Mahesh Babu Vusa

Prediction of Co2 emissions using linear regression model

Abstract

Using the data set provided, we built a simple machine learning model to predict the emission of Co2 gas (in grams per kilometer).In this model simple linear regression is used considering features like number of cylinders, engine size (L) and combined fuel consumption both in city and highway (in L per 100 kilometers).also multiple linear regression is used for same above features set.

Introduction: Predicting the emission of Co2 is important in the automobile sector .a company should be able to estimate the emission of co2 from the different vehicles that are producing and to pass the pollution control board tests. Data Science and predictive analytics play an important role in building models that can be used to predict the Co2 emission. In this project, we are provided with a data set fuel.csv containing 880 data points. Each data point represents a Co2 emission, and three features are provided as follows:

Target set:

- The column with header “Co2 emissions” indicates the tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving.

Features set:

- The column with header “cylinders” indicates the number of cylinders in a vehicle.
- The column with header “Engine size” indicates the capacity of engine in litres.
- The column with header “Comb” indicates City and highway fuel consumption ratings are shown in litres per 100 kilometres (L/100 km) - the combined rating (55% city, 45% high way) is shown in L/100 km.

Project Objective: The goal of this project is to use techniques of data science to estimate the Co2 emissions depending on the Engine size, cylinders and fuel consumption.

Exploratory Data Analysis: The data set was important in Python and calculations were performed using python (jupyter Notebook). We plot the following figures:

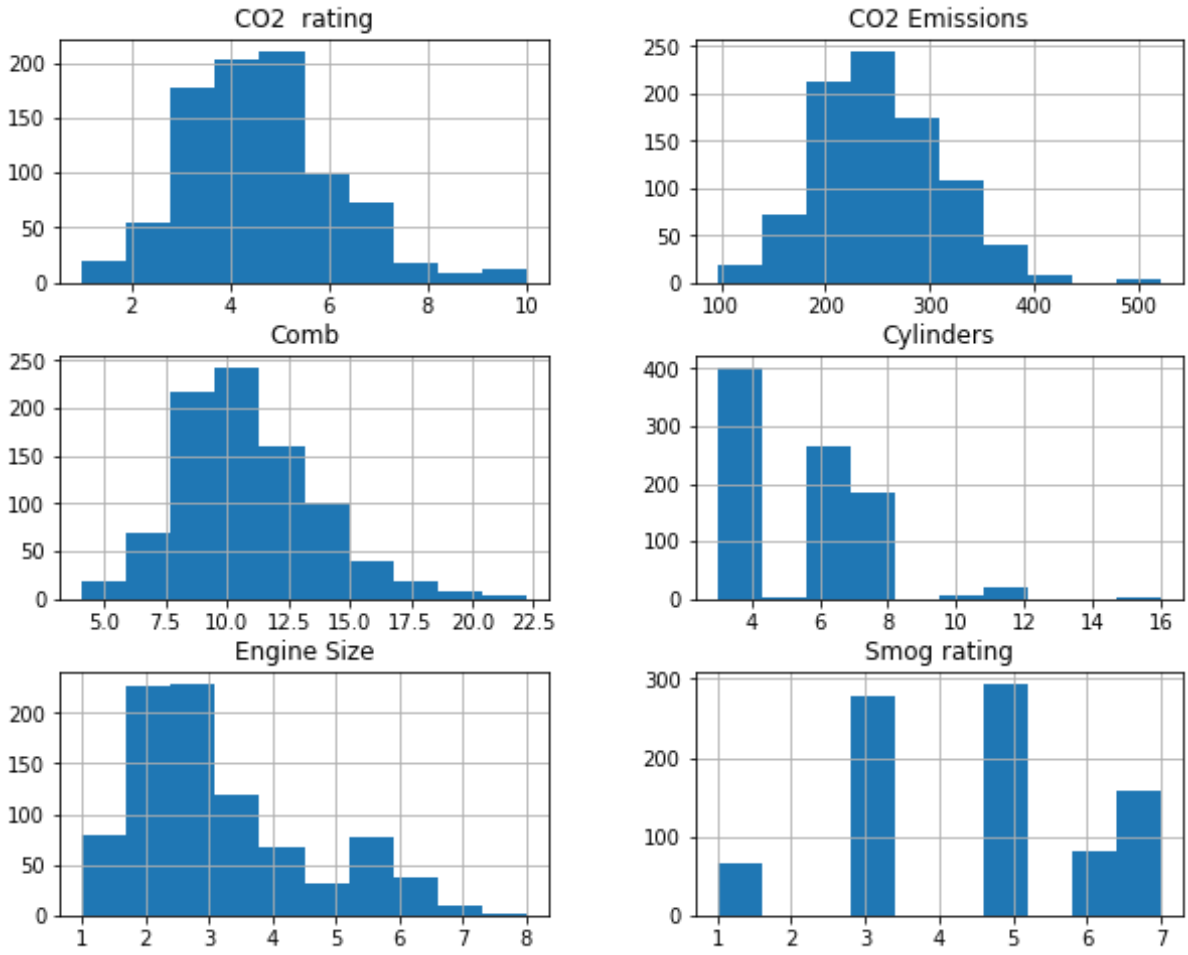

Fig1: histograms for different features and their variation

Fig1 - Analysis:

- Co2 emissions have Mean of 252.362 and median of 250, there were no considerable outliers.
- Cylinders have Mean of 5.65 and median of 6, there were no considerable outliers.
- Comb has Mean of 10.87 and median of 10.60, there were no considerable outliers.
- Engine size has Mean of 3.16 and median of 3.0, there were no considerable outliers.

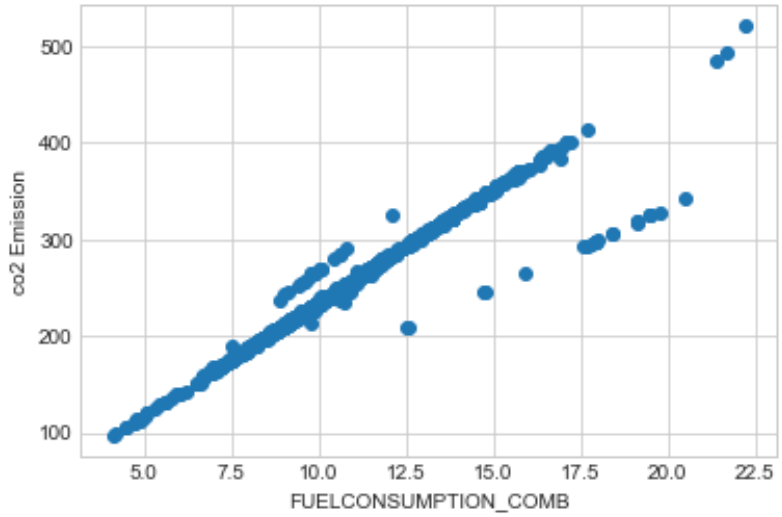

Fig2: Scatter plot combined fuel consumption (vs.) Co2 E missions

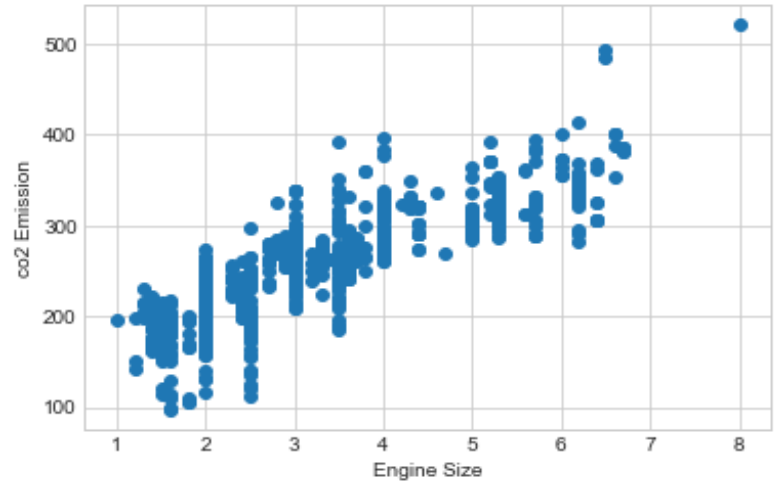

Fig3: Scatter plot Engine size (vs.) Co2 E missions

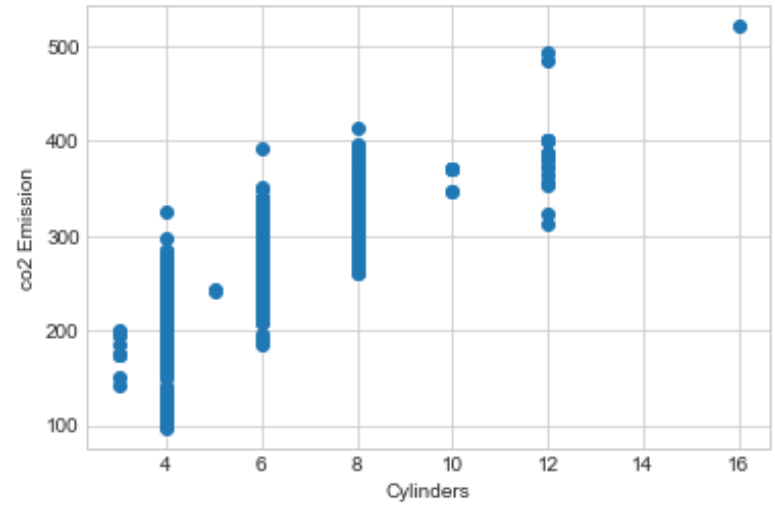

Fig4: Scatter plot Cylinders (vs.) co2 Emissions

Model selection: In the given data set by observing the scatter plots from fig2, fig3, fig4 the features showed the linear relation with the target set, we therefore choose linear regression model for predicting the emission of carbon dioxide for given features.

Model Calculations:

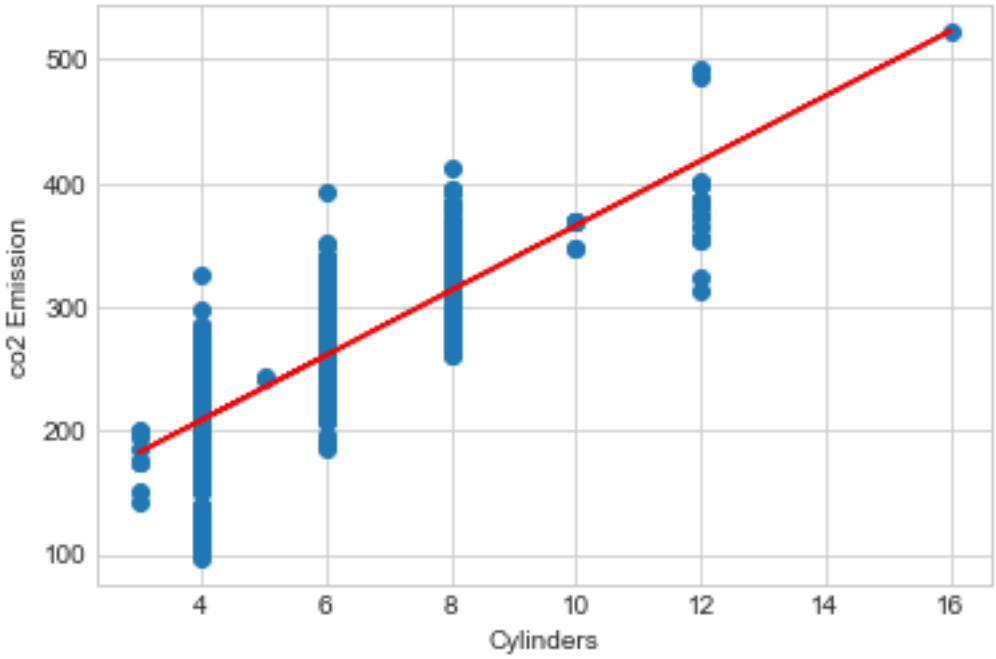

Fig5: Scatter plot after model fitting cylinders vs co2 emissions

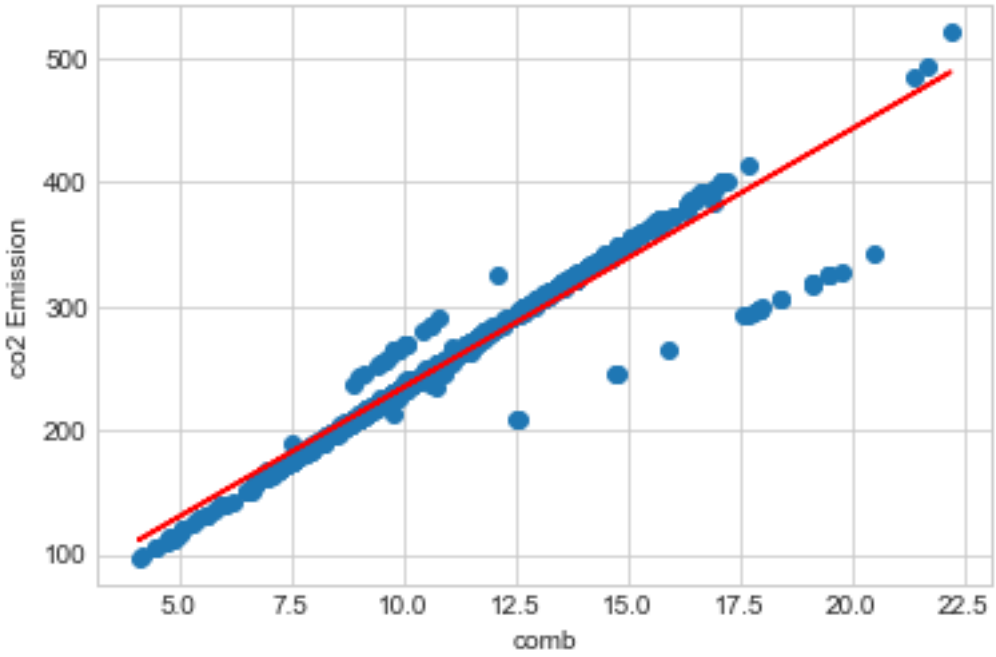

Fig6: Scatter plot after model fitting comb vs co2 emissions

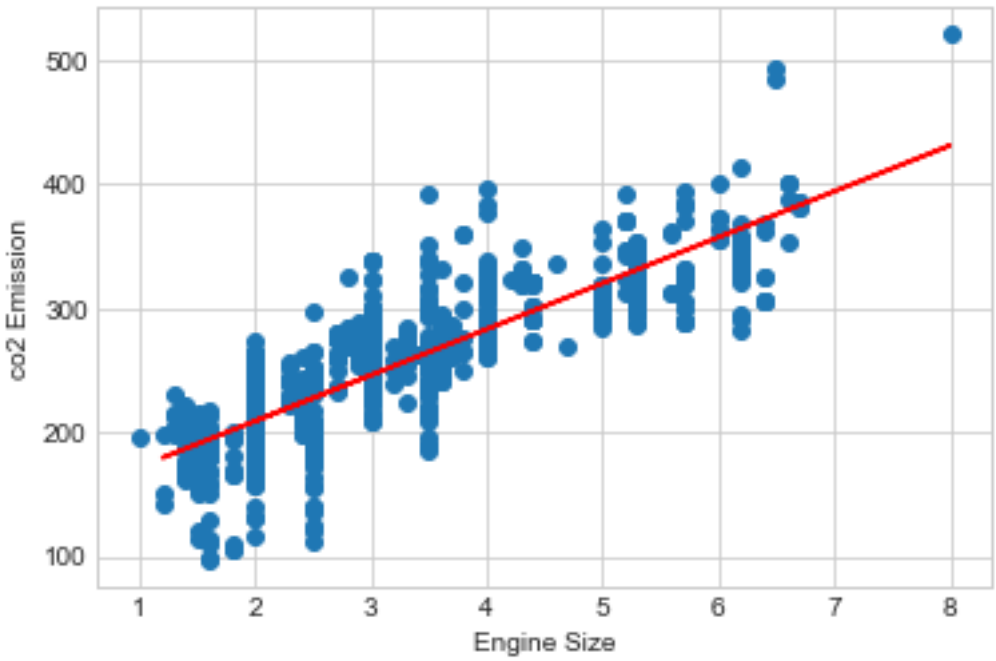

Fig7: Scatter plot after model fitting Engine size vs co2 emissions

Table explaining the linear relation between features and target

- **Simple linear regression:(Y=MX+C)**

Y	X	Coefficient(C)	Intercept(M)
Co2 emissions	Engine size	37.02	135
Co2 emissions	Cylinders	26.22	103.64
Co2 emissions	comb	20.84	26.13

- **Multiple linear regression:(Y=A1 X1+A2 X2+A3 X3+C)**

Y	X1	X2	X3	A1	A2	A3	C
Co2 emissions	Engine size	Cylinders	comb	2.53	4.06	17.51	31.18

- Residual sum of squares: 370.68
- Variance score: 0.89

Conclusions: We have presented a simple model based on the linear regression for predicting the Co2 emissions for different features like engine size, cylinders, fuel consumption (comb). Different models could be used such as KNN, decision forest regression, etc. I would like to try these different approaches to see if the results are comparable to the linear regression results.

APPENDIX: Python code for performing data analysis

```
# Python code for predicting Co2 gas emissions
# Author: Mahesh babu vusa
#Import Necessary Libarries

import matplotlib.pyplot as plt
import pandas as pd
import pylab as pl
import numpy as np
%matplotlib inline
from sklearn import linear_model
regr = linear_model.LinearRegression()
from sklearn.metrics import r2_score
# importation of data set

df1=pd.read_csv(r'C:\Users\Mahesh\fuel.csv')

# Exploratory data analysis

df1.head()
df1.describe()
df1.info()
df2 = df1[['Engine Size','Cylinders','Comb','CO2 Emissions','CO2  rating','Smog rating']]
df2.describe()
#histograms for different feautures and their variation
df2.hist(figsize=(10,8))
plt.show()
# scatter plots for different features vs Co2 emission
```

- plt.scatter(df2['Comb'],df2['CO2 Emissions'] , marker='o');
  plt.xlabel("FUELCONSUMPTION_COMB")
  plt.ylabel("co2 Emission")
  plt.show()
- plt.scatter(df2['Engine Size'],df2['CO2 Emissions'] , marker='o');
  plt.xlabel("Engine Size")
  plt.ylabel("co2 Emission")
  plt.show()
- plt.scatter(df2['Cylinders'],df2['CO2 Emissions'] , marker='o');
  plt.xlabel("Cylinders")
  plt.ylabel("co2 Emission")
  plt.show()

```
# training and testing by linear regression model (y=mx+c)

# multiple linear regression(y=a1 x1 +a2 x2+  .... +c)

msk = np.random.rand(len(df2)) < 0.8

train = df2[msk]
```

test = df2[~msk]

 # modeling

- from sklearn import linear_model
  regr = linear_model.LinearRegression()
  train_x = np.asanyarray(train[['Engine Size']])
  train_y = np.asanyarray(train[['CO2 Emissions']])
  regr.fit (train_x, train_y)
  # The coefficients
  print ('Coefficients: ', regr.coef_)
  print ('Intercept: ',regr.intercept_)
- # plot outputs which show the fit line to the data
  plt.scatter(df2['Engine Size'],df2['CO2 Emissions'] , marker='o');
  plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
  plt.xlabel("Engine Size")
  plt.ylabel("co2 Emission")
  plt.show()

 # Multiple linear regression model

- from sklearn import linear_model
  regr = linear_model.LinearRegression()
  x = np.asanyarray(train[['Engine Size','Cylinders','Comb']])
  y = np.asanyarray(train[['CO2 Emissions']])
  regr.fit (x, y)
  # The coefficients
  print('intercept: ', regr.intercept_)
  print ('Coefficients: ', regr.coef_)
- # Model evaluation - multiple linear regression
  y_hat= regr.predict(test[['Engine Size','Cylinders','Comb']])
  x = np.asanyarray(test[['Engine Size','Cylinders','Comb']])
  y = np.asanyarray(test[['CO2 Emissions']])
  print("Residual sum of squares: %.2f"
  % np.mean((y_hat - y) ** 2))

  # variance score: 1 is perfect prediction
  print('Variance score: %.2f' % regr.score(x, y))