.

1. **Simple Linear Regression report**

The feature selection method employed in the datasets were:

1. Correlation based selection – This method calculated the correlation between features and the basis is that features with high correlation which were selected were the most representative in the linear model. In my case I performed a manual selection since I don't want a car maker to influence the model. The final features selected were the following:

Table

Description automatically generated

1. Variance Threshold – This method consisted of calculating the variance of each feature and putting a threshold for selection on those variances. The threshold selected was 0.125 since it would filter out most of the features with little variance and would still leave out brands from the selected features.

Table

Description automatically generated

1. Select K-Best – Consists of an in-house model provided by scikit to select the k best features in a data frame according to their k-score. In this case, the score function utilized was f\_regression since according to scikit-learn this is the method to obtain F-value between label/feature for regression tasks. The number of k features in this case was set to 10 minus the feature ‘emissions’ since that is the feature we want to predict.

Table

Description automatically generated with low confidence

1. **Linear Regression with Lasso Model**

The top results for my alpha scores are shown in the following table:

Table

Description automatically generated

These calculations utilized all available features in the model and followed the same ratio for training and testing as all the other models. The values were selected based on the alpha that gives the lowest RMSE score and so an alpha of 0.00100 was finally found. This alpha provided the best results out of all the linear models and thus this model was chosen as the Linear Regression Model to make the final predictions.

1. **Analysis**

The results obtained are shown in the following figure (figure 6) and include the summary table of possible feature selection combinations and the scatter plot comparing the predicted value vs. the actual value for emissions:

Table

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 6. Summary Analysis and Predicted Price vs Actual Price scatter plot.

Findings on the linear model:

* I believe the best linear model which is the Linear Regression Model with Lasso since it has a high accuracy and score for RMSE and R2 values because the features used in this dataset follow various iterations that find the best alpha. Additionally, this method utilizes a generous number of features and I believe that plays a high part into the prediction since it has more information to make a model
* A disadvantage is that there are more coefficients needed to use this model and since the other models use less coefficients and have a combination which also produces a high accuracy model then it seems that perhaps in real life the less features used with the most accuracy would be the best answer for the linear regression model.
* The linear regression scatter plot seems to have a clear positive linear relationship and the values predicted are very similar or on par with the test values inputted into the model. For the next predictions I believe there needs to be more fine tuning for outliers, there are still some outliers in the ‘fuel\_hwy’ feature which might skew the data slightly to the left.
* It would be very interesting if one could also loop through how many features to use and fine tune it for the best possible number of features per feature selection method, in this case the manual selection would have to be relinquished. Finally, I think it’s also a good idea to experiment with the train\_test\_split parameters such as the test\_size to see how much the scores vary, experiment more with the alpha values and with the number of iterations it will be tested