Linear Regression model to predict the Value of Gas Emitted from Vehicle

Israa Abdulrauof Othman
Assistant Professor
Software Engineering Program
Elnasr Technical College

ABSTRACT
Global warming, jeopardizes the national security, endangers health and threatens other basic human needs. Some impacts such as rising seas, record high temperatures, and severe droughts and flooding are already increasingly common. Unfortunately, oil related emissions may rise in the coming years as refines "unconventional" oils, such as tight oil and tar sands and the oil industry extracts. Avoiding unnecessary emission from the oil we do use and using less oil is the real solution. This paper presents the application of machine learning model using linear regression techniques to supply fuel consumption of vehicles to a large dataset from IBM. The model gives:Mean Absolute Error (MAE) =22.78, residual sum square (RSS) =917.55, R2 score=0.067. The results contribute to the quantifying process of energyair pollution and cost caused by transportation, followed by proposing relevant recommendations for bothproducers and vehicle users. Future effort should aim towards developing larger datasets for building APIs and applications and higher performancemodels.

General Terms
Artificial Intelligence (AI), Machine learning, Linear Regression, CO2 Emission, greenhouse gas (GHG)

Keywords
Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines

1. INTRODUCTION
Nowadays, one of the biggest challenges to face is the reduction of greenhouse gas (GHG) emissions from the transport industry. In particular, the road transport sector accounts for about 80% of the whole energy demand required by transportation, due to its reliance on fossil fuels, represents one of the most important sources of GHG emissions in the world [1].Transport emits CO2, the most important greenhouse gas (GHG), andif global warming crosses the safety threshold of 2° C then the consequences could be anywhere between bad and catastrophic [2]. In fact, there isevidence already that the safety threshold may be 1.5[3]. To keep global warming below 2° C or 1.5atmospheric concentrations of GHGs must be stabilized and this willeventually require net zero annual emissions [2]. Worldwide,in 2014 transport as a whole was responsible for 23% of total CO2emissions from fuel combustion and road transport was responsible for20% [4].Since the United Nations Framework Convention on Climate Change(UNFCCC)came into force in 1994, global CO2 emissions have continuedto increase [5], withsome regions of the world increasing their total emissions substantially, such asChina and India, and others decreasing theirs, such as Europe[6].Machine learning and generally Artificial Intelligencehas become prevalent recently. People across different disciplines are trying to apply AI to make their tasks a lot easier. For example, economists are using AI to predict future market prices to make a profit, doctors use AI to classify whether a tumor is benign or malignant, meteorologists use AI to predict the weather, HR recruiters use AI to check the resume of applicants to verify if the applicant meets the minimum criteria for the job, etc. The impetus behind such ubiquitous use of AI is machine learning algorithms. The rudimental algorithm that every Machine Learning enthusiast starts with is a linear regression algorithm. [7]Fuelconsumption testingit would be difficult to drive every model of new vehicleon the road to measure fuel consumption. And it wouldbe impossible to get repeatable results that way becauseso many factors like weather and road conditions, to namejust two can affect a vehicle's performance.That's why vehicle manufacturers use standard,controlled analytical proceduresand laboratory testing to generate the fuel consumption data. [8]

2. LITERATURE REVIEW
Appl. Sci. 2022, 12, 803 4 of 29 INTEGRATION model for estimation emissions from measured fuel consumption. Moreover, it is also developed for the optimization and simulation a of trip based microscopic traffic [9]. A model named CMEM is proposed by a group of researchers to estimate parameters for a wide range of light-duty vehicles. Using more parameters, including 55 parameters. For dynamometer testing, this model uses emissions per second data of HC ,CO, CO2, and NO, along with physical vehicle features (vehicle mass , aerodynamic drag coefficient , and engine size) and operating features (acceleration and speed) [10]. Another example of using data intensive parameters is MEASURE, which was invented by the Georgia Institute of Technology. It calculates the emissions of VOCs, CO and NOx from vehicle operating modes, including acceleration, deceleration, idling and cruise. However, CO2 estimation is not included in this model, while it has over 30 features as its inputs [11]. Another one of the standard methodologies for road transport emission inventories in EEA member countries which is well known framework has been developed by the European Environment Agency (EEA) called COPERT [12]. It estimates primary air pollutants (VOC, SO2, CO, NOx, PM, NH3, heavy metals) and greenhouse gas emissions (CH4, CO2, N2O) using functions of the mean traveling speed throughout a complete driving cycle [13]. However, the framework neglected other characteristics while estimating the emissions of a specific

vehicle, such as engine model, cylinders and engine size.

3. METHODOLOGY

3.1 Simple Linear Regression

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the type of relationship between the independent and dependent variables and number of independent variables. [8] Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

y = a_0 + a_1 * x ## Linear Equation

The motive of the linear regression algorithm is to find the best values for a_0 and a_1. Before moving on to the algorithm, the following two important concepts used in linear regression:

3.1.1 Cost Function

The cost function helps to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since the best values for a_0 and a_1 is needed, this search problem converted into a minimization problem where the error between the predicted value and the actual value is minimized.

3.1.2 Minimization and Cost Function

$$minimize = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$j = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

The above function is used to minimize. The difference between the predicted values and ground truth measures the error difference. Then square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function. Using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

3.1.3 Gradient Descent

The next important concept is gradient descent. Gradient descent is a method of updating a_0 and a_1 to reduce the cost function (MSE). The idea is some values for a_0 and a_1 used initially and then these values iteratively will be change to reduce the cost. Gradient descent helps on how to change the values. To find these gradients, partial derivatives is taken with respect to a_0 and a_1. Now, to understand how the partial derivatives are found below some calculus is required.

$$j = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$j = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 . x_i - y_i)^2$$

$$\frac{\partial j}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1.x_i - y_i).x_i \Rightarrow \frac{\partial j}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i).\ x_i$$

$$\frac{\partial j}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1.x_i - y_i) \Rightarrow \frac{\partial j}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$a_1 = a_1 - a\,.\frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)\,.x_i$$

$$a_0 = a_0 - a\,.\frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

The partial derivate are the gradients and they are used to update the values of a_0 and a_1. Alpha is the learning rate a hyperparameter that mustbe specified. A smaller learning rate could getting closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that the minima could be over shouted.

3.2 The model building steps

Only the engine size property was used to create a simple linear prediction model to predict the amount of gas emitted by the vehicle, where x is the independent variable (Engine Size) and y is the dependent variable (Emission).

Table 1. Asample of the dataset records

ID	Engine Size (L)	Cylinders	Fuel Consumption City (L/100 km)	CO2 Emissions (g/km)
1	2.4	4	9.9	200
2	3.5	6	12.6	263
3	2	4	11	232
4	2	4	11.3	242
5	2	4	11.2	230
6	2	4	11.3	231
7	3	6	12.3	256

Firstly the dataset is loaded then checking for any missing or NA values in the training and testing dataset. After that exploring the relationship between the dependent variable and the independent variable as shown in the figure bellow:

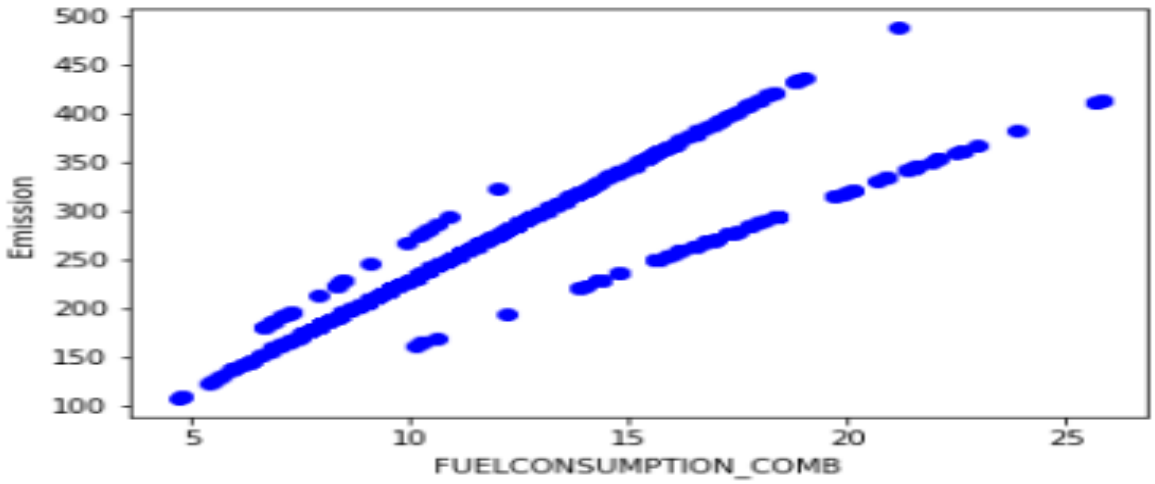

Fig 1: Realationship between dependent and independnt variables

The next step is visualising the regression equation fitment on the data as shwon in Fig 2.

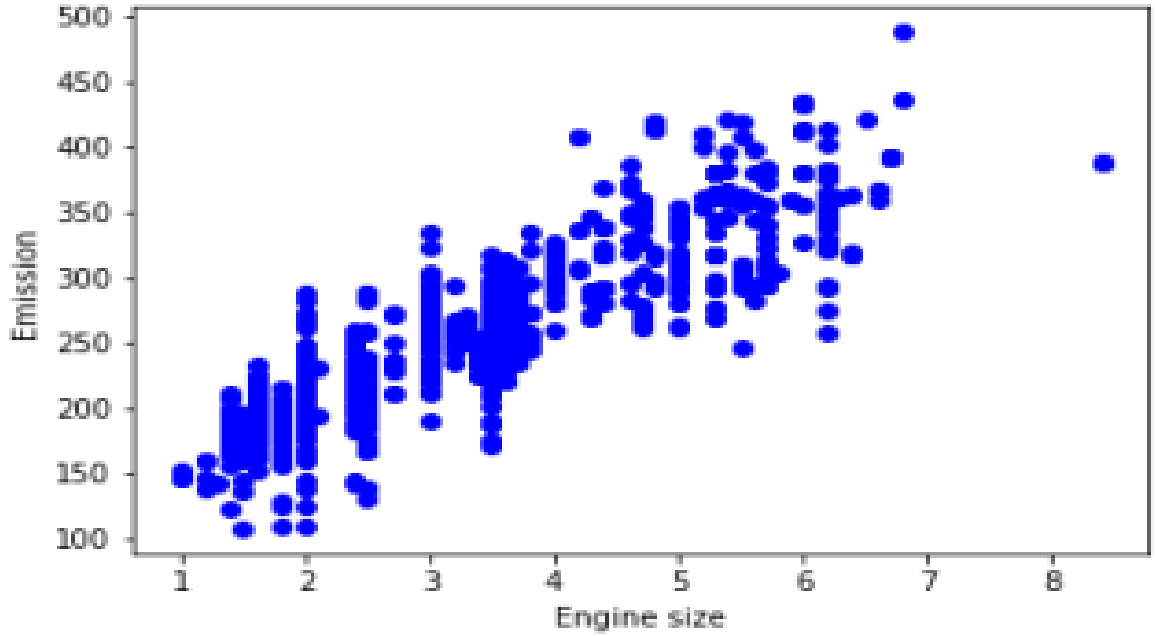

Fig 2: Visualizing the regression equation fitment

Figure below shows and plotting the features: cylinder, co2 emissions, engine size and fuel consumption.

```
viz = cdf[['CYLINDERS','ENGINESIZE','CO2EMISSIONS','FUELCONSUMPTION_COMB']]
viz.hist()
plt.show()
```

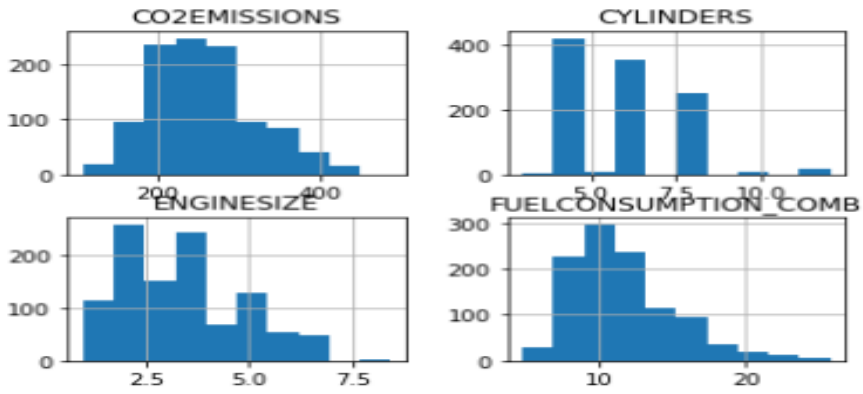

Fig 3:Plotting the dataset features

3.3 Model Evolution

In machine learning, in order to avoid overfitting of regression models, cross-validation is usually performed. This is done by splitting the data into training and test datasets (usually 75% and 25% of data respectively) [14].

3.3.1 Creating train and test

In order to define more reliable models, which make predictions independent from how the available data are subset, 10-fold cross-validation has been used in this study [14]. This means that the splitting process has been repeated randomized 10 times and 10 different models have been generated). The average performances, as root mean squared error (RMSE), and mean absolute error (MAE) have been used. Obtaining similar performances of the models for each split of the data indicates that the available information is not affected by bias and that the final results are not affected by how the data are split. On the other hand, variations between data splits indicate lack of reliability of the prediction model.Train /test splitting involve spletting the dataset into traing and testing set respectivly ,which are mutally ezlusive , aftet train with traing set and test with the testing set this provide more accurate evaluation on out of sample accuracy becase the testing dataset is not part of the dataset that have been used to train the data,this mean that the outcome of each point in this dataset is now known,making it grait to test with ,and scince this data has not been used to train the model has no knwlwge of the outcom of this data points so in essence it is truly an out of sample testing .

3.4 Modelling

After train and test spliting the model gives the cofficint and intercept as shown below :

```
[14]: from sklearn import linear_model
      regr = linear_model.LinearRegression()
      train_x = np.asanyarray(train[['ENGINESIZE']])
      train_y = np.asanyarray(train[['CO2EMISSIONS']])
      regr.fit (train_x, train_y)
      # The coefficients
      print ('Coefficients: ', regr.coef_)
      print ('Intercept: ',regr.intercept_)

Coefficients:  [[39.1757159]]
Intercept:  [125.66702924]
```

Fig 4. shows how the fit line over the data

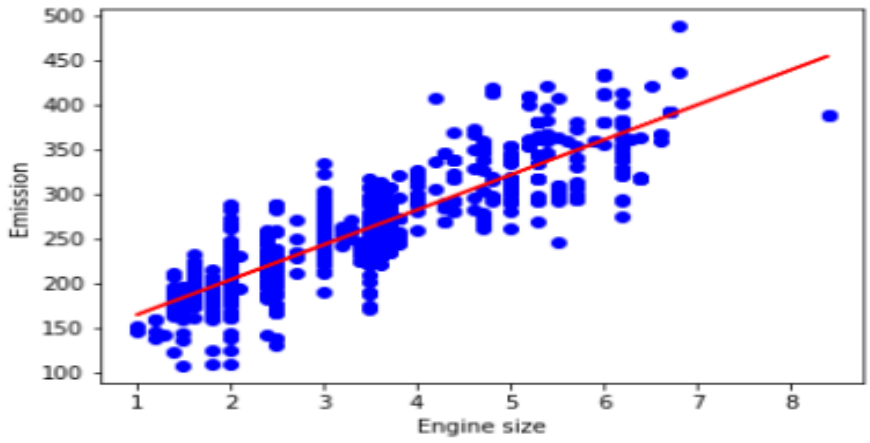

Fig 5. Plotting the liner line

Finally the vlue of mean absulue error ,residual sum of aqare and R2 scor is clauated using the model as desciped in Fig 5.

```
[21]: from sklearn.metrics import r2_score

      test_x = np.asanyarray(test[['ENGINESIZE']])
      test_y = np.asanyarray(test[['CO2EMISSIONS']])
      test_y_ = regr.predict(test_x)

      print("Mean absolute error: %.2f" % np.mean(np.absolute(test_y_ - test_y)))
      print("Residual sum of squares (MSE): %.2f" % np.mean((test_y_ - test_y) ** 2))
      print("R2-score: %.2f" % r2_score(test_y_ , test_y) )

      Mean absolute error: 22.78
      Residual sum of squares (MSE): 917.55
      R2-score: 0.67
```

Fig 6.Model getting MES, RSS and R2 scor

The following table (Table 2) shows the MAE and RMSE for the training and test sets, and the computed R2 for the developed machine learning regression models.

Table 2. Performance for the developed linear regression model

MAE	RMSE	R2
22.78	917.55gb	0.067

4. CONCLUSION

Machine learning is popular in solving the prediction problems of complex systems such as fuel consumption prediction. By making the model learn the training set, it is possible for the model to show a better prediction effect on the test setsimple linear prediction was used in this study ,because of the is ease of handling, as it lies in finding the equation of the straight line. One of the advantages of this method is that it is fast and easy to understand, as it does not need to adjust complex equations such as KNN, in addition to being easy to interpret.The dataset includes 25 variables in all and 947 record. In order to avoid over fitting, among all the parameters available only the most significant have been selected and included in the regression analysis. After an initial cut off of the variables performed by analyzing the correlation of each parameter with fuel consumption theGradient descent was used to updating a_0 and a_1 and reduce the cost function (MSE. Train /test splitting technique was used in order to provide more accurate evaluation on out of sample accuracy.as aresult The model gives:Mean Absolute Error=22.78, residual sum square RSS=917.55, R2 score=0.067.). The study recommend that it is important to include information on the level of congestion in the driving patterns to get more accurate emission predictions ,Future studies include estimating fuel consumption/emissions in a larger scale and a larger time span and studying approaches to reduce fuel consumption/emissions

5. ACKNOWLEDGMENTS

I would like to thank IBM for rendering help under IBM Academic Initiative.

6. REFERENCES

[1] EPA, 2017. Inventory of U.S. Greenhouse Gas Emissionsand Sinks 1990–2015. EPA-43-P-17- 001.

[2] Intergovernmental Panel on Climate Change, 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, [Core Writing Team: R.K. Pachauri and L.A. Meyer (eds.)], IPCC, Geneva, Switzerland. http://ar5-syr.ipcc.ch/.

[3] Science and policy characteristics of the Paris Agreement temperature goalC.-F. Schleussner,W. HareNat. Clim. Change, 6 (2016), pp. 827-835

[4] International Energy Agency, 2016. CO_2 emissions from fuel combustion by sector in 2014, in CO_2 Emissions from Fuel Combustion, IEA, 2016. In CO_2 Highlights 2016 - Excel tables. http://www.iea.org/publications/freepublications/publication/co2-emissions-from-fuel-combustion-highlights-2016.html.

[5] United Nations Environment Programme (2016)A UNEP Synthesis Report, Nairobi, November.http://uneplive.unep.org/theme/index/13#egr

[6] Trends in Global CO_2 Emissions: 2014 ReportJ.G.J. Olivier and M. Muntean PBL Netherlands Environmental Assessment Agency and Institute forEnvironment and Sustainability of the European Commission's Joint

Research Centre, The Hague (2014). http://edgar.jrc.ec.europa.eu/news_docs/jrc-2014-trends-in-global-co2-emissions-2014-report-93171.pdf

[7] Published in Towards Data Science https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a\23/8/2022, 1:55 PM

[8] Natural Resources Canada https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

[9] So, J.; Motamedidehkordi, N.; Wu, Y.; Busch, F.; Choi, K. Estimating emissions based on the integration of microscopic traffic simulation and vehicle dynamics model. Int. J. Sustain. Transp. 2018, 12, 286–298.

[10] Hung, W.T.; Tong, H.Y.; Cheung, C.S. A modal approach to vehicular emissions and fuel consumption model development. J. Air Waste Manag. Assoc. 2005, 55, 1431–1440.

[11] Fomunung, I.; Washington, S.; Guensler, R. Comparison of MEASURE and MOBILE5a predictions using laboratory measurements of vehicle emission factors. In Transportation Planning and Air Quality IV: Persistent Problems and Promising Solutions; American Society of Civil Engineers: Reston, VA, USA, 2000.

[12] Ntziachristos, L.; Gkatzoflias, D.; Kouridis, C.; Samaras, Z. COPERT: A European road transport emission inventory model. In Information Technologies in Environmental Engineering; Springer: Berlin/Heidelberg, Germany, 2009; pp. 491–504.

[13] Ntziachristos, L.; Samaras, Z.; Eggleston, S.; Gorissen, N.; Hassel, D.; Hickman, A. Copert iii. In Computer Programme to Calculate Emissions from Road Transport; Methodol. Emiss. Factors (Version 2.1), Eur. Energy Agency (EEA), Cph.; European Energy Agency: Copenhagen, Denamrk, 2000.

[14] James, G., Witten., D., Hastie, T., and Tibshirani, R., 2013. An Introduction to Statistical Learning, New York, NY: Springer Science+Business Media New York. Available at: http://link.sprin-ger.com/10.1007/978-1-4614-7138-7.