Predicting and Forecasting Carbon Emissions with Machine Learning

Rose Zhang, Sherry You

April 17th, 2023

1 Introduction

Climate change has been a global concern for decades. Not only does it have devastating impacts on the environment such as causing floods and droughts, it also threatens to displace millions into poverty, expand inequalities, limit economic growth, and much more. (United Nations, n.d.) The United Nations reports that, as of 2021, the global mean temperature was about 1.1°C above the pre-industrial level. (United Nations, n.d.) The Paris Agreement was established in 2015 and adopted internationally in order to band nations together to curb the effects of climate change, stressing the need to limit global warming to 1.5°C above the pre-industrial level. ("The Paris Agreement", n.d.) Given the current trends in rising temperatures, however, the world must substantially increase its efforts in reducing carbon emissions to meet this goal. ("The Paris Agreement", n.d.)

This paper develops a model to predict and forecast carbon dioxide emissions (CO2), which are a direct cause of climate change. Answering this prediction problem will not only help forecast future emission levels, but can help identify ways in which governments can pursue more effective paths to mitigate the effects of climate change. We build our model based on a wide range of economic and environmental datasets at the country level, including but not limited to measures of population, GDP, land area, and agricultural production. We first used our data to construct random forest (RF) models, and our model with optimized tuning parameters had a test performance of approximately 81%. Furthermore, we identified GDP, land area, meat production (chicken, beef, pig), and cereal production to be the main drivers of CO2 emissions. Then, we built a machine learning model using recurrent neural networks (RNN) in order to forecast emissions. Of the 4 different models tested, the single LSTM layer model with dropout had the best prediction performance with a mean absolute error of 7.753.

2 Literature Review

The literature regarding this topic is quite expansive, and there are many economic variables that have been investigated to be potential influencing factors of CO2 emissions. For example, Musah et al. (2020) study the predictors of carbon emissions for NAFTA countries and find economic growth to be a positive predictor of carbon emissions, while population and foreign direct investments had no significant effects. Ito (2017) investigates a panel of 42 developed countries over the period 2002–2011 and finds that consumption of fossil fuel energy and GDP lead to an increase in CO2 emissions, whereas renewable energy consumption is associated with a decrease in emissions. Ren et al. (2014) investigate carbon emissions in China and find that foreign direct investment was one of the key contributors to China's rapidly rising CO2 emissions.

It is only in recent years that machine learning methods have become more popular in the literature. For example, Sun and Ren (2021) used emissions data from China to develop a model for the short-term prediction of carbon emissions, using ensemble empirical mode decomposition (EEMD) and the backpropagation neural network based on particle swarm optimization (PSOBP). Sun and Huang (2022) collected emissions panel data from 2000-2016 on China and its provinces to construct a model to predict carbon emission intensity based on influencing factors. They use the newly developed whale optimization algorithm (WOA) to optimize an extreme learning machine to build their prediction model. They find that the level of economic development, industrial structure, urbanization level, and government intervention all decrease carbon emission intensity.

Finally, Niu et. al (2020) use a general regression neural network (GRNN) forecasting model based on improved fireworks algorithm (IFWA) optimization to forecast total carbon emissions and carbon emissions intensity in China. They use a series of technological and economical influencing factors in their forecasting model. To reduce the redundancy in the data used for the model, they use a random forest (RF) model to select only the most important factors before training their neural network.

Our research contributes to the literature by using influencing factors in addition to past emissions data to forecast CO2 emissions. We also examine data on a global level, looking at how country level characteristics can predict emission levels rather than simply using data from one country.

3 Model

Our target variable of interest for our models is the per capita carbon emissions for a specific country in a specific year. The features, or input variables for our model, are a variety of economic and environmental variables that we specify in the Data section. We have two goals for this paper: (1) to predict per capita carbon emissions given these country specific input variables, and (2) to forecast per capita carbon emissions for a country at some point in time in the future.

3.1 Random Forest Model

We first use a Random Forest (RF) model (Breiman 2001) to predict CO2 emissions given data on several influencing factors. An RF model is an ensemble of decision trees, which are prediction algorithms known for their abilities to handle mixed data types and missing values, be scale independent, and are robust to outliers and irrelevant inputs. Decision trees predict target values by partitioning the input space recursively into increasingly homogenous groups, where the partitions are created by splitting on an input variable (i.e. conducting a binary test). Since simple decision trees use the best variable it can at each split, other typical ensemble methods like bootstrap aggregation tend to produce similar, correlated trees as they consider all input variables. RF models on the other hand only consider a random subset of input variables for each split, resulting in an ensemble of trees that have splits on different variables. This generates an ensemble of unbiased, decorrelated decision trees that results in a model with reduced variance. RF models also allow us to investigate feature importance; we will be able to determine which features are the most influential in predicting CO2 emissions.

3.2 Long Short-Term Memory Model

To address our second question of forecasting CO2 emissions, we settled on the Long Short-Term Memory (LSTM) model, which we believed would be a better fit than the RF model. The LSTM model is a type of recurrent neural network (RNN). Artificial neural networks are able to handle relationships between variables that are very complex and nonlinear, which is suitable for our case here. In contrast to those of a traditional neural network, the hidden layers of a RNN depend on inputs from the current time and inputs passed from the previous time point. Thus, RNNs are specifically optimized for sequential data, and are typically used for language processing and time-series forecasts. However, since states of the hidden units in RNN models are updated recursively, key information from previous time points can gradually disappear during forward propagation

and back propagation. Thus, these models tend to struggle with vanishing gradients and long-term dependencies.

The LSTM model, developed by Hochreiter and Schmidhuber (1997) is capable of handling both of these problems. It incorporates memory cells that connect to themselves at the next time point with a weight of one, effectively copying their states over and 'remembering' past inputs. This self-connection is also gated by other units that selectively determine what to to forget and remember from the current state that will be passed onto the next time point, establishing a network that will only 'remember' important information for prediction.

4 Data

We collected all our data from the website Our World in Data. Our target variable of interest is the amount of carbon emissions for a specific country in a specific year measured in millions of tonnes. For our independent variables, we refer to the existing literature on influencing factors of CO2 emissions and merge together a variety of economic and environmental datasets. Our final list of independent variables is below:

- Population
- GDP
- GDP Per Capita
- Trade (
- GDP per capita growth (annual
- Fruit Consumption
- Vegetable Consumption
- Agricultural Land Area
- Land area (sq. km)
- Population density
- Population Growth Rate Estimates
- Rice Production
- Apple Production
- Avocado Production
- Banana Production
- Barley Production
- Bean Production
- Cassava Production
- Cocoa Production
- Coffee Production
- Pig Production
- Chicken Production
- Beef Production
- Milk Production
- Seafood Production
- Aquaculture Production
- Cereal Production

Before merging together the datasets, we created some graphs to get a better understanding of our dataset. In our first graph, we plotted the carbon emissions of each country over time. From this plot, we found that a majority of the countries have less than 2000 million tonnes of CO2 emissions by 2018. We then plotted just the top five countries with the most carbon emissions and we could see that by 2018, China had almost doubled the amount of carbon emissions of the second leading country, with approximately 10,000 million tonnes of CO2. In our last graph, we plotted the total carbon emissions by year and we can also see the rapidly increasing amount of emissions as time passes. From these graphs, we noted that there are some drastic differences in carbon emissions between countries and over time, which may pose as a challenge for our models.

After merging together the datasets above, we needed to select a subset of the data that can minimize the null values. Variables such as "Palm Oil Area Harvested", "Forest Area" and a few others had over half of their values recorded as missing. Therefore, we removed these variables from our analysis. Our final dataset only included data for the years after 1960, as this was the subset that we believe minimized the number of null values without overly decreasing the number of records. Our final dataset had 9091 records with 31 variables.

Once we established our final dataset, we began to prepare the data for our chosen models. First, we split 75% of our data into the training set and left 25% for our test set. Since we have data recorded for each country over a span of years, we created a 75-25 split for each country, then merged the training/test sets for each country together to get our full train and test sets. After splitting our data, we normalized the values with min-max scaling as we believe our data does not follow a Gaussian distribution. Finally, we decided to fill the null values for variables related to production with 0 as we assumed the null values indicated no production for the country at this time. The remaining null values in the dataset were filled with the mean of the column.

As our goal was to forecast carbon emissions using multivariate time-series panel data, we needed to lag the variables before running our RNN models. Since we are working with annual data, we only lagged our variables by one year so that we do not lose any degrees of freedom. (Adeleye 2018) Before inputting our data into our RNN models, we also needed to reshape our data into a three-dimensional array in the form of [samples, timesteps, features]. We decided that we will be using the previous years data to help form predictions for the current year, therefore the value for timesteps for our new shape is 1.

5 Estimation

For our analysis, we decided to run six different models including one random forest and five variations of the RNN model. For our random forest model, we used a RandomForestRegressor with 25 estimators and a max depth of 10. After tuning the parameters with grid search, we found that the optimal model should include 40 estimators, a max depth of 15, and max features set to 'log2'.

The first RNN model we implemented was a simple RNN model with a single SimpleRNN layer between the input and output layers. This model can be used as a baseline comparison for the other LSTM models. The next two models were LSTM models with a single LSTM layer, however, the second model also includes one Dropout layer. This Dropout layer randomly drops a number of the layer's output nodes which can help prevent overfitting in the model. Finally, we tried

implementing two stacked LSTM models with two LSTM layers, adding Dropout layers to the second model. By stacking the LSTM layers, we are increasing the depth and abstraction of the model, allowing the model to tackle more complex sequence predictions. (Brownlee 2017) As we are connecting an LSTM layer to another LSTM layer, we set the return_sequences parameter to True in the first LSTM in order to return a 3D output sequence to the following LSTM layer.

Typically the number of nodes in an LSTM layer range from 32 to 128, however we found that when we increased the number of nodes to 128, the model had a slightly better predictive ability. For the activation function in our LSTM layers, we decided to implement ReLu as we found that there seems to be a larger difference in the training and testing errors when we use Sigmoid, implying potential overfitting. Our output layer is a Dense layer with one unit, returning a single prediction for carbon emissions.

When compiling our model, we set our loss function to the mean absolute error function so that our error will be on the same scale as our target variable. This in turn will provide better interpretability for our results. (Allwright 2022) We used Adam for our optimizer as it is able to converge to the minimum quickly while following the shortest route. We will be using mean absolute error as the metric to evaluate our models as well.

When fitting each model, we set the number of epochs to 50 with a batch size of 64 samples. We found that when we increased the batch size to 80, the model performance decreased as it used more samples per update, which in turn decreases the models ability to generalize. However, when we tried running the models with much smaller batch sizes of 15 or 30, the resulting validation loss had much more noise and took more time to converge. As a result, we decided to use 64 as our batch size as the model is able to generalize without much added noise or convergence time.

6 Results

Our first RF model, using a maximum of 25 trees and maximum depth of 10, had a training accuracy of 1.00 and a test accuracy of 0.76. We can see that this is a standard case of overfitting, as the model performs perfectly on the training data but only at 76% accuracy for the test. After optimizing the parameters of our model using grid-search 5-fold cross validation, our new RF model performs better with a test accuracy of 0.81. Training accuracy is still at 1.00, so our model is still overfitting the data slightly, but it has decreased in bias and variance with the new RF model. Through this optimized RF model we can also identify important features in the prediction of CO2 emissions. In our visualization of important features, we can see that GDP, followed by land area, and then meat (chicken, beef, pig) and cereal production are the most important features in predicting CO2 emissions. Intuitively, these are reasonable results. Countries with higher GDP have higher amounts of economic activity, which likely consists of activities that require a high use of energy such as manufacturing and production, which will emit CO2 into the atmosphere. Similarly, meat production and cereal production require a lot of land and machinery to operate, which also demands the use of a lot of energy depending on the burning of fossil fuels, for example. Land area can also be an indication of a country's potential for agricultural production and economic activity.

We implement 4 different RNN models. First, the simple RNN model had a mean absolute error of 7.8189 million tonnes, which is just under 7% of the mean CO2 emission levels (113.4 million

tonnes) in our data. Our single layer LSTM model performed slightly better, as expected, with a mean absolute error of 7.7572. Adding a dropout layer to this model improved its performance ever so slightly to an error of 7.753, making this our best performing RNN forecasting model. We also tested out a model that stacked two LSTM layers, resulting in a model with error 7.7988, and including a dropout layer actually increased the error to 8.0267.

Ultimately, the single layer LSTM model with a dropout layer had the best performance. This makes sense, as including more layers increases the complexity of the model and makes it harder to train. Given our rather low sample size, our model would have a harder time learning more complex relationships between the variables.

7 Conclusion

In our research, we have determined through our RF model that GDP, land area, and meat and cereal production are the main influencing factors of CO2 emissions in a given country. This RF model was also able to predict emissions based on country characteristics at 81% accuracy on our test sample. Furthermore, our best performing LSTM model was able to forecast CO2 emissions given certain influencing factors with a mean absolute error of 7.753. This performance is decent, with the error being just number 7% of our sample's mean CO2 emission levels.

A main limitation of our research is the small sample size of around 9000 observations. Deep learning typically requires a lot of data to ensure that the model is trained properly, particularly for more complex models, so our small sample size could be a reason why our models are not incredibly accurate in their prediction performance. Our research can be extended by training these models on larger datasets to improve their performance. It would also be interesting to include other economic variables as inputs, such as some measure of trade, exports, innovation, etc. as these features have also been found in the literature to have a significant influence in predicting CO2 emissions.

8 References

Adeleye, Bosede Ngozi. "Time Series Analysis (Lecture 2): Choosing Optimal Lags in EViews." CrunchEconometrix, February 12, 2018. https://cruncheconometrix.blogspot.com/2018/02/time-series-analysis-lecture-2-choosing.html.

Allwright, Stephen. "MSE vs Mae, Which Is the Better Regression Metric?" Stephen Allwright, July 7, 2022. https://stephenallwright.com/mse-vs-mae/. Breiman, Leo. "Random Forests." Machine Learning 45, no. 1 (2001): 5–32. https://doi.org/10.1023/a:1010933404324.

Brownlee, Jason. "Stacked Long Short-Term Memory Networks." MachineLearningMastery.com, August 18, 2017. https://machinelearningmastery.com/stacked-long-short-term-memory-networks/.

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." Neural Computation 9, no. 8 (1997): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Ito, Katsuya. "CO2 Emissions, Renewable and Non-Renewable Energy Consumption, and Economic Growth: Evidence from Panel Data for Developing Countries." International Economics 151 (2017): 1–6. https://doi.org/10.1016/j.inteco.2017.02.001.

Musah, Mohammed, Yusheng Kong, and Xuan Vinh Vo. "Predictors of Carbon Emissions: An Empirical Evidence from NAFTA Countries." Environmental Science and Pollution Research 28, no. 9 (2020): 11205–23. https://doi.org/10.1007/s11356-020-11197-x.

Niu, Dongxiao, Keke Wang, Jing Wu, Lijie Sun, Yi Liang, Xiaomin Xu, and Xiaolong Yang. "Can China Achieve Its 2030 Carbon Emissions Commitment? Scenario Analysis Based on an Improved General Regression Neural Network." Journal of Cleaner Production 243 (2020): 118558. https://doi.org/10.1016/j.jclepro.2019.118558.

Our World In Data. Accessed March 21, 2023. https://ourworldindata.org/.

Ren, Shenggang, Baolong Yuan, Xie Ma, and Xiaohong Chen. "International Trade, FDI (Foreign Direct Investment) and Embodied CO2 Emissions: A Case Study of Chinas Industrial Sectors." China Economic Review 28 (2014): 123–34. https://doi.org/10.1016/j.chieco.2014.01.003.

Sun, Wei, and Chenchen Huang. "Predictions of Carbon Emission Intensity Based on Factor Analysis and an Improved Extreme Learning Machine from the Perspective of Carbon Emission Efficiency." Journal of Cleaner Production 338 (2022): 130414. https://doi.org/10.1016/j.jclepro.2022.130414.

Sun, Wei, and Chumeng Ren. "Short-Term Prediction of Carbon Emissions Based on the EEMD-PSOBP Model." Environmental Science and Pollution Research 28, no. 40 (2021): 56580–94. https://doi.org/10.1007/s11356-021-14591-1.

"The Paris Agreement." United Nations Framework Convention on Climate Change. Accessed March 21, 2023. https://unfccc.int/process-and-meetings/the-paris-agreement.

United Nations. "Climate Change - United Nations Sustainable Development." United Nations. Accessed March 21, 2023. https://www.un.org/sustainabledevelopment/climate-change/.