DSCI431-Project 2 Final Report

Name: Botsung Huang

Abstract

This report mainly presents my results in detail and some interesting findings. Including the Process of Data Analysis and shows the description of my findings for the Linear Regression model I used.

Determine the significance of each feature

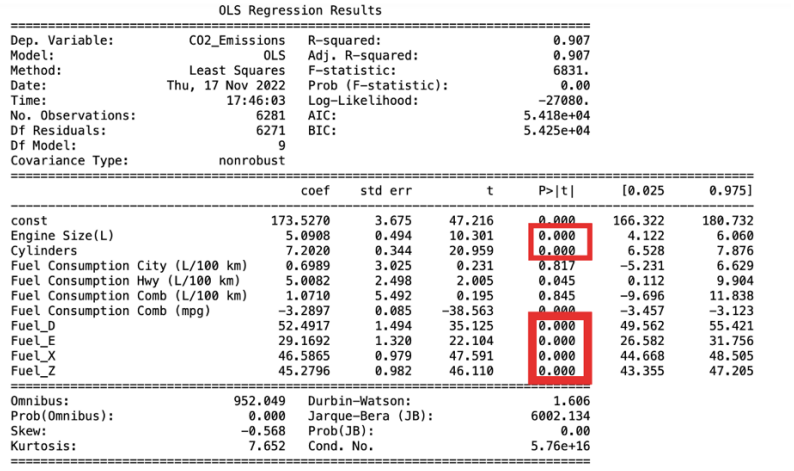
The first result I want to mention below is Ordinary Least- Square (OLS) Regression Results to show which features are significant when building the Linear Regression model.

As you can see below the codes I use, the main purpose of this is to define the target feature we want to use to predict the model and show performance. The target I chose here is "CO2_Emissions" which we are most concerned about.

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
x = df[df.columns[df.columns != 'CO2_Emissions']]
y = df.CO2_Emissions

# Statsmodels.OLS requires us to add a constant.
x = sm.add_constant(x)
model = sm.OLS(y,x)
results = model.fit()
print(results.summary())
```

The graph below is the OLR result. As you can see there are several metrics we can use to build the linear regression model. What I'm focusing on is P-value. If the feature got a low p-value which is lower than 0.05, it means the feature is significant to the model building. From the result below we can determine Engine size, Cylinders, and Fuel Type are significant features.



OLS Regression Results

Dep. Variable:	CO2_Emissions	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	6831.
Date:	Thu, 17 Nov 2022	Prob (F-statistic):	0.00
Time:	17:46:03	Log-Likelihood:	-27080.
No. Observations:	6281	AIC:	5.418e+04
Df Residuals:	6271	BIC:	5.425e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	173.5270	3.675	47.216	0.000	166.322	180.732
Engine Size(L)	5.0908	0.494	10.301	0.000	4.122	6.060
Cylinders	7.2020	0.344	20.959	0.000	6.528	7.876
Fuel Consumption City (L/100 km)	0.6989	3.025	0.231	0.817	-5.231	6.629
Fuel Consumption Hwy (L/100 km)	5.0082	2.498	2.005	0.045	0.112	9.904
Fuel Consumption Comb (L/100 km)	1.0710	5.492	0.195	0.845	-9.696	11.838
Fuel Consumption Comb (mpg)	-3.2897	0.085	-38.563	0.000	-3.457	-3.123
Fuel_D	52.4917	1.494	35.125	0.000	49.562	55.421
Fuel_E	29.1692	1.320	22.104	0.000	26.582	31.756
Fuel_X	46.5865	0.979	47.591	0.000	44.668	48.505
Fuel_Z	45.2796	0.982	46.110	0.000	43.355	47.205

Omnibus:	952.049	Durbin-Watson:	1.606
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6002.134
Skew:	-0.568	Prob(JB):	0.00
Kurtosis:	7.652	Cond. No.	5.76e+16

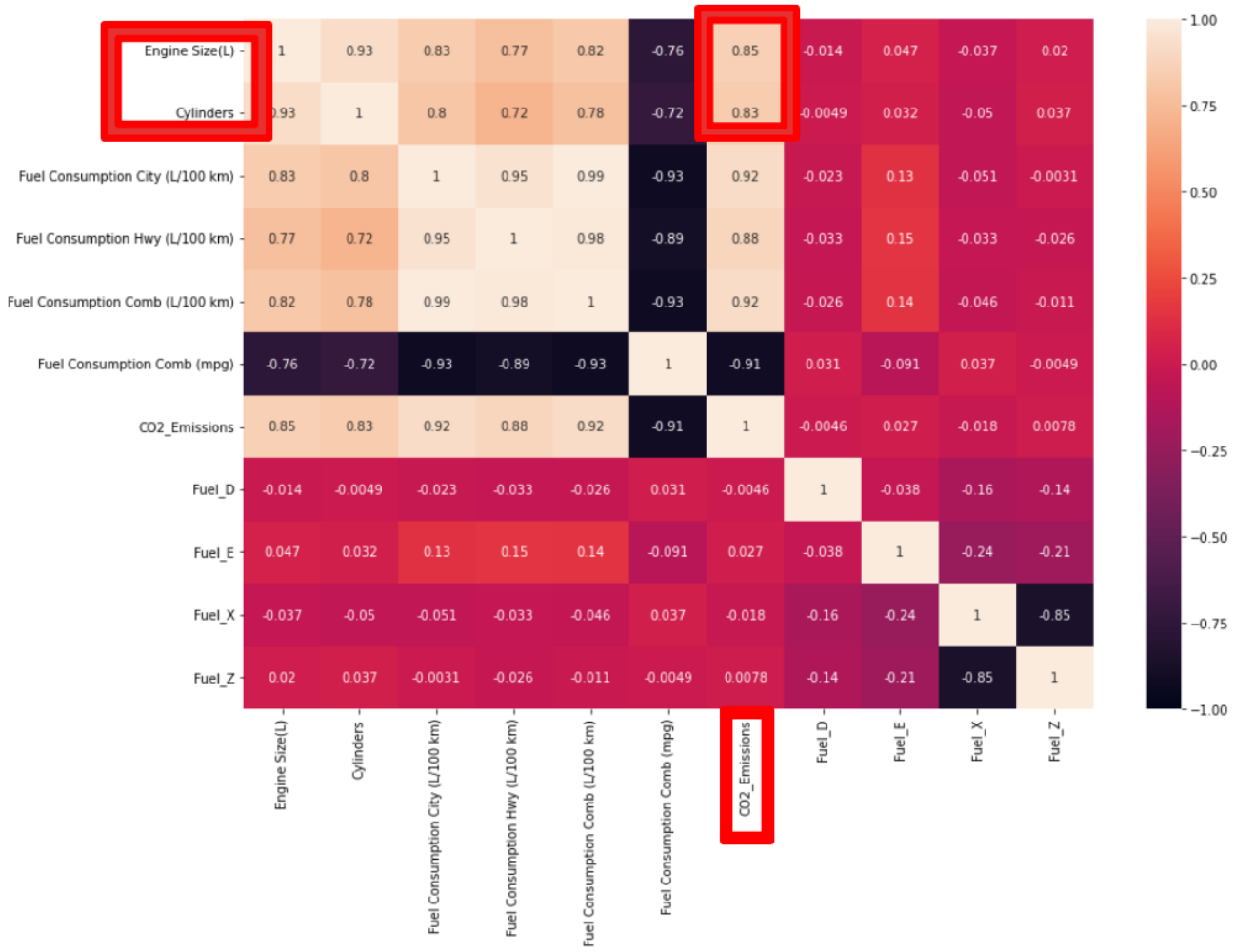
Correlation Analysis (2 variables)

The next step is visualizing the correlation analysis into a heatmap. In this part, we can check the correlation between 2 variables for each.

```
# Analysing the correlation between all variables

plt.figure(figsize = (15,10))
sns.heatmap(df.corr(), vmin = -1, vmax = 1, annot = True);
```

The result of the correlation analysis heatmap shows below. As you can see Engine size and Cylinders have a strong correlation with CO2 emissions, with the high separate value of 0.85 and 0.83. This result helps us to understand which features cause more CO2 emissions.



Analyze the correlation between all variables

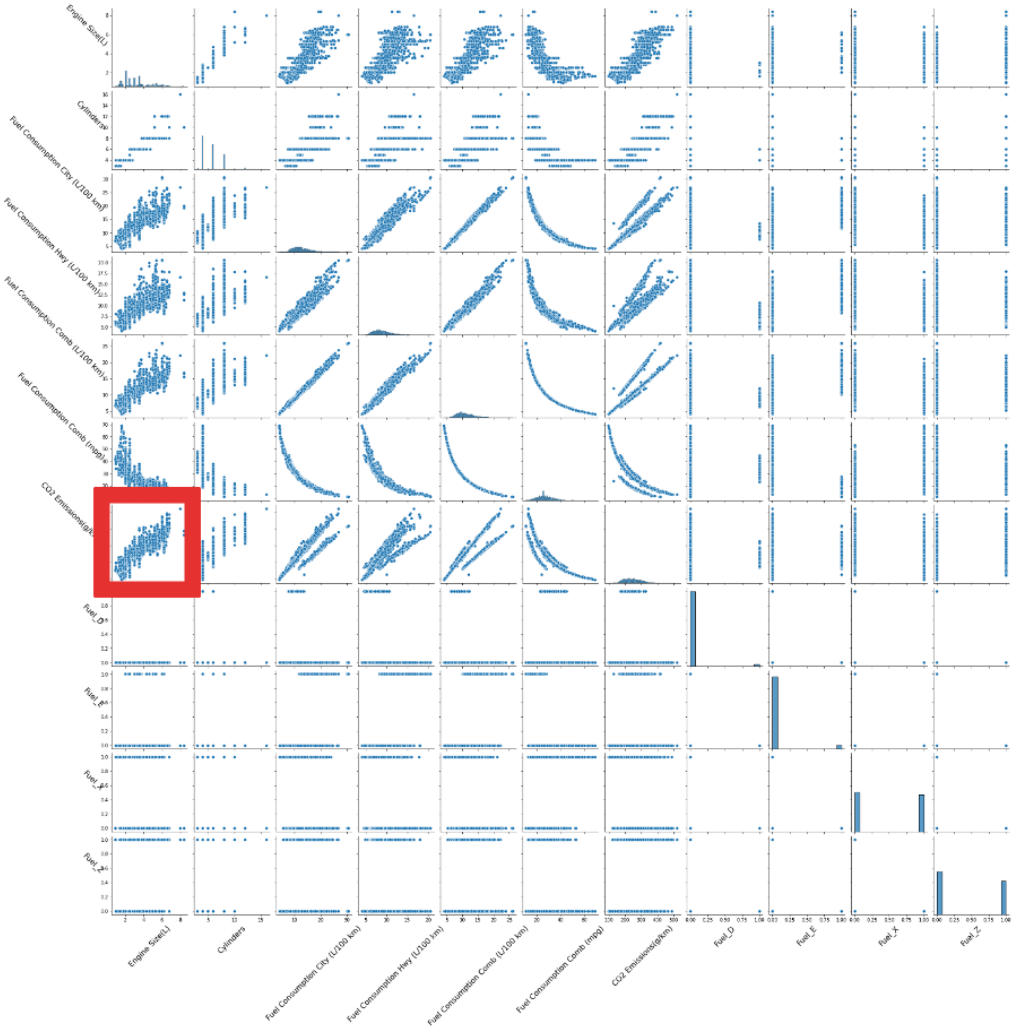
This part shows the result of the correlation between all variables rather than the previous just showing 2 variables. The code for this analysis shows below.

```
# Analysing the correlation between all variables – pairplot

all_pairs = sns.pairplot(df)

for ax in all_pairs.axes.flatten():
    ax.set_xlabel(ax.get_xlabel(), rotation = 45, fontsize = 16)
    ax.set_ylabel(ax.get_ylabel(), rotation = -45, fontsize = 16)
    ax.yaxis.get_label().set_horizontalalignment('right')
```

As the result shown below, we can find that Engine size has a positive correlation with CO2 emissions. Although fuel consumption also has a positive correlation with C02 emissions, it is a fact that we already know that the more fuel we consume, the more CO2 generate. So here I want to do further analysis for engine size.

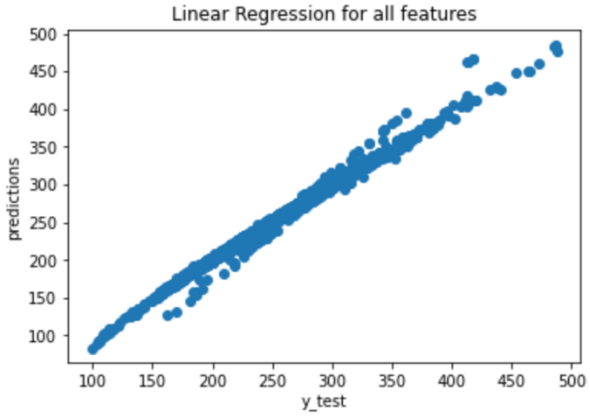


Linear Regression Model with all features & significant features

This first result shows the linear regression model which I consider all features to build it. The R2 score is high, with a value of 0.991. This means the model performed well. This result also shows every coefficient for each feature and shows MAE, MSE, and RMSE to evaluate the model performance. The second graph below shows only considered significant features. Although all features have better performance, both models perform well with high R2 scores.

	Coefficient
Engine Size(L)	0.288766
Cylinders	1.025957
Fuel Consumption City (L/100 km)	7.538225
Fuel Consumption Hwy (L/100 km)	6.090936
Fuel Consumption Comb (L/100 km)	6.356243
Fuel Consumption Comb (mpg)	-0.867125
Fuel_D	49.961919
Fuel_E	-88.698214
Fuel_X	19.722363
Fuel_Z	19.013931

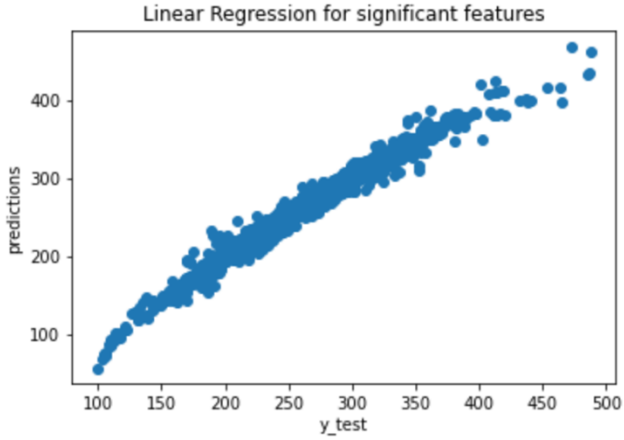
```
R2 score is: 0.9914341432555234
```




```
Mean Absolute Error: 3.2110675416674686
Mean Squared Error: 30.284374517116497
Root Mean Square Error: 5.503124068846395
```

	Coefficient
Engine Size(L)	3.217308
Cylinders	4.322204
Fuel Consumption Hwy (L/100 km)	15.725581
Fuel Consumption Comb (mpg)	-2.750741
Fuel_D	44.462616
Fuel_E	-74.015912
Fuel_X	14.504442
Fuel_Z	15.048854

```
R2 score is: 0.9762418427573324
```

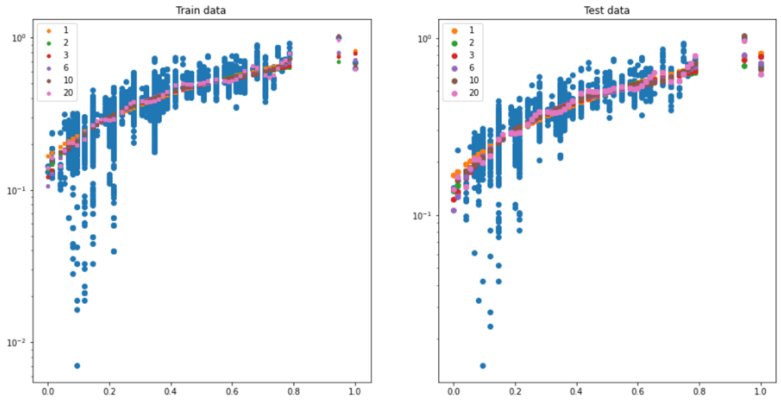



```
Mean Absolute Error: 6.513684266096663
Mean Squared Error: 83.99637692253394
Root Mean Square Error: 9.164953732700123
```

Polynomial Regression

In this part, I generate a polynomial regression model to pursue better model performance between Engine size and CO2 emissions. I split data into 70% for training data and the rest 30%

for testing data. The polynomial Regression degree is 1,2,3,6,10, and 20. As you can see from the graph below, the more degree I use, the better performance I got. The best performance is a 20-degree polynomial degree with a train score of 0.75 and the test score of 0.78. From the graph, you can also find that the more degree I use, the better fitting condition to the model.

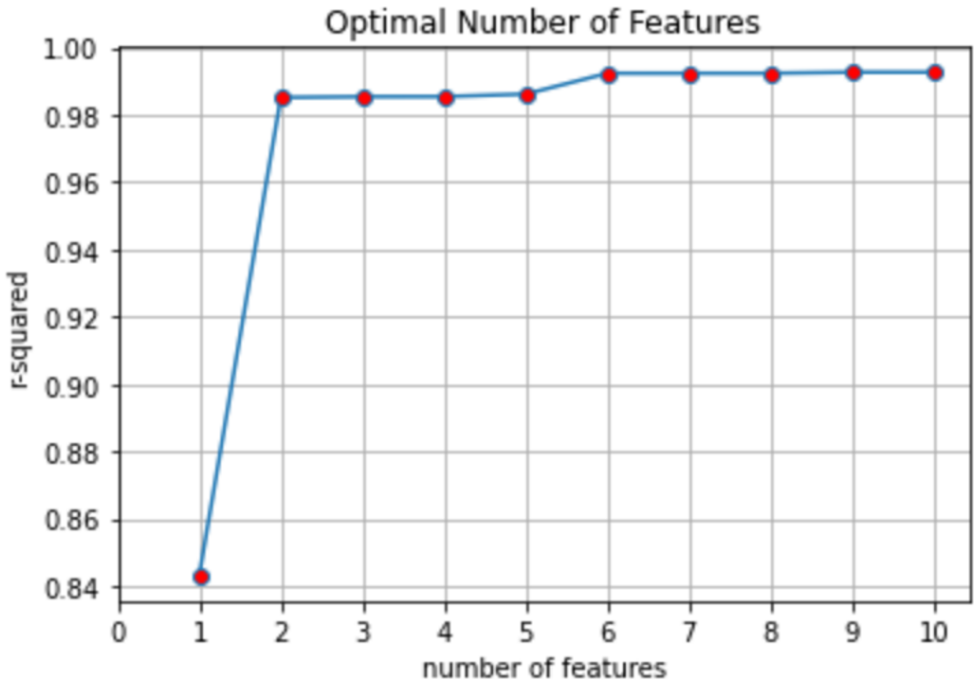



```
R-squared values:

Polynomial degree 1: train score=0.72, test score=0.75
Polynomial degree 2: train score=0.73, test score=0.75
Polynomial degree 3: train score=0.73, test score=0.75
Polynomial degree 6: train score=0.73, test score=0.76
Polynomial degree 10: train score=0.73, test score=0.76
Polynomial degree 20: train score=0.75, test score=0.78
```

Grid Search and k-fold Cross-Validation

In the last part, I establish grid search and 5-fold cross-validation to prevent bias and overfitting happens. Before building the model, I use Minmax to rescale all variables from 0 to 1. Then I use a dummy method to transfer text variables like fule_type. Then I build a hyperparameter data frame by using recursive feature elimination(RFE) to select relevant variables. As the result shows below, the R-squared value increase when more features use and reach nearly 0.99 when 10 features use.



Conclusion

- The observed data fit well with the regression model with a high R-squared value.
- Engines Size is the most influencing feature that affects CO2 emission.
- For Polynomial Regression, the more degrees we used, the higher R-squared we can get.
- We can get a higher R-squared score for more features according to Grid-Search.
- For the influence feature that affects CO2 emission of fuel type, diesel has the biggest impact on CO2 emissions. But demand for diesel remains high for medium and long-haul vehicles, especially long-haul trucks.
- Although ethanol fuel has a negative coefficient on carbon emissions, it does not mean that the use of ethanol fuel is encouraged, because its energy conversion efficiency is low, and it will consume more energy and pollute the production process.
- Dataset reference: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles.