CO_2 Emission Prediction Based on Prophet, ARIMA and LSTM

Yiqi Luo*

School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China

*Corresponding author: 22013231@mail.ecust.edu.cn

Abstract. As a matter of fact, with the carbon emissions continuing to rise in recent years, various problems caused by global warming has become more and more serious contemporarily. With this in mind, it is necessary to construct certain models to realize the prediction of the emission. To be specific, this study has come up with an idea to predict the amount of CO_2 emission in the future with Python programs. At the same time, to have a further knowledge of the algorithms used, this study has also compared the reliability of different models. According to the analysis, the results show that CO_2 emissions will keep rising at a steady pace in the future. In addition, the comparisons show that models based on machine learning tend to be more accurate. Overall, the investigations shed light on not only the environmental protection issue but also a future vision of time series forecasting.

Keywords: Time series forecasting; ARIMA; Prophet; LSTM.

1. Introduction

In recent decades, the amount of CO_2 emissions has been rising steadily, and it has already triggered a lot of side effects. CO_2 is one of the major contributing factors of global warming [1]. The rise of carbon emissions has led to changes in agriculture, economy and various other aspects of the lives. Maintaining a good environment would be key to improving the living standards. Therefore, it would be necessary to evaluate the trend of CO_2 emissions and predict the future conditions. The predictions of the future would help us establish better ways to protect the environment in the long run. Researches have made various attempts to predict the CO_2 emission in the future. Support vector machine (SVM) has been applied to predict future carbon emissions. The prediction made by the SVM had an error value of 0.004, which was very close to reality [2]. A group of researchers in Iran also tried to predict the CO2 emission in the agricultural sector of Iran based on Inclusive Multiple Model. In their research, the residuals of MLR4, GPR4, ANN4, and IMM models were compared, and it turned out that the errors of IMM models were smaller than those of the other models [3].

The research conducted is generally aimed at making time series forecasts of the future trend of CO_2 emissions with three kinds algorithms. It is believed that the analysis of the performances from the models would not only provide a valuable insight on the growth trend in CO_2 but it would also act as simulated data in other research involving future CO_2 emission as data basis.

Two conventional ways of forecasting are included, which are the Prophet model and the ARIMA model. Additionally, one tried to take machine learning into practice by applying the LSTM model. Finally, one calculates the mean absolute errors of the predictions given by the three models and compare them so as to find out the time series forecasting model with the highest accuracy. In this essay, the first part is the introduction of the data and methods applied, the second part is the results and discussions of the predictions, and the last part is about the issues one has yet to overcome and the future envisions.

2. Data and Method

The prediction is based on the dataset created by Yoann Boyere and was published on Kaggle (https://www.kaggle.com/) based on the sthe ce data from The WorldInData [4]. The dataset contains the CO_2 emission data in tonnes from 1750 to 2017 sorted by countries, with the name of the country, the Iso code of the country, the year of the emission and the values of CO2 emissions in tonnes respectively recorded in each group of data. One has set the yearly CO_2 emission amount as the only training data.

The autoregressive integrated moving average, known as the ARIMA model, is one of the models that one has applied in the prediction. The ARIMA model is constructed by combining the autoregressive (AR) terms and moving average (MA) terms, and the model also includes three crucial parameters p, d, q as well. The AR part indicates that the variable of the model is regressed on its own lagged value, and the MA part indicates that the error which forms in the regression process is a linear combination of error terms whose values occurred at various times in the past. The combination is expressed by the following equation:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}\epsilon_t + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \cdots + \phi_q\epsilon_{t-q} \tag{1}$$

In this equation, $\beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}\epsilon_t$ is the linear combination of lags of Y up to p lags (AR) and $\phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \cdots + \phi_q\epsilon_{t-q}$ is the linear combination of lags of lags of error up to q lags (MA). The ARIMA model fits the data based on its own past values, including lags and lagged forecast errors. Hence, the model does not work well with seasonal data, but does well when predictors are independent.

The Prophet model is based on business time series modeling including growth, periodic seasonality and holiday events. The model mainly follows the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \tag{2}$$

In this equation, the y is the observed data. The g(t) refers to the sum of growth, including parts of the non-linear growth and the linear growth with changepoints. The s(t) function refers to the seasonal changes, which is modeled with a Fourier series. The h(t) stands for changes caused by various events that take place in holidays, which is an accumulation of changes caused by each event in the holiday. The $\epsilon(t)$ stands for noise [5]. The Prophet model features a unique Simulated Historical Forecast (SHF). The model simulates the past with part of the known data and compares it with the recorded past. The SHF would help identify errors that the simulation has made and would try to minimize the errors when predicting the future. This greatly improves the accuracy of the prediction.

The long short-term memory, known as the LSTM network, is a variant of Recurrent Neural Network (RNN). The model has excellent long-term dependency modeling capabilities and is a popular method for processing sequence data. Compared to regular RNN structure, LSTM model has gated cells for the hidden layers. A typical LSTM unit consists of a cell, an input gate, an output gate and a forget gate. The figure below shows the structure of a typical LSTM hidden layer.

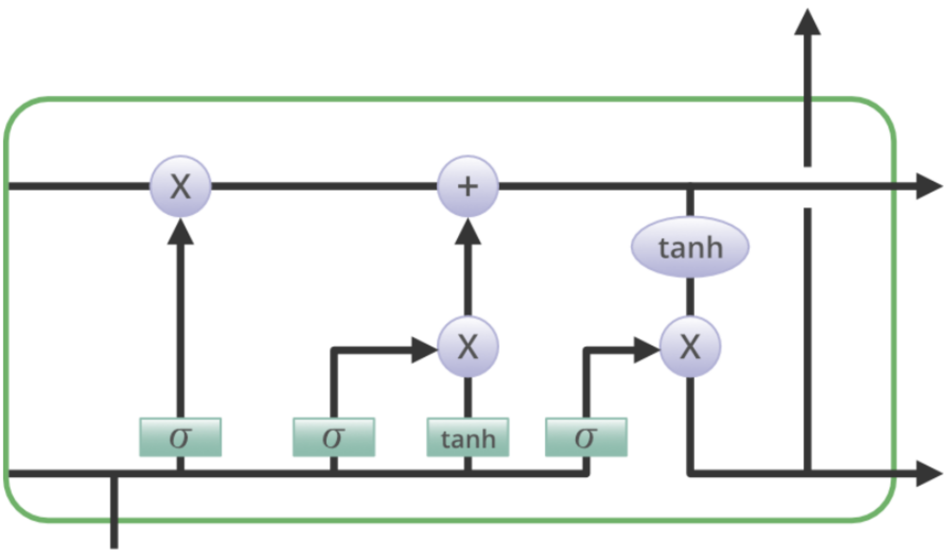

Fig. 1 Structure of LSTM. Three gate cells with sigmoid functions σ controls the information input and output to the tanh layer.

In Fig. 1, a LSTM model has three unique gates named Forget, Input, and Output gate to adjust the features into the layers, which is an additional part compared to conventional RNN models. The gate cells enable the model to store information in a long period of time and control information forwarded or forgotten, avoiding gradient explosion in backtrack process when training RNN models. This makes LSTM a better type of model than conventional RNN and hidden Markov model in terms of sequence learning. Besides being popular in text generation and image processing, LSTM model is also effective in time series forecast, especially with a large dataset [6-8].

To evaluate and compare the errors of different models, the Mean Absolute Error (MAE) of the results of the predictions given by the three algorithms were also respectively calculated [9, 10]. The mathematic formula is mainly expressed as:

$$MAE = (|\text{ predicted } 1 - \text{observation } 1| + \ldots + |\text{ predicted } n - \text{observed } n|) / n \tag{3}$$

3. Results and Discussions

To apply the model into practice, three steps are taken. Firstly, the d of the difference is taken, then the p and q are obtained from the dataset. Afterwards, model fitting and prediction are carried out on the time series data according to p, d and q. The result of the ARIMA model indicates that the amount of CO_2 emission would be on a steady rise in the future, at a rate similar to the speed at which the amount of CO_2 emission grows in the period from 2000 to 2020. The visualization of the result is demonstrated in the Fig. 2.

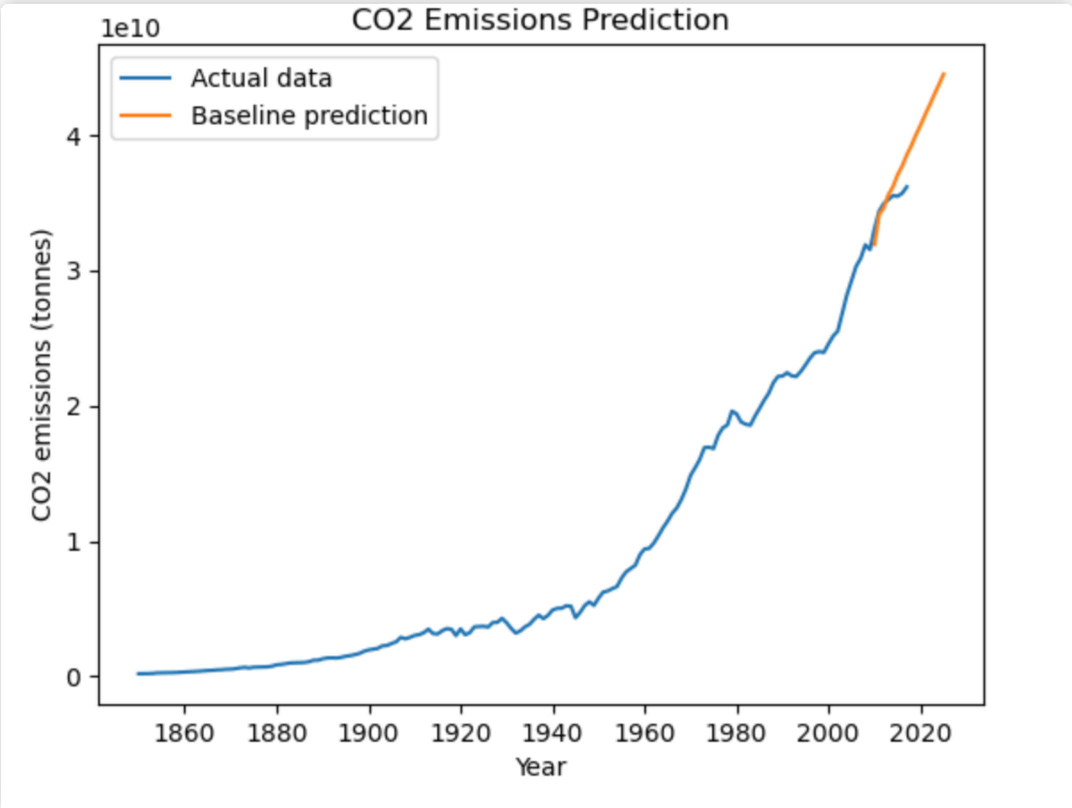

Fig. 2 Prediction by the ARIMA model

In the prediction with the Prophet model, one took factors including the additive model, seasonality and holiday effects into account. The prediction made by the Prophet model indicates that the amount of CO_2 emissions would be on a steady rise in the coming years, and the model predicts that the speed at which the amount of CO_2 emission grows has remained stable since the 1950s. The result is shown in Fig. 3.

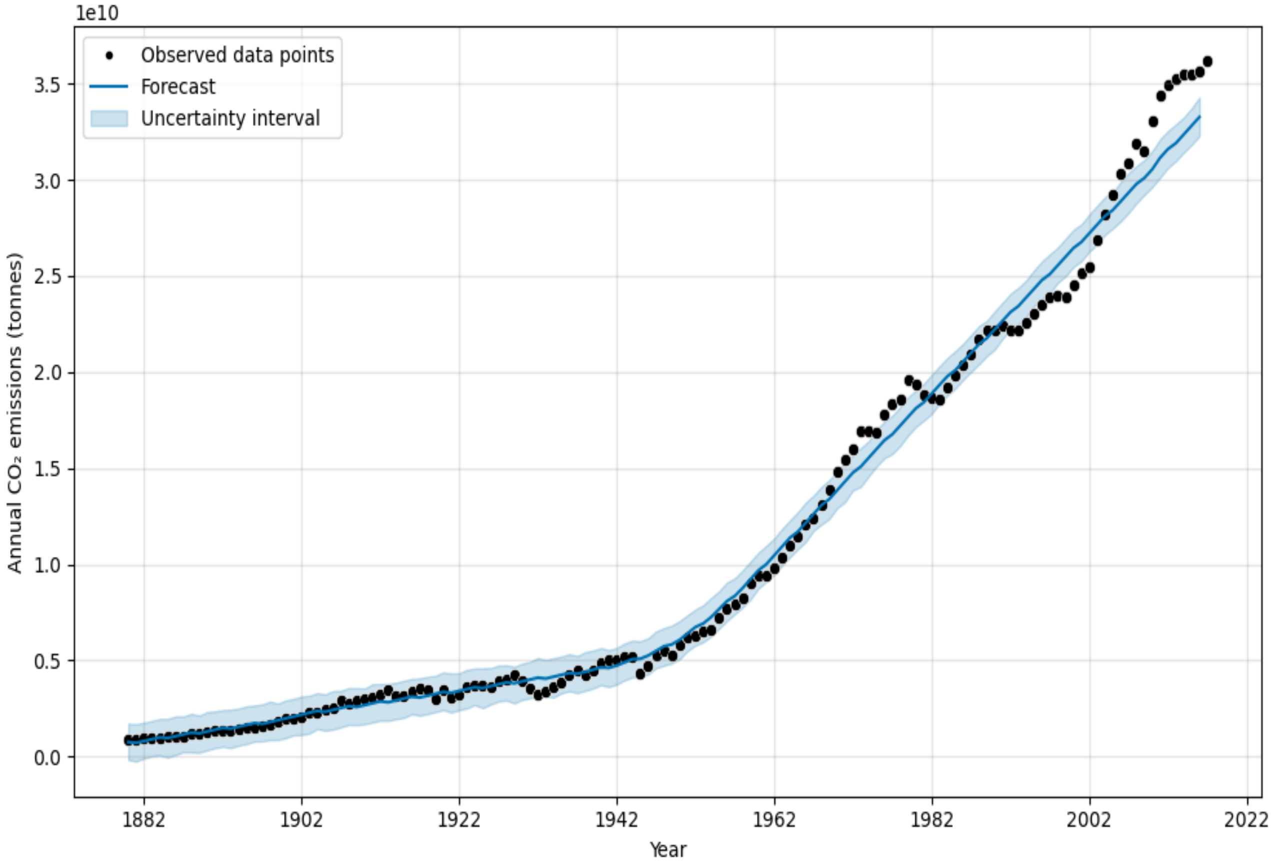

Fig. 3 Prediction made by the Prophet model

To implement the LSTM model, three steps are taken. Firstly, the dataset is divided into various segments. Afterwards, the dataset is converted into a data format that is suitable for the LSTM model. Finally, the LSTM model is built and trained based on the dataset. The prediction made by the LSTM model suggests that the amount of CO_2 will continue to rise in the future, but with a fluctuating velocity. The graph indicates that the speed at which the amount of CO_2 emissions rise tends to be larger than before. The model shows that the amount of CO_2 emissions would dramatically rise after 2000, and there would be a fluctuation in about 2010, which is very close to reality. The detailed prediction given by the LSTM model is demonstrated in the Fig. 4.

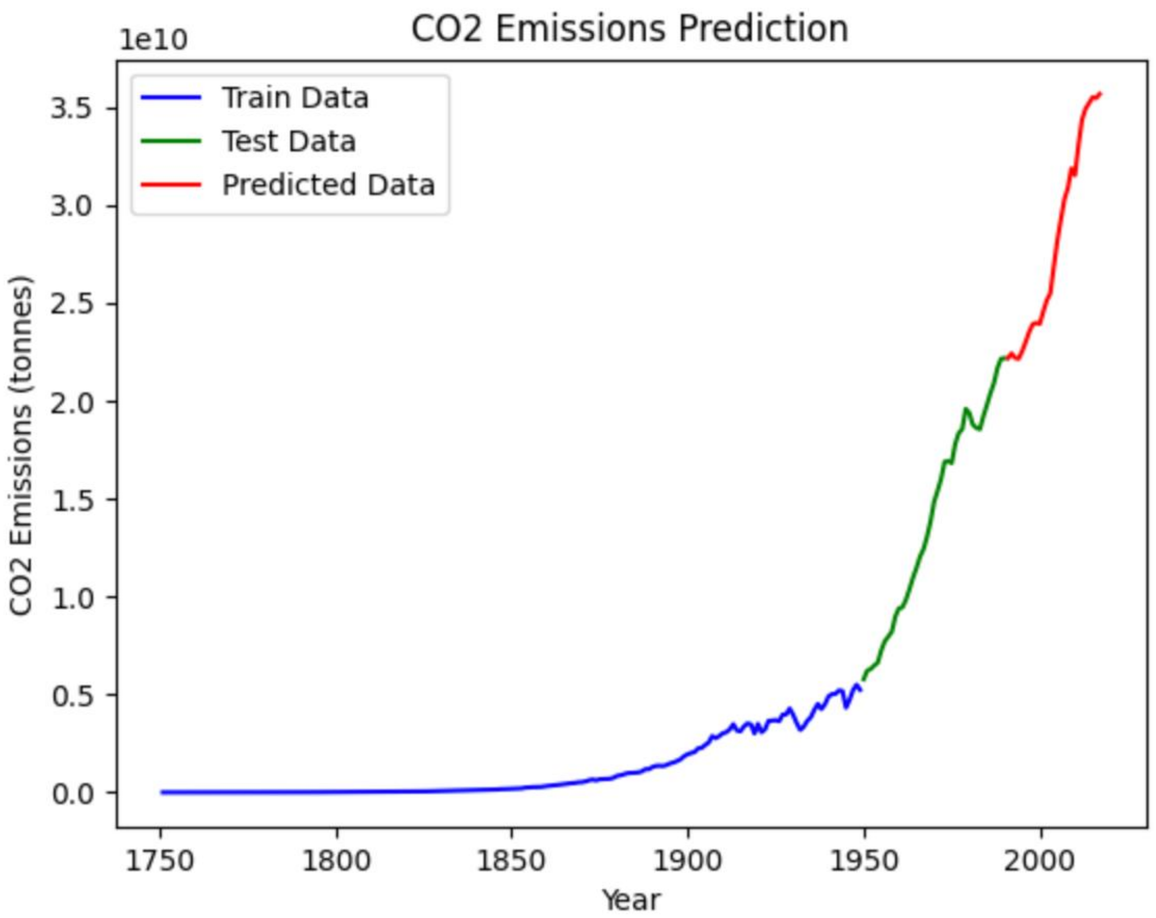

Fig. 4 Prediction made by the LSTM model.

The predictions made by the three models all show a similar trend that the amount of CO2 emission would continue to rise in the future, and the result given by the LSTM model is at first sight more accurate than the predictions given by the other two models. The mean absolute error of the three models are respectively calculated, and the results turn out that the MAE of the ARIMA model is 1.86×10^{10}, the MAE of the LSTM model is 1.26×10^{10}, and the MAE of the Prophet model is 1.39×10^{10}. Compared to the Prophet model and the ARIMA model, the LSTM model has a smaller MAE, which sheds a light on the idea that models based on machine learning tend to be more reliable than other types of models. As the MAEs of all models are in an acceptable spectrum and have relatively small differences, all three models are regarded as suitable models for this dataset. This study has set the yearly CO_2 emission amount as the only training data, which is to some extent different from the real world. In the future, one would try to take more realistic factors into account, such as the promotion of renewable energy and some countries' limitations on CO_2 emissions, which would further optimize the accuracy of the prediction. In addition, as one has observed that models based on machine learning tend to be more reliable than conventional models, one would try to search for machine learning models that are more suitable for the dataset than the LSTM model in the future, which would contribute to better predictions.

4. Conclusion

To sum up, this study has focused on predicting CO_2 emission in the future based on three models, the ARIMA model, the Prophet model and the LSTM model, and this study has compared the reliability of the three models by calculating and comparing their MAEs. The results of the forecasts turn out that the amount of CO2 emission would rise steadily in the future. The results have also indicated that the LSTM model is more accurate than the ARIMA model, which means that time series forecasting models which are based on machine learning tend to be more reliable than traditional time series forecasting models. To further improve the predictions in the future, one plans to add more sets of training data, such as the growing popularity of renewable energy etc. One would also try more types of time series forecasting models and look for better ones.

References

[1] IPCC 2007 Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007.

[2] Saleh C, Dzakiyullah N R, Nugroho J B. Carbon dioxide emission prediction using support vector machine. IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2016, 114(1): 012148.

[3] Shabani E, Hayati B, Pishbahar E, et al. A novel approach to predict CO2 emission in the agriculture sector of Iran based on Inclusive Multiple Model. Journal of Cleaner Production, 2021, 279: 123708.

[4] Kaggle dataset. Retrieved from: https://www.kaggle.com/datasets/yoannboyere/co2-ghg-emissionsdata

[5] Taylor S J, Letham B. Forecasting at scale. The American Statistician, 2018, 72(1): 37-45.

[6] Yang G, Yuan E, Wu W. Predicting the long-term CO2 concentration in classrooms based on the BO–EMD–LSTM model. Building and Environment, 2022, 224: 109568.

[7] Liu T, Bao J, Wang J, et al. A hybrid CNN–LSTM algorithm for online defect recognition of CO2 welding. Sensors, 2018, 18(12): 4369.

[8] Zuo Z, Guo H, Cheng J. An LSTM-STRIPAT model analysis of China's 2030 CO2 emissions peak. Carbon Management, 2020, 11(6): 577-592.

[9] Oyedeji A O, Salami A M, Folorunsho O, et al. Analysis and prediction of student academic performance using machine learning. JITCE (Journal of Information Technology and Computer Engineering), 2020, 4(01): 10-15.

[10] Schwalbert R A, Amado T, Corassa G, et al. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 2020, 284: 107886.