International Journal of Computer Applications (0975 - 8887)
Volume 183 - No.5, May 2021

CO_2 Emission Prediction and Identification of Relevant Factor to the Emission based on Machine Learning Analysis: A Study in Bangladesh

Fazle Mohammad Tawsif
Shahjalal University of Science and Technology
Institute of Information and Communication Technology
Sylhet, Bangladesh

Md. Jubair Ibna Mostafa
Islamic University of Technology
Computer Science and Engineering
Dhaka, Bangladesh

B.M. Mainul Hossain
University of Dhaka
Institute of Information Technology
Dhaka, Bangladesh

ABSTRACT

Carbon Dioxide (CO_2) is the major contributing factor to global warming and climate change. Growing industry and civilization increase the amount of CO_2 emissions rapidly. Being a developing country in South Asia, Bangladesh is also facing the consequences like climate change for the last few decades due to CO_2 emissions. It is essential to monitor CO_2 emissions to take necessary steps towards reducing the emission rate by identifying contributing factors. Authors have analyzed a time series data of 42 years CO_2 emission of Bangladesh. Diverse factors of CO_2 emission covering multiple areas like environment, fossil consumption, and energy production are considered. By analyzing these data, it is showed as a prediction model for CO_2 emission rate. In this literature, authors have identified the relevant factors that have much impact on CO_2 emission in Bangladesh along with prediction. Different machine learning algorithms like Linear regression, Multi-Layer Perceptron (MLP) are applied in this study to build the prediction model to address the issue.

Our result depicts that CO_2 emission follows a linear model, and environmental factors are mostly related to CO_2 emission. Few of these factors show high relation with CO_2 emission for the past few years. It shows that the amount of rainfall is decreasing due to overall emission escalation. According to the data linearity found in CO_2 emission in burning natural gas and Solid Fuel, a regression model is built based on these features. It successfully predicts the emission with significantly low RMSE. However, rainfall is not affected by the CO_2 emission as the Correlation matrix does not provide any meaningful information. Instead, decreasing of Forest and Agricultural land have an impact on the emission. The effect of the overgrowing population in the last few decades has exponentially increased in CO_2.

General Terms

CO_2 Prediction using regression, CO_2 Emission

Keywords

Linear Regression, CO_2 Emission Prediction, CO_2 Emission, CO_2 relevant factor, Parametric Method Analysis

Introduction

Carbon Dioxide (CO_2) is one of the main elements in the Chlorofluorocarbons (CFC) gas. CFCs are responsible for climate change, and the impacts of climate change on weather, human beings, animals, and plants are critical and disastrous. The literature already mentioned many consequences like floods, extreme weather, increases in temperature, heatwaves, storms, etc. are some of the effects of climate change on weather [7]. Moreover, animal extinction, forest reduction, and human sufferings are also the consequence of this change. The gradual change in rainfall leads to warming in some areas, which results in droughts, water scarcity. It also harms the water supply, agriculture, and overall environment [1]. Such warming extensively environmental change can antagonistically influence to high paying nations just as low-paying nations [12]. As Bangladesh is a low-income developing country, the impacts of climate change due to CO_2 escalation should be considered important.

Covering a small area of 147,570 sq km, Bangladesh is the most populous country in the world. Though it is a small developing country, emission of CO_2 is proliferating, and the country goes through adverse effects of climate changes. Finding the significant sources from enormous CO_2 emitting medium can reduce the emission. The increasing number of people, industries, vehicles, and decreasing forest area, agricultural land, and river areas indicate national growth to modern civilization. This scenario also denotes the emission of greenhouse gas, more specifically carbon dioxide emission severely. Consideration of different factors and previous data that may affect the emission of CO_2 can help minimizing CO_2 emission locally and globally, leading to a positive impact on the environment. In this research, the authors mainly focus on identifying the leading cause of CO_2 emission in Bangladesh. Visual-

izing the data and finding the significant components using machine learning models and techniques will help people understand the overall scenario. A prediction model is fitted with the data that may help to identify the future impacts. In the analysis, output for rainfall is considered a result of CO_2 emission escalation. It shows a gradual yearly decrease in rainfall.
The Pearson correlation matrix is used for the data analysis to find the correct related data and eliminate unwanted features. Component residual plotting is implemented in gathering information data linearity. This plotting helps to find a model for further analysis. An attempt has been made to analyze the features where it can predict the output. Burning Natural age and solid fuel is the major contributing factor in CO_2 emission. It is discussed in the result discussion section. Also, the Linear Regression model is used for prediction in CO_2. From the correlation matrix, rainfall does not show any relation with CO_2. Correlation coefficient analysis showed in Table 4.
But apart from these factors, there may be other factors such as electricity production, forest and agricultural land, total population, industrial development that can affect the CO_2 emission. In this literature, the factors as data related to many of these factors are available are focused. It is experimented to explore how these features contribute to the CO_2 emission. One of our main goals was to identify the significant factors related to CO_2 emission and predict a model that can provide further assumption. These findings can help a large group of people like policymakers and regulatory bodies make a decision. Moreover, some action can be taken from these findings for environmental well-being, like preventing deforestation and controlling the industrial emission of harmful gases. Industrial development is one of the critical factors with a high correlation value to CO_2 emission.
The paper is organized in the following order. Section two discusses the related works done in predicting and analyzing CO_2. The third section provides the Data description. The fourth section explains the methodology used to design the model, which includes data collection, preprocessing of data, feature extraction, the definition of the class labels, and analyzing the performance of different classifiers. Finally, an overview of the developed method is given, and scopes for future improvements are discussed.

1. LITERATURE REVIEW

CO_2 is the main element of greenhouse gas that is responsible for climate change. World wide industrial revolution to urbanization, all countries emit a large number CO_2 every day. Substantial changes in climate responsible for the disaster and global warming. Literature has tried to predict CO_2 emission from a different approach, applying different machine learning algorithms, analyzing prediction results such as [21].

Many research has already been conducted regarding climate change caused by human damage. Some changes cannot be reversed. The severity of these changes depends on their magnitude and not changeability [20]. An in-depth assessment was conducted to find out the problems due to CO_2 emission. This showed the review for demographic differences [16]. They have used economic changes due to energy consumption change. Over time, Technological advancement and globalization have become a potential cause for rocketing the CO_2 emission. Reduction the population growth in some particular area of the world and regularizing the energy consumption can lead to a decrease in the rate of adverse climate changes [16]

In recent years, [3] showed that developing countries impact on CO_2 for their growing economy. Their findings that developing countries are shifting towards fossil fuel energy consumption whereas the developed countries already started moving to renewable energy. This imbalance is not helping to reduce CO_2 emission. They suggested that as economic growth is significantly related to energy consumption, they should skip a step and move towards renewable energy. Another research is based on data for 14 years of carbon emission by using renewable energy and foresting. It showed that CO_2 emission is reduced by using renewable energy and foresting [22] Deforestation is one of the leading cause of the increase in CO_2 emission [13]

Increased use of Nonrenewable energy and economic growth all over the world had a severely adverse effect on the atmosphere. A review of over 128 countries carbon dioxide emissions and an increase in nonrenewable energy found that the amount got worse day by day along with the growth of population. They considered data for 14 years over countries from different region [10] Moving apart from using gasoline fuel vehicles which is nonrenewable energy toward electric cars, can reduce carbon emission. Fossil fuel is impacting the atmosphere significantly. [11] showed that it would likely fall steadily shortly if it is moved from fossil fuel to renewable energy.

Pre-Industrial development and Post Industrial development made us building many Industrial structures. The population increasing sharply, and people are building in parallel. Analyzing eight different scenario, [23] presented that population growth is contributing highly in CO_2 emission.

Real-time CO_2 emissions are analyzed through mobile devices. Different sensors from mobile devices are used to gather information. The author followed an autonomous way to classify a model using the users relocation time by time. In this algorithm, they used eight classes to fit the model in the decision tree. The algorithm uses the Fast Fourier Transform (FFT) to compute the features. This FFT considers the total speedup measured using the accelerometer of the phone. Later, these coefficients are used for that algorithm. To evaluate the proposed method, an application had also been developed [15].

The Petroleum exporting countries comprise a community where they provide estimation for carbon emission concerning the export ratio. The community was known as The Organization of Petroleum Exporting Countries (OPEC). This CO_2 emission ratio is related to the economic growth of the importing countries. The more economic growth develops, the more fuel burn increases. So a perfect prediction from this organization can provide a reference to this development. By regularizing the export, the OPEC community can provide a view for decreasing the evergrowing global warming phenomena [7]

A case study in china to predict carbon dioxide emissions have conducted. In their analysis, they used an extraordinary merged model combining the principal component analysis (PCA) with regularized extreme learning machine (RELM) to determine CO_2 emissions forecast dependent on the information from 1978 to 2014 in China [21]. They have shown that their proposed hybrid model outperforms most of the impactful models. They made a comparison for the resultant errors. They used the RELM model, the extreme learning machine (ELM), backpropagation neural network (BPNN), GM(1,1), and the Logistic model and performs

better than any other mentioned models in terms of computing time.

2. DESCRIPTION OF DATA

There are many factors related to carbon dioxide (CO_2) emissions. These factors are increasing the density of CO_2 in the atmosphere. Besides harming the atmosphere, it is decreasing the oxygen supply for breathing, particularly in compact spaces. If the concentration level rises as high as 10 percent or greater, carbon dioxide can be the reason for death, convulsions, or unconsciousness. This rise of concentration can cause harm to the unborn child in the mothers ovary. The lessen amount of CO_2 in the atmosphere is also harmful. It can result in vision problems, nerve injury, muscle problems, high blood pressure, and breathing problem. It can likewise create headache, fatigue, numbness, memory loss, muscle tremor, vomiting, confusion, skin and eye burns, and ear issue.

Last few decades, a large number of Industrial area has been established. Various multinational companies are building their factories in Bangladesh. As a result of this, the burning of fuels, natural gas, using electricity has increased, which is related to CO_2 emission. There is a significant change in rainfall amount over the year. They are also causing unusual high temperatures.

For our analyses on CO_2 emission, these related factors data are collected from various authentic sources. Most data CO_2 and CO_2 emission factors are collected from World Bank Data library[4]. Rainfall and Temperature data of Bangladesh is collected from BARC [8]. Vehicle data is collected from BRTA [2]. Data from 1970 to 2013 is formulated for this analyses.

Table 2 states the data are used for this analysis. CO_2 emission data are distributed in six different predictor classes. They are Environmental, Energy Production, Energy Consumption, Population, Industrial investment, and Emissions. Factors like Agricultural land, rainfall, etc., are related to the Environmental class. The energy Production class contains all sorts of sources pertaining to everyday energy needs. Among the other prediction classes, emission provides insights into CO_2 production and releasing the toxic part into the environment.
In the data, initially, 57 rows indicate the yearly data. Except for the annual population count, all the data are decimal values. Each year has 16 features. These features are mentioned in this table 1. There is some missing information in CO_2 data. As Bangladesh got independent in 1971, data before that time were missing for CO_2 emission from electricity production and fuel burning. As these features are important factors, as explained in Figure 3, the Pearson correlation graph, rows before 1971 are omitted. However, these data possessed significant information in land and population information. Table 2 shows the insights of each features considered in this where Min, Max, Mean value, and variance are presented.
Finding relation among factors and predicting CO_2 emission is needed for human awareness and control. Because most of the elements are human-created, causing environmental hazards, considering data of these factors from 1960 to 2016 would provide enough knowledge to identify highly related factors and suggest optimal control over these factors.

3. METHODOLOGY

The features were selected from the data collected from the World Bank statistics of Bangladesh. A feature selection is performed to find out those to find significant and non-signification features from all collected features. This research follows the step-by-step identification process with Akaike Information Criterion (AIC) from different feature selection techniques. In this procedure, Features are introduced and truncated at every stage dependent on some model, which is, AIC [14]. Section 4.1 discusses the details of the collected data and shows the selected features. In this section, a description of our research is given. Figure 1 provides an overview of the whole system. Each step is explained in the subsequent subsections.
The training data was used to predict the carbon dioxide (CO_2) in the next year using algorithms, specifically linear regression, random forest, and multilayer perceptron with backpropagation. To choose the best-fitted model by applying these algorithms, all of those are compared and analyzed with a Paired T-Test by considering Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Correlation Coefficient. The test has been performed to know the difference of results among the three techniques or get the statistically significant impact. The following sections briefly describe the data organization process and linear regression, random forest, and multilayer perceptron with backpropagation to understand this work.

3.1 Data Collection

In this research, All the data are collected from renowned sources. These sources provide a wide range of data related to the environment, lands, vehicles, etc., for Bangladesh. As mentioned in 2 section, this dataset is prepared with data from 1960 to 2016. They contain information for CO_2 emission-related factors. Most of the factors collected individually with some predefined parameters related to CO_2. Those with some unrelated information are truncated from the dataset.

3.2 Data Processing

The data sources contain data in both CSV and Excel formats. As multiple sources is used to collect data, some data showed noise in some years. Due to a mismatch of the data collection year in some data sources, reduced the range for analysis years from 1960-2016 to 1972-2014. Because some data sources did not contain data before 1972, as Bangladesh became an independent country in 1971, this data includes more than 30 variables. Each one of these factors can be utilized as an element. Yet, while handling information, a portion of the estimations of these factors are discovered to be missing. In this way, those variables are taken out of our dataset. Some variables were irrelevant in our research context. They were removed too. After all the clearing, 16 variables remain in the dataset with data from 1972-2014.

3.3 Feature Extraction

As mentioned earlier, the dataset has more than 30 variables for each year. These 30+ variables were later reduced to 16 variables due to the context relevancy problem. After that, they are converted into 16 features for each year. All these features have continuous value.

3.4 Linear Regression

Linear regression is one of the pioneer methods for machine learning. It accepts a training data set as a model for prediction. It follows $\hat{y} = \alpha x + c$ where $\hat{y}$ is the output estimation of the variable y using x as a predictor or input, c is the intercept, and α is the slope. The difference between the training result and predicted value, $y - \hat{y}$ is denoted as the residual of the linear model. It tries to

Table 1. : Predictors, Names, Description and Data Source

Type	**Predictor Class**	**Name**	**Acronym***	**Definition**
Predictor	Environmental	Agricultural Land (sq. km)	al	The amount of agricultural land
		Forest Area (sq. km)	fa	Total area of forest
		Adjusted savings: net forest depletion (current US$)	afd	Saving in US$ by forest depletion
		Rainfall (mm)	rf	Average yearly rainfall
	Energy Production	Power from coal source (% of total)	epc	The percentage of Electricity production from coal sources
		Power using natural gas (% of total)	epg	The ratio of electricity from natural gas sources
		Power using oil sources (% of total)	epo	The percentage o Electricity production from oil sources
	Energy Consumption	Fossil fuel usage (% of total)	fec	The percentage of Fossil energy consumption
	Population	Total Population	p	Total number of population
	Industrial investment	Industry value added	iva	Industry value added (US$)
	Emissions	Gas fuel (kt)	gfe	The amount of CO_2 from gaseous fuel
		Liquid fuel (kt)	lfe	The amount of CO_2 from liquid fuel
		Solid fuel (kt)	sfe	The amount of CO_2 from solid fuel
		Methane production	ame	The amount of methane emission from agriculture
Response		CO_2 emissions	cem	The amount of CO_2 emissions in next year

* Abbreviations are given for every one of the predictors for presentation
Rainfall data collected from Bangladesh Agricultural Research Council and All other data collected from World Bank's statistics for Bangladesh

Table 2. : Predictor Classes, Names and Description

Type	**Predictor Class**	**Name**	**Min**	**Max**	**Mean**	σ^2
Predictor	Environmental	Agricultural Land (sq. km)	90990	104400	96580	3712.925
		Forest Area (sq. km)	14290	14940	14620	198.8618
		Adjusted savings: net forest depletion (current US$)	19390000	531800000	145900000	107526659
		Rainfall (mm)	147.4	235.9	200.5	21.55048
	Energy Production	Power from coal sources (% of total)	0	3.215	0.4559	0.9248901
		Power using natural gas (% of total)	34.69	92.7	72.8	19.36065
		Power using oil sources (% of total)	1.765	43.11	15.85	12.34774
	Energy Consumption	Fossil fuel usage (% of total)	20.71	73.77	48.26	16.59389
	population	Total Population	48200000	163000000	103200000	36304650
	Industrial investment	Industry value added	298100000	60550000000	10700000000	13699741703
	Emissions	Gas fuel (kt)	715.1	45970	14800	13611.97
		Liquid fuel (kt)	2428	15230	7545	3414.843
		Solid fuel (kt)	99.01	3751	1088	1064.399
		Methane production	0	81570	69010	12375.4
Response		CO_2 emissions	0.05191	0.4591	0.1941	0.1164843

reduce the summation of mean squared error represented as MSE using a predictor response training data set. Instead of MSE, the sum of squared residual symbolized as RSS can be used to provide better insights.

$$RSS = \sum_{i=1}^{k} (y_i - \hat{y_i})^2 \tag{1}$$

In linear regression, prediction utilizing different predictors are performed by the using the model below.

$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n + c \tag{2}$$

Similar to the single parameter linear regression called as univariate regression, this multiparameter or multivariate one aims to derive α_1, α_1 etc. by reducing the sum of squared residuals RSS using the Least-Squares method. Equation 3 shows the RSS by substituting the value of $\hat{y}$ into Equation 1.

$$RSS = \sum_{i=1}^{k} (y_i - (\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n + c)) \tag{3}$$

Fitting a model into linear regression, it is required to satisfy six predefined hypotheses, which are predicting variable and output without any calculation problem, without any correlation between predictors, data linearity, homoscedasticity, which is data with no noise, and relationship with a dependant and independent variable, normality of residuals and no collinearity between predictors[17]

3.5 Random Forest

Random forest is an advanced learning process for data classification, regression. This performs by building a multi-dimension of decision trees during training and results in the class that is the mode (classification) or mean prediction (regression) of the different trees. It is an advanced decision tree that performs better than bagged decision trees [5]. Bagging is a concept for reducing variance to avoid overfitting [9]. If we try to fit the data into a specific decision tree in the training stage, It can result in overfitting. For this reason, training data is bootstrapped. That is, a chunk of data is consistently collected from the training sample, and a new model of the decision tree is built for every set of these. In order to get a lower variance throughout prediction, the result for all the decision trees is considered by using the average value over them. Consequently the variance found from the average of n different tree model $x_1, x_2, x_3, \cdots, x_n$ with σ^2 variance is $\frac{\sigma^2}{n}$. However, suppose the data contains a strong independent variable that plays as a predictor. In that case, the bagging trees always tend to generate similar output because the heavyweight variables will always try to stay on the top. So, the variance will be similar. They will not have a significant change from the average of different trees due to similarities. From this problem, Random forest comes to play. It avoids the similar variance by taking a subset of predictors from the data and running it multiple times. In general, $k = \sqrt{n_p}$ number of data is selected at different run where n_p denotes amount of predictors is used for every execution.

3.6 Multilayer Perceptron

The artificial neural network is also used in the form of Multilayer Perceptron(MLP). It is another area in the neural network. It uses a feed-forward network. Like other neural network models, it uses three layers. They are an input layer, multiple hidden layers, and a final layer for output [18]. It uses additional perceptron models, which are named Multilayer Perceptron(MLP). These additional perceptrons are incorporated in the hidden layers. It receives actual input values as the input layer. Multiple perceptrons process the values in the middle layer by multiplying them with some wight, which plays as a different neuron in the hidden layer. The final layer provides the output for predictors by using some activation function. A backpropagation algorithm is applied in each layer by passing the weighted sum from one neuron to another [19]. The mathematical equation 4 for MLP is given below, where this equation is applied at each perceptron in all the layers.

$$y_k(i) = \theta \times \Big[\sum_{j=1}^{n} x_j(i) \times w_{jk}(i) - th_k \Big] \tag{4}$$

In the equation, $y_k(i)$ is the value for each perceptron k. i is the iteration number for which it is calculating. $x_1, x_1, \cdots, x_n$ is passed from the earlier layer and acts as input in the current layer. $w_{jk}(i)$ is the multiplied weight for each neuron, th_k is the accepted threshold, and θ is the mentioned function for output. We excluded the threshold for regression. Our research preferred using the sigmoid function for activation function in the final layer as many of the refereed papers suggested this function in their models. We considered the following Equation 5.

$$y_s = \frac{1}{1 + e^{-x}} \tag{5}$$

MLP uses backpropagation. It helps to reduce the error between layers. Error from the output layer is traverse back to top neurons to tweak the weights for with it can reduce the error between layers. The difference between output prediction and training set result is considered an error. It back propagates to the previous neurons if it crosses a certain threshold. Equation 6 is used for calculation.

$$e_o(i) = y_{xp}(i) - y_p(i) \tag{6}$$

In the equation, $y_{xp}(i)$ is the expected training dataset output. $y_p(i)$ is the calculated result output from $i-th$ iteration of the final layer. Later on, Using backpropagation, weight of the previous layers is recalculated using this formula 7.

$$w_{jo}(i+1) = w_{jo}(i) + \delta w_{jo}(i) \tag{7}$$

Here, $\delta w_{jo}(i+1)$ is calculated from Equation 8 and 9.

$$\delta w_{jo}(i) = \alpha \times y_j(i) \times \delta_o(i) \tag{8}$$

$$\delta_o(i) = y_o(i) \times (1 - y_o(i)) \times e_o(i) \tag{9}$$

However, for the hidden layer, weights are updated as follows.

$$w_{sj}(i+1) = w_{sj}(i) + \delta w_{sj}(i) \tag{10}$$

$$\delta w_{sj}(i) = \alpha \times x_s(i) \times \delta_j(i) \tag{11}$$

$$\delta_j(i) = y_j(i) \times (1 - y_j(i)) \times \sum_{k=1}^{o} \delta_k(i) \times w_{jk}(i) \tag{12}$$

Here, $x_s(i)$ is the passed input value for s number of neuron. $y_j(i)$ is predicted result for $i-th$ iteration. The backpropagated result is represented as the summation, which is denoted as errors in MLP. o indicates the neuron from which is collected the errors. By following this, each iteration is evaluated, and error from the output layer

propagated backward to the corresponding neuron. This process is repeated until a certain threshold is reached.
After finalizing the algorithm, the prediction model is used to analyze how different predictors influence the emission of carbon dioxide. Next, Principle Component Analysis is performed to find which principal components explain most of the variances of the data and how these components influence the output.

4. EXPERIMENTATION

In this section, the details of the data are described and explored. Moreover, the Paired T-Test performed to select the best predictor is also presented.

4.1 Data

Most of the data were collected from the latest data from the World Bank. From those data only Bangladesh, related data was extracted. To form the training data, different features-based data of Bangladesh was collected and aggregated. The sources of the selected predictors and the output for the predictors are presented in Table 1. The table shows that 14 predictors were selected, divided into six different classifications according to their type of similarities. The environmental predictors are related to agricultural land, forest area, forest depletion saving, and weather-related rainfall. Since rainfall is influenced by forest area and forest depletion saving is related to the forest area, both of this information is considered in environmental class. The energy production class consists of three different electricity production sources: coal sources, natural gas sources, and oil sources. A considerable amount of CO_2 is producing from the electricity production of these sources. In energy consumption, fossil fuel energy is considered. All livings are breathtaking every day. There might be a relation between CO_2 emission with the number of people. Industrial value add referred to urbanization and industrialization. Since industrialization means more fuels, oil is burnt, this also affects emissions. Lastly, different sources of CO_2 emissions were classified into emissions classes to relate with total carbon dioxide CO_2 emissions. The response is the emission of carbon dioxide in upcoming years.

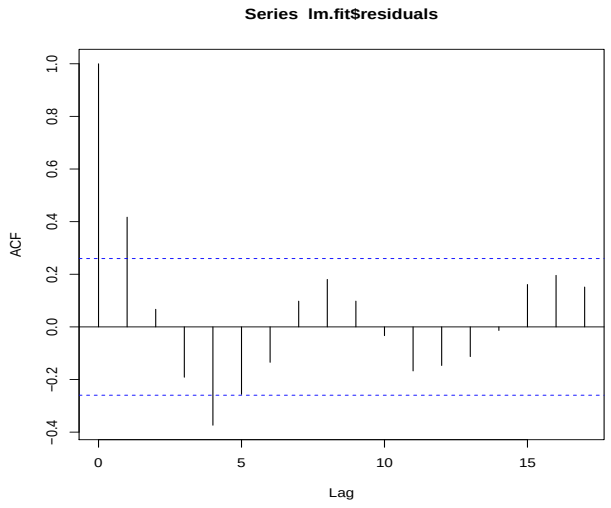


Fig. 1: Autocovariance and Autocorrelation Functions Plot

Table 2 demonstrates various insights among the data. They are Mean, standard deviation, maximum, and minimum values. An Adjusted saving of forest depletion, total population, and industry value added to have a larger value range than other predictors from the data. However, Section 4.2 shows that the data satisfy the normality assumption.

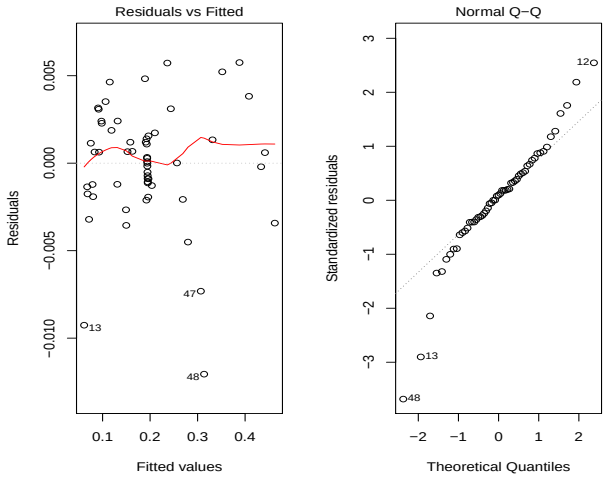


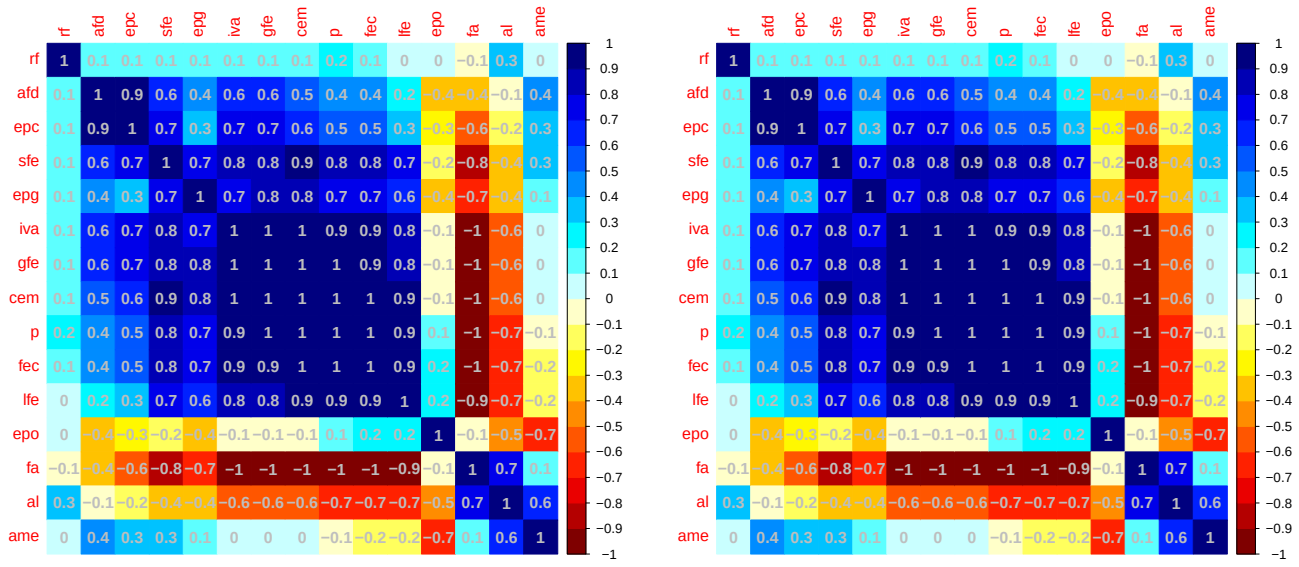
Fig. 2: Regression Diagnostic Plots

Figure 3 represents the Pearson Correlation coefficient of the features in the dataset. Here, for the amount of CO_2 emission, Considering .05 confidence level, 11 predictors are found statistically significant. It is also considered the 0.1 confidence level, and they showed similarities with the .05 confidence level. All the 11 predictor is also important here. Fossil fuel emission(1), gaseous fuel emission(1), industrial investments(1), and the population (1) etc. are highly positively correlated with the data.

On the other hand, the amount of forest lands (-1) is highly negatively correlated with the data. Besides these features, Electricity production from gas(.9), solid fuel burn(.8), etc. also related to the response. The Spearmans Rank Correlation test and the Pearson correlation coefficient are also evaluated for comparison. It is found that both of these tests provide similar correlation values for each predictor. In this specific case, the Pearson correlation coefficient for 11 indicators is more noteworthy than Spearmans Rank Correlation. This similarity and more excellent value in the Pearson test provide more evidence that the data is linear.
As some of the features are insignificant, feature selection is performed to remove these. For that reason, the Akaike Information Criterion (AIC) is followed. AIC follows a stepwise process to select a feature. At different stages, randomly picked variables are incorporated into the stage and reduced. This selection process follows a criterion defined explicitly for this particular model [14].

4.2 Paired T-Test

Paired T-Test result is presented in Table 3. Table 3 depicts a comparison view using linear regression, random forest, and multilayer perceptron(MLP). All of these algorithms were evaluated at the 0.05 confidence level. From this table, it is found that the MAE of linear regression, MLP, and the random forest is 0.0046, 0.0055, and 0.0078, respectively. The mean absolute error suggested that



(a) 0.05 Significance Level

(b) 0.1 Significance Level

Fig. 3: Pearson correlation of the Features (significant features are marked with cyan color to dark blue)

Table 3. : Mean Absolute Error(MAE), Root Mean Square Error(RMSE) and Correlation Coefficient values

Topic	**Linear Regression**	**MLP**	**Random Forest**
MAE	0.0046	0.0055	0.0078
RMSE	0.0064	0.0068	0.012
Correlation Coefficient	0.9985	0.9983	0.9964

linear regression has comparatively less error. RMSE of these three techniques has values 0.0064, 0.0068, and 0.012. Although linear regression and multilayer perceptron have almost the same error, linear regression has less root mean square error.

Moreover, linear regression has 0.9985 for correlation coefficient, whereas the multilayer perceptron and the random forest have a coefficient of 0.9983 and 0.9964 value of 0.9964. It is also suggested linear regression is better than others. In Section 3, it was mentioned that parametric models are more preferable for analysis as those are easier to explain. In this case, although the mean correlation coefficient from linear regression and the random forest is not significantly different, linear regression performs equally considering the correlation coefficient and better considering the other metrics. Thus, Linear Regression using the parametric model is selected for better reasonableness with adequate prediction power.
In Section 3, it is mentioned that linear regression depends on six assumptions. To model a data set into linear regression, all six should be satisfied. Figure 5 shows the residual component plots for each of the features. The difference between the residual line and

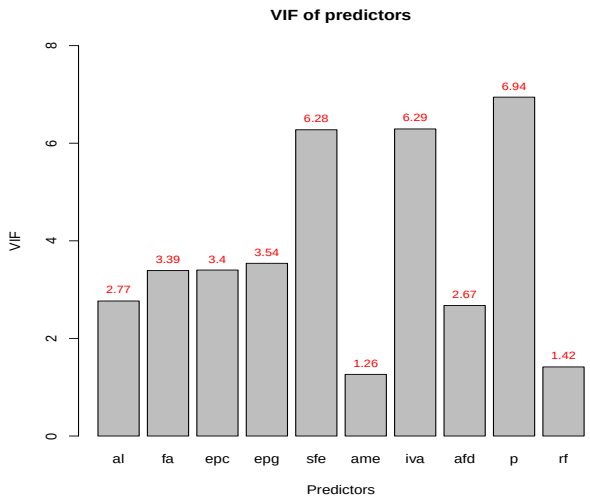


Fig. 4: Variance Inflation Factors of the Predictors

the component line is small for most of those, indicating linearity. However, al, epc, epg, epo, and rf have shown some deviation from the component and residual line. So, these predictors are needed to be transformed for establishing a linear relationship.
Figure 1 shows the autocovariance and autocorrelation functions for residuals time series. The first line represents the interrelation among the residuals. For this reason, it is more significant than others. The successive lags of the residuals are slight and reside within blue dotted but two of those cross blue dotted lines. It does not provide sufficient visualizing patterns among the predictors. Since this

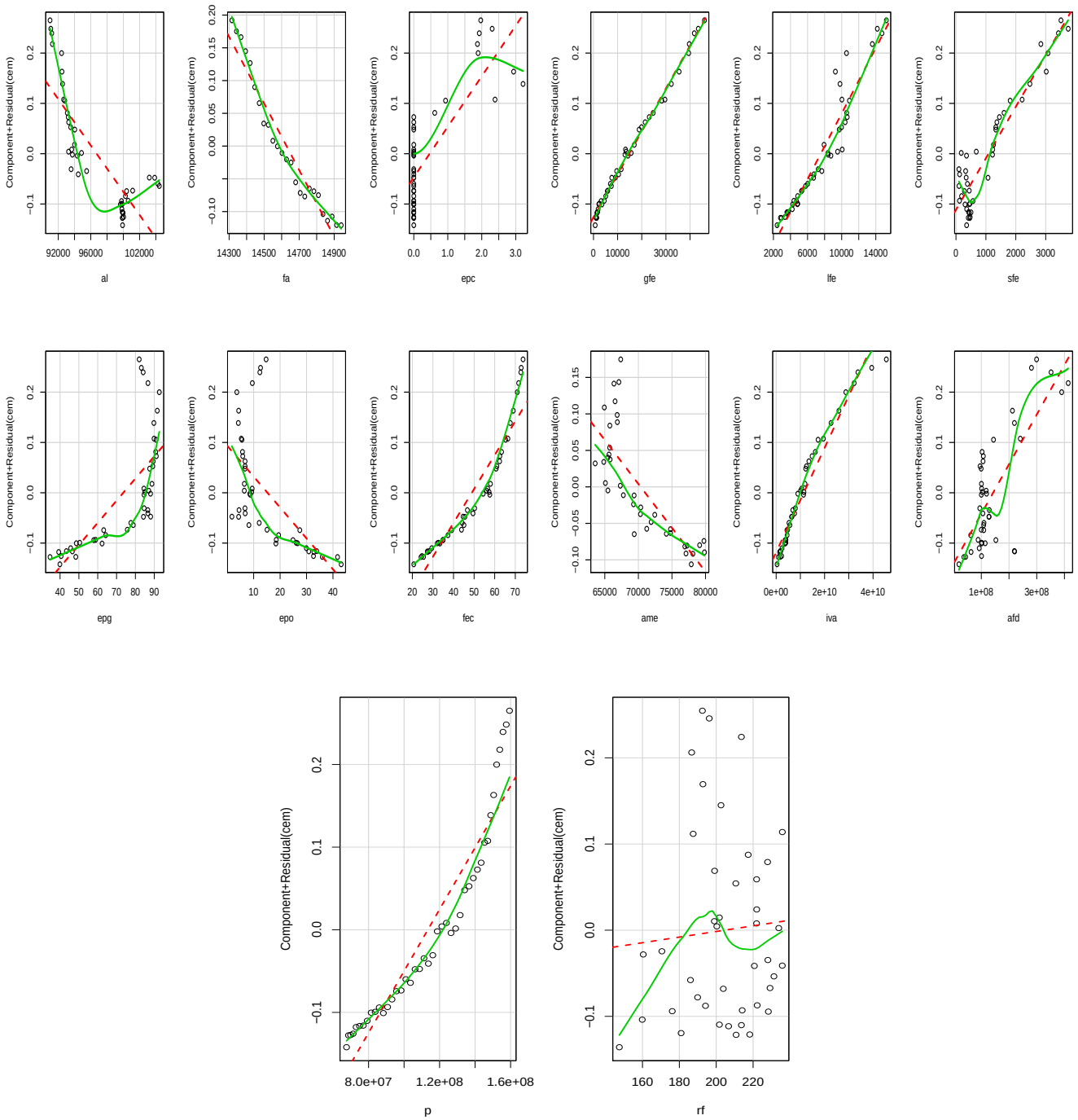


Fig. 5: Component Residual Plots for the Predictors

is a time-series data of 42 years, a pattern might be found on those data.

Figure 2 depicts the residual vs. fitted and regular Q-Q plots. In this figure, the residuals vs. fitted plot do not provide evidence towards heteroscedasticity because it does not show any specific pattern. Standardized residuals closely follow a linear trend in the Normal Q-Q plot. This evidence the normality of residuals. At last, the multicollinearity test is needed to be performed by the variance inflation factor. In the literature, Variance Inflation Factor (VIF) is considered to perform multicollinearity [6]. Multicollinearity is as-

Table 4. : Analysis for Colinear Coefficients (****p <0.001, ***p <0.01, **p <0.05, *p <0.1)

Predictors	Estimate	Std. Error	t value	Pr(>—t—)
(Intercept)	1.500e+00	5.524e-01	2.716	0.00928 ***
al	-4.180e-06	1.243e-06	-3.362	0.00157 ***
fa	-7.546e-05	3.741e-05	-2.017	0.04956 **
epc	-4.262e-03	8.038e-03	-0.530	0.59852
epg	1.764e-03	2.994e-04	5.891	4.22e-07 ****
sfe	6.887e-05	7.337e-06	9.387	2.93e-12 ****
ame	-4.922e-08	2.938e-07	-0.168	0.86768
iva	1.992e-13	4.944e-13	0.403	0.68884
afd	-1.589e-11	4.581e-11	-0.347	0.73026
p	2.253e-11	1.960e-10	0.115	0.90896
rf	1.004e-05	1.532e-04	0.066	0.94802

sumed to be present if VIF is greater than ten [6]. Figure 4 shows that VIF is less than 10 for all cases. This ensures that multicollinearity does not exist in the data. So, all the assumptions of linear regression are satisfied closely or partially.

5. RESULTS AND DISCUSSION

From the earlier discussion, It can be concluded that the linear regression model provides enough insight for this data. For evaluation, it is followed some validation processes. For this reason, the model is fitted into 10-fold cross-validation. The R^2 value and Adjusted R^2 value are 0.9671 and 0.9599. These values provide us enough evidence that it is nicely fitted into a linear model. Moreover, by using 10 and 46 for degrees of freedom, The F-statistic provides 135. The P-value is also evaluated for the F-statistic test to verify the confidence interval is below. These test results are $2.2\text{e}{-12}$, which is significantly below for our model. That is why the null hypothesis for the fitted model and the model fitted without any variables is similar, and considered rejecting it. The residual standard error is 0.0202 on 46 degrees of freedom, which is also low.

Result analysis for the regression is presented in Table 4. The null hypothesis for the T-test is that there are zero relationships between the output and predictors. From table 4, for p values considering confidence level 0.001, 0.01, 0.05 and 0.1, probabilities greater than t is considered. It is seen that the null hypothesis can not be rejected for some of the predictors. At 0.001 confidence level, epg and sfe predictors are strongly linked with the predicted output and positively impact CO_2. It means that the increase in electricity production from gas and increase in CO_2 emissions from solid fuel consumption increase the amount of total carbon dioxide (CO_2) emissions. The signs of their slope estimates indicate that response and predictors are related positively means value increase in predictors increases the response value. In agricultural land, the confidence level of 0.001 suggests that it is significantly related to CO_2 emission. The increasing value of agricultural land decrease the amount of CO_2 emission. Forest area is also related to response with a 0.05 confidence interval. It has the most negative value in the coefficient value. It means more the forest area decrease CO_2 emission will be increased and vice versa. For intercept value, it means without any predictor, response variable means CO_2 will be positive because other factors may have the effect that is not considered. This describes the real-world scenario. So the analysis shown four components are mostly related to CO_2 emissions.

This table 4 provides valuable information about $\mathbf{CO_2}$ emission on the environment. From the *t value* column and *p value* column indicate the impact over $\mathbf{CO_2}$. The features with high significance put a large impact on the environment negatively or positively. Such as, burning solid fuels or electricity production from natural gas has an adverse impact. Their high positive *t value* and small *p value* provide the relationship with CO_2 emission. If this source can be optimized or controlled to some extent, then the emission of CO_2 can be reduced from its linearly growing rate over the years. The forest area(fa) and agricultural lands(al) are also showing a negatively significant value. As forest area or agricultural lands, which consists of trees, consumes CO_2 from the environment, the negative impact of these features means the decrease of these elements from environment cause rise of CO_2. Forest areas and agricultural lands are needed to be preserved to sustain a minimal level of CO_2. The population has a positive *t value*, which means it also affects the CO_2 emission. For the overgrowing population, agricultural lands are reduced to provide accommodations, which results in large CO_2 emissions. To control it, a growing population needs to be controlled. Finally, these feature needs to be maintained as described above to reduce the increment of CO_2 emission.

6. CONCLUSION

This paper presents a carbon dioxide (CO_2) emission model of Bangladesh for analyzing relevant features that have a significant influence on CO_2 emission. It has been shown that considering essential factors, and this model can make a prediction with an acceptable error like 0.0046 in MAE, 0.0064 in RMSE, and 0.9985 in correlation coefficient using a multiparameter linear regression model. The analyzed result has also shown that CO_2 emission is most significantly related to electricity production from natural gas sources, solid fuel consumption, whereas agricultural land and forest depletion saving also significant impact.

7. REFERENCES

[1] Myles R Allen, William J Ingram, and David A Stainforth. Constraints on future changes in climate and the hydrologic cycle. *Nature*, 419(6903):224, 2002.

[2] Bangladesh Road Transport Authority. Registered vehicle data of bangladesh, accessed on: 27 September 2017.

[3] Goodness C. Aye and Prosper Ebruvwiyo Edoja. Effect of economic growth on co2 emission in developing countries: Evidence from a dynamic panel threshold model. *Cogent Economics & Finance*, 5(1), 2017.

[4] World Bank. Bangladesh statistical data, accessed on: 27 September 2017.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] S Chatterjee and B Price. Regression analysis by example. 1977.

[7] Haruna Chiroma, Sameem Abdul-Kareem, Abdullah Khan, Nazri Mohd Nawi, Abdulsalam Yau Gital, Liyana Shuib, Adamu I Abubakar, Muhammad Zubair Rahman, and Tutut Herawan. Global warming: predicting opec carbon dioxide emissions from petroleum consumption using neural network and hybrid cuckoo search algorithm. *PloS one*, 10(8):e0136140, 2015.

[8] Bangladesh Agriculture Research Council. Climate data of bangladesh, accessed on: 27 September 2017.

[9] Glenn De'Ath. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251, 2007.

[10] Kangyin Dong, Xiucheng Dong, and Cong Dong. Determinants of the global and regional co2 emissions: What causes what and where? *Applied Economics*, 51(46):5031–5044, 2019.

[11] Xue Dong, Bin Wang, Ho Lung Yip, and Qing Nian Chan. Co2 emission of electric and gasoline vehicles under various road conditions for china, japan, europe and world averageprediction through year 2040. *Applied Sciences*, 9(11):2295, 2019.

[12] Andy Haines, R Sari Kovats, Diarmid Campbell-Lendrum, and Carlos Corvalán. Climate change and human health: impacts, vulnerability and public health. *Public health*, 120(7):585–596, 2006.

[13] Richard A Houghton. Tropical deforestation as a source of greenhouse gas emissions. *Tropical deforestation and climate change*, 13, 2005.

[14] Joseph B Kadane and Nicole A Lazar. Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290, 2004.

[15] Vincenzo Manzoni, Diego Maniloff, Kristian Kloeckl, and Carlo Ratti. Transportation mode identification and real-time co2 emission estimation using smartphones. *SENSEable City Lab, Massachusetts Institute of Technology, nd*, 2010.

[16] Brian C O'neill, Michael Dalton, Regina Fuchs, Leiwen Jiang, Shonali Pachauri, and Katarina Zigova. Global demographic trends and future carbon emissions. *Proceedings of the National Academy of Sciences*, 107(41):17521–17526, 2010.

[17] Michael A Poole and Patrick N O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, pages 145–158, 1971.

[18] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[19] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[20] Susan Solomon, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingstein. Irreversible climate change due to carbon dioxide emissions. *Proceedings of the national academy of sciences*, 106(6):1704–1709, 2009.

[21] Wei Sun and Jingyi Sun. Prediction of carbon dioxide emissions based on principal component analysis with regularized extreme learning machine: The case of china. *Environmental Engineering Research*, 2017.

[22] Rida Waheed, Dongfeng Chang, Suleman Sarwar, and Wei Chen. Forest, agriculture, renewable energy, and co2 emission. *Journal of Cleaner Production*, 172:4231 – 4238, 2018.

[23] Yang Yu, Yu-ru Deng, and Fei-fan Chen. Impact of population aging and industrial structure on co2 emissions and emissions trend prediction in china. *Atmospheric Pollution Research*, 9(3):446–454, 2018.