Machine Learning to Predict Energy Usage Outcomes

James Zhang
Sir Winston Churchill Highschool

Abstract

How realistic are the sustainability standards that we impose on countries? Using data from countries all around the world from 2000-2020, this paper answers this question. Using a regression model, we predicted the renewable energy share in the total final energy consumption and CO2 emissions using various different factors that impact a country. The model has an r-squared value of 0.94 for renewable energy share in the total final energy consumption and 0.99 for CO2 emissions. The developed model has the potential to be used as a tool to achieve realistic progress towards sustainable development.

Introduction

According to the UN, sustainable development is one of the most prevalent issues being discussed today[1]. Sustainable development is an approach to development that sustainably uses environmental resources. Fearing further destruction caused by global warming, international organizations like the UN have set up sustainability goals. Many people in Western countries complain about these goals because many of the countries that they cannot control such as China contribute far more pollution than they do. According to the World Population Review, China had the most CO2 emissions by far, with almost triple the emissions of the second-place country[2]. However, how plausible are these environmental goals in countries that are less developed? This research aims to determine what environmental goals are plausible for countries based on many different qualities. First, a suitable dataset containing various information about most countries in the world was found. Next, irrelevant data was

removed, missing data was filled in, and the dataset was normalized to make the data suitable for a regression model. Finally, the model outputs its prediction of the renewable energy share in the total final energy consumption and CO2 emissions based on its r-squared value.

Methodology

Dataset

The data set used for this project is a Kaggle dataset that included data from most countries in the world from 2000-2020 [3]. The dataset was compiled mostly using data from the World Bank and the International Energy Agency. The dataset had many empty cells because some of the data was unavailable for certain countries. When the empty cell is in the columns that we predict, the entire row is deleted. When the empty cell appears elsewhere, it is treated as a zero. The entire dataset is then normalized in order for the model to produce more accurate results. Ninety percent of the data is used to train the model and the remaining ten percent is used to test the model.

Model

The prediction model used for this project is an MLP regressor model. We ran experiments prediction column 4 and column 11. By creating loss graphs for various values of learning rate, hidden layer size, and batch size, we estimated optimal values for each of these hyperparameters. Using Figures 1 and 2 shown below, I determined that the best batch size is around 80.

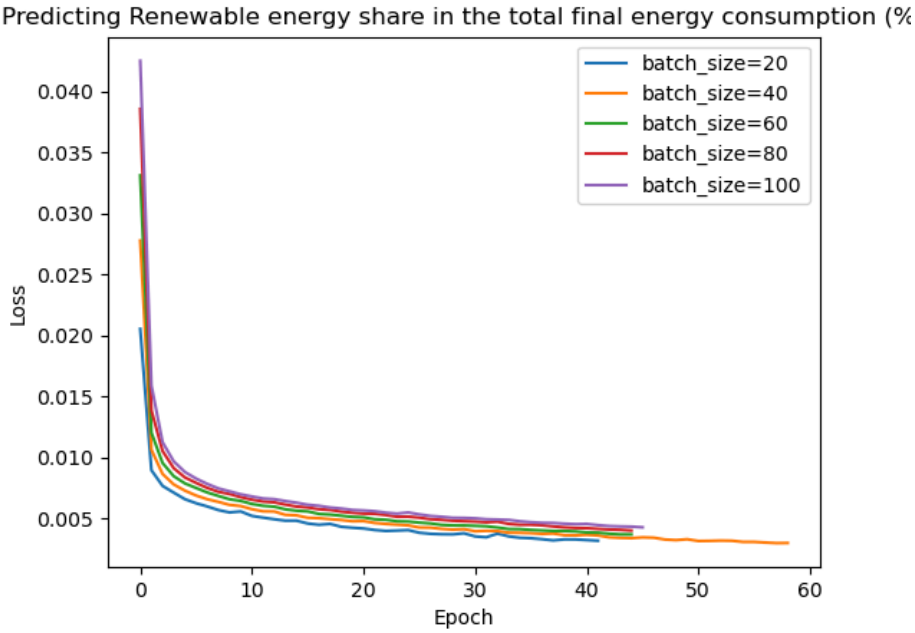

Figure 1: Loss vs Epoch Graph of the Model Predicting Renewable energy share in the total final energy consumption while trying different batch size values

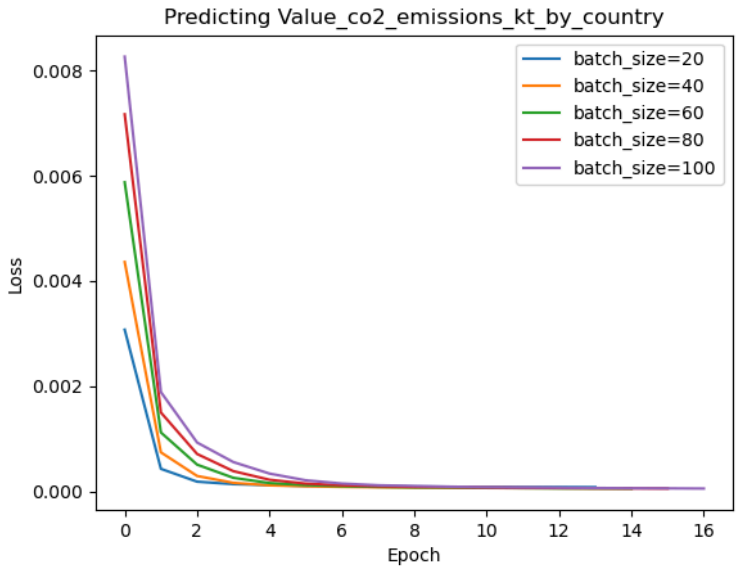

Figure 2: Loss vs Epoch Graph of the Model Predicting CO2 emissions while trying different batch size values

Using Figures 3 and 4, we determined that the best-hidden layer size is around (10000,).

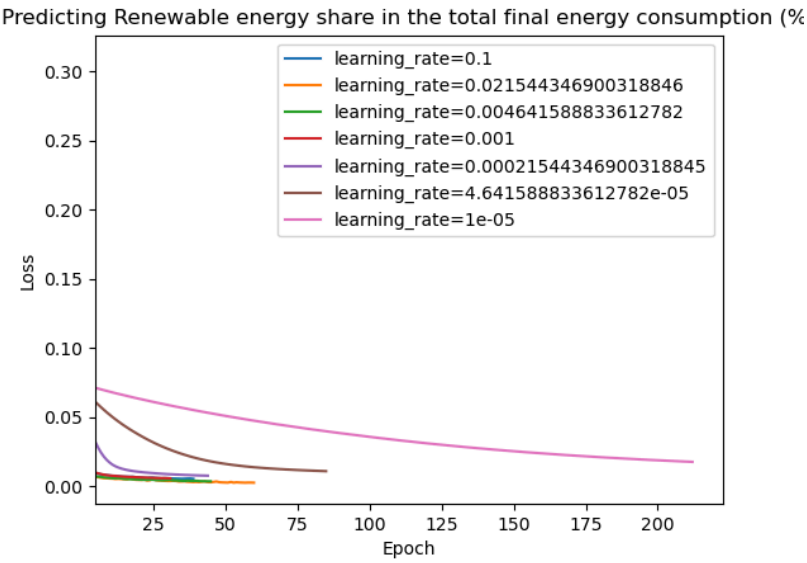

Figure 3: Loss vs Epoch Graph of the Model Predicting Renewable energy share in the total final energy consumption while trying different learning rate values

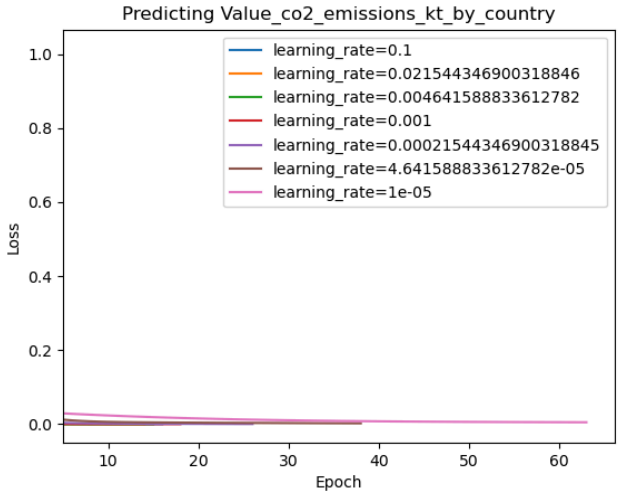

Figure 4: Loss vs Epoch Graph of the Model Predicting CO2 emissions while trying different learning rates.

Using Figures 5 and 6, we determined that the best learning rate is around 0.00681292.

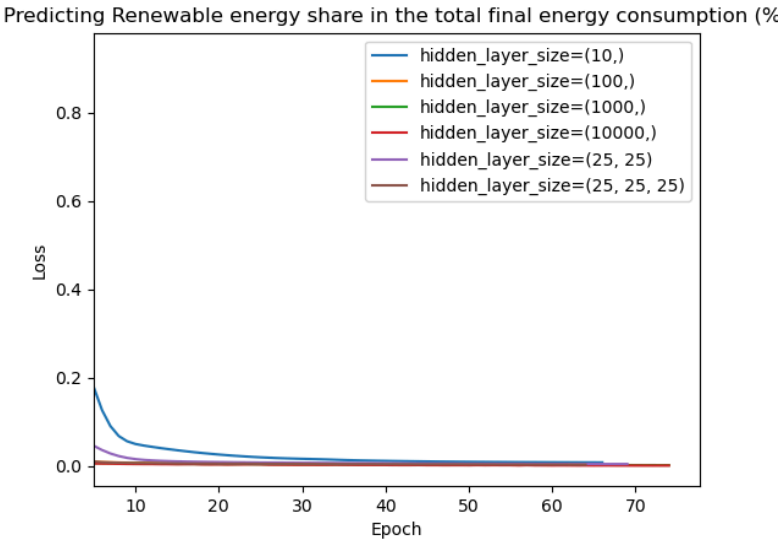

Figure 5: Loss vs Epoch Graph of the Model Predicting Renewable energy share in the total final energy consumption while trying different hidden layer size values

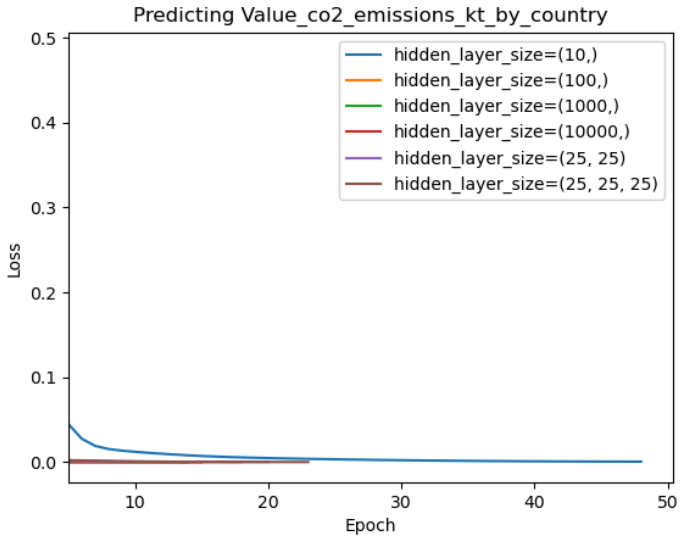

Figure 6: Loss vs Epoch Graph of the Model Predicting CO2 emissions while trying different hidden layer sizes

Results

The r-squared value is a way of measuring how accurate a regression model is. When the r-squared value is equal to the maximum value of 1, which means that the model perfectly matches the data. The closer the r-squared value is to 1 the more accurate the model is. The r-squared values for the normalized and unnormalized predictions have a very minimal difference, meaning that the model can actually be useful for making predictions. The r-squared value ended up being 0.94 for predictions for renewable energy share in the total final energy consumption and 0.99 for predictions for CO2 emissions. However, there are some outliers where negative numbers appear where they should not. This program can be used for many other datasets in order to solve other types of problems.

Limitations

The model can only predict realistic environmental goals based on known and predictable factors. It cannot take into account any unforeseen events or obstacles that are bound to arise in the real world. The model can also only take into account data that works in a regression model, limiting the factors that the model can consider.

Conclusion

This model helps determine realistic sustainability goals countries can set for themselves. Beyond that, it can help ease the feeling of discontent among those unhappy with the environmental habits of some developing countries by informing them that their expectations for those countries are not feasible for poorer countries to meet. The data used to train the model

needs to be updated if it is to be used in the future as the model does not use the most recent data.

References

1. United Nations. (n.d.). *Global issues*. United Nations. https://www.un.org/en/global-issues
2. *Pollution by Country 2024*. Pollution by country 2024. (2024a). https://worldpopulationreview.com/country-rankings/pollution-by-country
3. Tanwar, A. (2023, August 19). Global Data on Sustainable Energy (2000-2020). Kaggle. https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy