Global CO_2 Emissions Analysis and Prediction [⋆]

Yiqi Xiong, Chidambaram Veerappan, Piwat Wiphanurat, and Oscar Mui

University of Southern California, Los Angeles, CA 90007, USA

Abstract. In this paper, we introduce a system that predicts the CO_2 emission of a country based on its population. Population directly relates to a country's CO_2 emission and its impact on climate change. Here we analyze the data and visualize the impact of population on CO_2 emission and build a model that predicts the emission based on population. We talk about the approach and architecture of the model and our website. The visualization is displayed on the website using Vega and D3 JavaScript. We demonstrate the results of our model and go through the reason of choice for our model while suggesting areas of improvement that can be worked upon for future work.

Keywords: CO_2 emissions · Prediction · XGBoost .

1 Introduction

Growing population and human activities lead to a rapid increase concentration of anthropogenic greenhouse gases, like CO_2 (carbon dioxide), in the atmosphere, which results in climate change and global warming. It is important to visualize and analyze the historical data to understand the trend of CO_2 emission and make strategies to control the emission amounts.

A huge expansion in atmospheric CO_2 concentration breaks the Earth's carbon cycle. Our visualization shows the snowball effect of global CO_2 emission, in which the amount of CO_2 quickly increases year by year. It can raise the awareness of reducing carbon footprint.

In section 2, we present related work in the field to predict the future of greenhouse gas emissions. In section 3, we outline the features used to build our model and metadata information regarding our dataset. In section 4, we explain the approach used in designing our system. In section 5, we detail the tools used to build out the visualization that displays our system. Finally, the remaining sections conclude our model and system findings.

2 Related Work

Kaya et al. [2] have developed a widely used empirical estimate metric called Kaya Identity which is used to estimate the current and future greenhouse gas

[⋆] DSCI554 Project

emission. The calculation involves population and indexes includesGDP and energy efficiency. Skön et al. [4] has proposed a multilayer perceptron neural network to predict indoor CO_2 concentration using measurements of relative humidity and temperature. Zuo et al. [5] used an integrated LSTM-STIRPAT model to predict the peak of CO_2 emission in China using data collected from different regions. The time series analysis is also well-used in predicting CO_2 emission. Gopu et al. [1] implemented ARIMA time series model for air pollution prediction in hyderabad city of India based on the historical data. Pan [3] also applied the XGBoost algorithm to predict the PM2.5 concentration in air, and he stated that the XGBoost model has high accuracy and low over-fitting probability. In this project, we implement the XGBoost regression to predict how the global population can affect CO_2 emission as XGBoost is one of the pioneer models for supervised learning.

3 Data

Table 1. Descriptive statistics.

	count	mean	std	min	25%	median	75%	max
population	226	2.58e+09	1.87e+09	7.45e+08	1.28e+09	1.75e+09	3.25e+09	7.79e+09
CO_2	271	6.26e+03	9.94e+03	9.35e+00	4.95e+01	1.00e+03	6.55e+03	3.67e+04

The data that we used for prediction analysis includes global population and global CO_2 emission from 1750 to 2020, and there are 45 missing population data, and we consider to drop them as they do not have much influence to our model. As year increase, the number of population and total amount of CO_2 emissions are also increase, which can imply there are positive relationship between population and CO_2 .The statistics of the population and CO_2 emission are given in the table (Table 1) above.

For better understanding data, we did following for visualization: Filter out the required data from the dataset if the country was in the world geojson file. Filter out only the 2 features - population and CO_2. Plot various graphs to understand the nature of data visually: Boxplot (Data is right skewed), Histogram (Data is right skewed), Line chart (can tell how rapid increase the population and CO_2 emissions especially from 1900 to 2020) Drop null values, split the data into train and test set for the model.

4 Approach

The design process begins by finding an appropriate dataset and extracting proper covariates to utilize as inputs to our model. For this paper, we chose to use global population as the covariate, and CO_2 as the output. After the proper transformations and cleaning are done to the dataset, the data is used to train our model using a cross-validation methodology. Hyperparameter optimization is performed afterwards, and the final model results are evaluated using a series of 4 metrics.

Appropriate visualizations help readers interpret data. To internalize this, we further build a website containing different visualizations regarding the dataset. In particular, we explore the effects that population has on CO_2 emissions throughout the years. These visualizations are created to be interactive as well as offer a more detailed view into the dataset, such as allowing users to narrow into particular countries or regions, rather than solely a global view.

5 System

The infographic is based on Vega and D3 JavaScript library. Vega is a visualization grammar, a declarative language for creating, saving, and sharing interactive visualization designs. With Vega, we can describe the visual appearance and interactive behavior of a visualization in a JSON format, and generate web-based views using Canvas or SVG. In our case, we specify the visualization grammar called spec and use it in JavaScript file. For D3, D3 is a JavaScript library for manipulating documents based on data. D3 helps us bring data to life using HTML, SVG, and CSS as well as combining powerful visualization components and a data-driven approach to DOM manipulation. We manipulate the DOM by adding Bar Chart along with Map to HTML canvas through SVG.

Our website shows all the infographic including multiple line chart, bar graph, and map (Choropleth map). Multiple line charts display the annual amount of CO_2 emissions segmented by world region along with the annual amount of CO_2 per capital. For bar graph, it exhibits the CO_2 emission of different continents in 2017 in company with many interactions; filtering and sorting by CO_2 amount. Finally, choropleth maps manifest the world population and CO_2 per capital in 2020.

6 Results

We adapt the XGBoost regression model to predict the relationship between the population and CO_2 emissions on the global level. Since XGBoost algorithm has both linear model solver and tree learning algorithms, which gives the characteristics of efficiency, accuracy, and feasibility for implementation.

The dataset has dimension of 2×226 after dropping unnecessary missing values, and it split to 80% of training set and 20% of test set. After splitting to train and test set, the 5-fold grid search cross-validation with evaluation metric RMSE is used to resampling the train set and avoid the overfitting and selection bias issue. The best hyperparameter of XGBoost also has been generated from cross-validation.

There are 4 metrics to measure model prediction performance by comparing actual data and model fitted data from the test set: the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2), and Accuracy.The XGBoost regression model performances are also give in Table 2. The R^2 shows the model is well performed. The model can explain 99% of the dataset. Since the dataset is not standardized, the RMSE and MAE are relatively high. Fig. 1. of fitted values and actual values can also show the prediction result is quite good and generally follow the pattern of actual values.

Table 2. Model Evaluation Metrics.

Metrics	Values
R^2	0.9905
RMSE	1014.9337
MAE	539.1847
Accuracy	79.8384

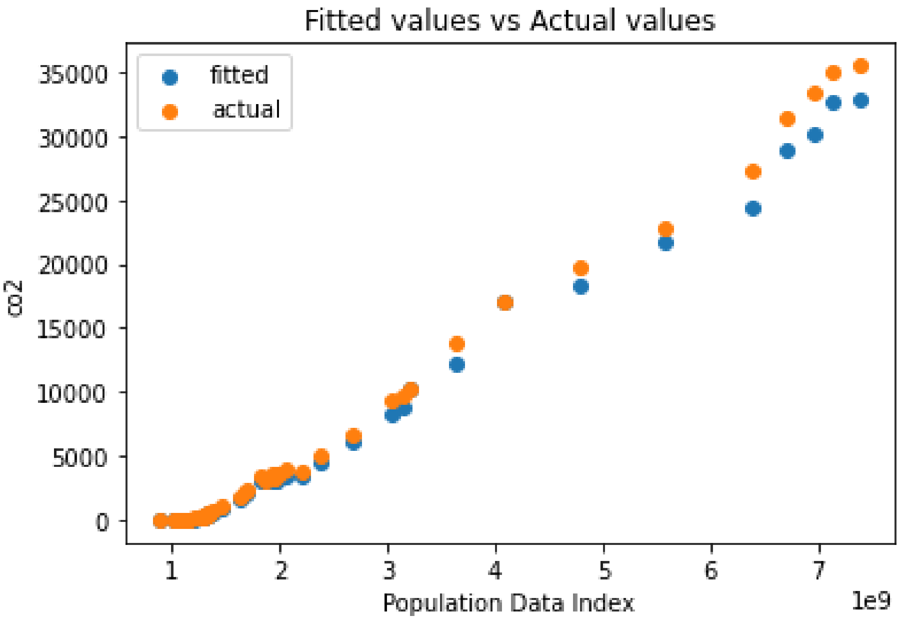

Fig. 1. Comparing true CO_2 data and model predicted data

7 Conclusion

We find global population to be a potent predictor in the rise of CO_2 emissions with global population explaining 99.05% of the variability in the model. Since global population is highly associated with GDP and energy usage, our findings are consistent with related empirical estimates that others have found in their work. Along with an interactive website that displays regional CO_2 emissions, our results offer a platform for users to explore and be increasingly conscious of their individual contributions to global warming.

Future considerations for our work include using React as a framework to implement front-end components along with D3, considering additional covariates in our model, incorporating model results into the website, and exploring additional frameworks to display data.

References

1. Gopu, P., Panda, R.R., Nagwani, N.K.: Time series analysis using arima model for air pollution prediction in hyderabad city of india. In: Soft Computing and Signal Processing, pp. 47–56. Springer (2021)
2. Kaya, Y.: Impact of carbon dioxide emission control on gnp growth: interpretation of proposed scenarios. Intergovernmental Panel on Climate Change/Response Strategies Working Group, May (1989)
3. Pan, B.: Application of xgboost algorithm in hourly pm2. 5 concentration prediction. In: IOP conference series: earth and environmental science. vol. 113, p. 012127. IOP publishing (2018)
4. Skön, J., Johansson, M., Raatikainen, M., Leiviskä, K., Kolehmainen, M.: Modelling indoor air carbon dioxide (co2) concentration using neural network. methods **14**(15), 16 (2012)
5. Zuo, Z., Guo, H., Cheng, J.: An lstm-stripat model analysis of china's 2030 co2 emissions peak. Carbon Management **11**(6), 577–592 (2020)