Machine learning to predict sustainable energy usage outcomes

James Zhang

Abstract

This study evaluates the practicality of global sustainability targets by predicting renewable energy share and CO_2 emissions based on a range of socio-economic, environmental and geographical factors. Using country-level data from 2000-2020, we developed a machine learning model to estimate the renewable energy percentage within total energy consumption and forecast CO_2 emissions. We trained a multi-layer perceptron (MLP) regressor on a dataset sourced from the World Bank and International Energy Agency, achieving an r-squared value of 0.94 for renewable energy share predictions and 0.99 for CO_2 emissions predictions. These high accuracies suggest that this model could support policymakers in setting achievable sustainability goals tailored to specific national circumstances.

Keywords
Machine learning, sustainable development, sustainable energy

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 enacted 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 CO_2 emissions by far, with almost triple the emissions of the second-place country [2]. This begs the question: how plausible are these environmental goals in less-developed countries? This research aims to determine what environmental goals are plausible for countries based on many different qualities and whether or not machine learning can be a useful tool in determining this. 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 CO_2 emissions based on its r-squared value.

Methods

Dataset

The data set used for this project is a Kaggle dataset that included observations from most countries in the world from 2000-2020. 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 cells were in the columns that we predicted, the entire row was deleted. When the empty cell appeared elsewhere, it was treated as a zero. The entire dataset was then normalized in order for the model to produce more accurate results. Ninety percent of the data was used to train the model and the remaining ten percent was used to test the model. The columns in the dataset contained various different characteristics of a country, including data that has a direct relationship to sustainability such as CO_2 emissions to those that do not seem to have a direct relationship with sustainable energy such as latitude [3].

Model

The prediction model used for this project is an MLP regressor model. We ran experiments to predict both the total share of renewable energy in final energy consumption and the total CO_2 emissions. 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. The graph is a loss epoch graph, which shows the relationship between the difference between predicted values and actual values and the number of iterations that the model has gone through all the training data. The lower the loss is, the more accurate the model is.

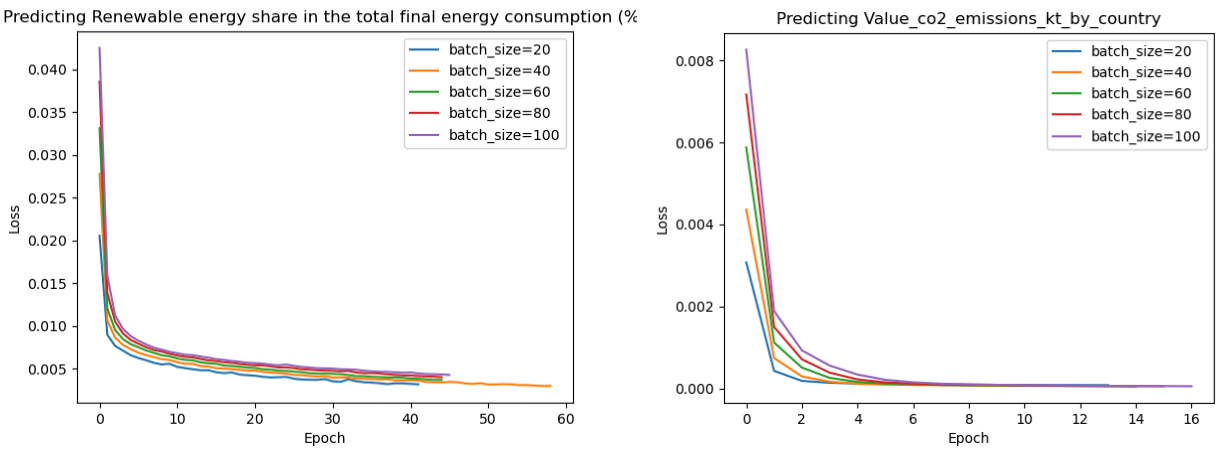

Figure 1. *Left side*: Loss vs epoch graph of the model predicting renewable energy share in the total final energy consumption while trying different batch size values. *Right side*: Loss vs epoch graph of the model predicting CO_2 emissions while trying different batch size values.

Using Figure 1 shown above, we determined that the best batch size is around 80. The batch size represents the amount of samples that the model goes through before changing its decision-making process. A model with a batch size of 80 reaches the lowest loss at a relatively quick rate, making it the best choice for the batch size.

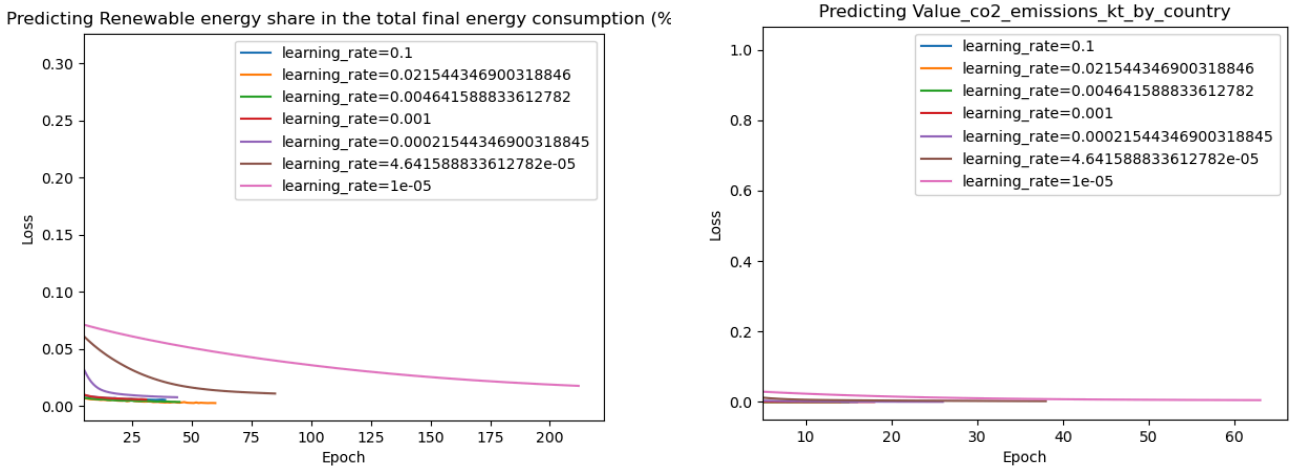

Figure 2. *Left side:* Loss vs epoch graph of the model predicting renewable energy share in the total final energy consumption while trying different learning rate values. *Right side:* Loss vs epoch graph of the model predicting CO_2 emissions while trying different learning rates.

Using Figure 2, we determined that the best-hidden layer size is a tuple at approximately (10000,). Hidden layers are the stages between the input and output. It is here that the model makes decisions and goes from the information given in the input to the conclusion in the output. The hidden layer size refers to the amount of layers in the hidden layer and the number of neurons used for each layer.

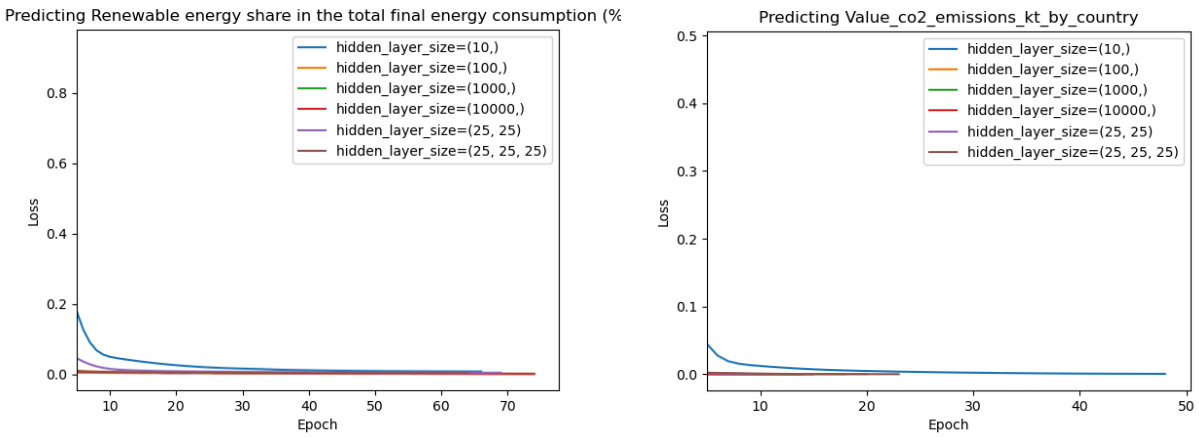

Figure 3. *Left side:* Loss vs epoch graph of the model predicting renewable energy share in the total final energy consumption while trying different hidden layer size values. *Right side:* Loss vs epoch graph of the model predicting CO_2 emissions while trying different hidden layer sizes.

Using Figure 3, we determined that the best learning rate is around 0.00681292. The learning rate affects how much the model changes based on its error during training.

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 accounts for all the variance observed in the independent variable. The closer the r-squared value is to 1 the more closely the model can account for variance in the outcomes. 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 CO_2 emissions. However, there were some outliers where negative numbers appeared where they should not have. This model could be used for many other datasets in order to solve other types of problems.

Discussion

This study successfully proved the hypothesis that machine learning can be a useful tool for predicting realistic sustainability goals and policy making. The model achieved r-squared values of 0.94 and 0.99, which are close to the maximum accuracy of 1. There is very little change in the r-squared value of the model after de-normalizing its predictions, meaning that it could be used in real-world circumstances.

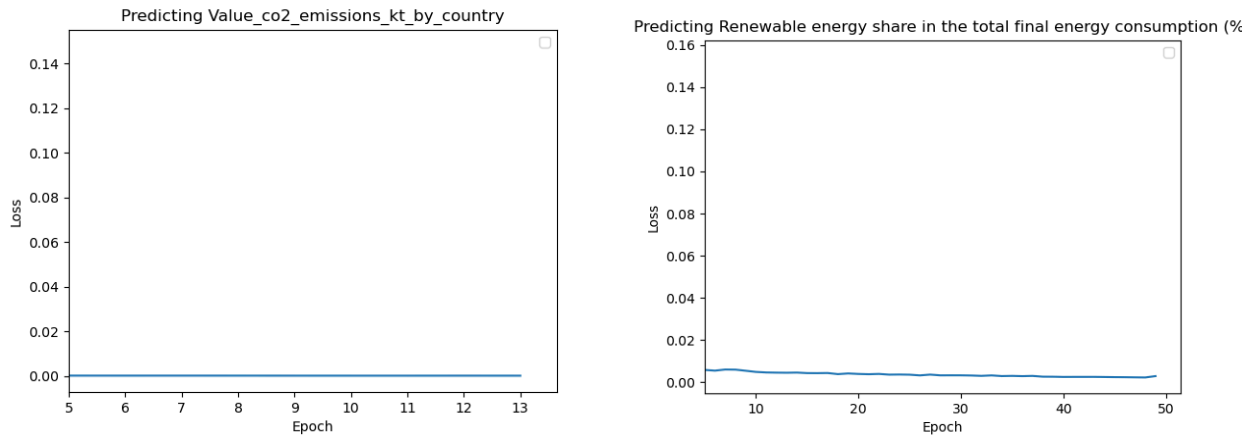

Figure 4. *Left side:* Loss vs epoch graph of the final model predicting renewable energy share in the total final energy consumption. *Right side:* Loss vs epoch graph of the final model predicting CO_2 emissions.

As shown in Figure 4 above, the loss epoch graph for the final model stops training at a small loss value. The loss represents how far off the model's predictions are from the actual value. The lower the loss of the model, the more accurate it is.

Existing research validates the findings of this study. In 2024, Astrom et. al. conducted a similar study where they used various types of machine-learning algorithms to estimate ground-level CO_2. The study used 4 different types of prediction algorithms: simple linear regression as a baseline, category boosting and extreme gradient boosting, and MLP. When comparing the accuracies of each algorithm, MLP, the one used in this paper, performed by far the best. However, some of the most important features that they used to predict ground-level CO_2 were unavailable in the dataset that we used. The most important factors in predicting ground-level CO_2 were Orbiting Carbon Observatory (OCO) data as well as surface temperature. While both of these characteristics are missing from the dataset that was utilized, the accuracy of the model was not negatively affected as we used non-environmental factors such as GDP [5].

Li et. al. used machine learning to create high-resolution CO_2 data for 1970-2018 based on daily emissions data from 2019-2023. The model specifically uses variations in temperature and temporal variables (day of the week, month) and the model predictions were confirmed with data from Vulcan from 2010-2015 [4]. Rather than the MLP regression model that was used for this project, they opted to use Extreme Gradient Boosting models [6]. Instead of the many monetary factors looked at in this study, Li et. al. focused mainly on the effect of extreme weather on carbon emissions. Their end goal was to assist with policy-making, which is similar to this study [4].

While this study builds upon the research of many comparable studies in literature, it is not without limitations. In particular, 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.

Overall, this model is useful as a reference for policymakers and environmental conservationists to get an understanding of what goals are achievable and realistic for their circumstances. There are many other factors that the model could leverage in order to make more realistic and accurate predictions. In particular, sustainability cannot be quantitatively understood without considering the international relationships and political dynamics of a country.

Conclusion

This study assessed whether a machine learning model can be developed to predict renewable share and CO_2 emissions for countries worldwide based on socio-economic and environmental factors. With r-squared values of 0.94 for renewable energy share predictions and 0.99 for CO_2 emissions predictions, our model demonstrates a high level of accuracy. These results indicate that machine learning can be an effective tool for evaluating realistic sustainability targets and guiding policy that aligns with each country's capabilities. Future research could expand this model to incorporate more dynamic factors or explore the inclusion of additional environmental indicators to increase its predictive robustness.

References

1. United Nations. Global Issues [Internet]. United Nations. 2021. Available from: https://www.un.org/en/global-issues
2. World population review. Pollution by country 2020 [Internet]. worldpopulationreview.com. 2023. Available from: https://worldpopulationreview.com/country-rankings/pollution-by-country
3. Global Data on Sustainable Energy (2000-2020) [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy
4. Li T, Wang L, Qiu Z, Ciais P, Sun T, Jones MW, et al. Reconstructing Global Daily CO_2 Emissions via Machine Learning [Internet]. arXiv.org. 2024 [cited 2024 Nov 13]. Available from: https://arxiv.org/abs/2407.20057
5. Åström O, Geldhauser C, Grillitsch M, Hall O, Sopasakis A. Enhancing Carbon Emission Reduction Strategies using OCO and ICOS data [Internet]. arXiv.org. 2024 [cited 2024 Nov 13]. Available from: https://arxiv.org/abs/2410.04288
6. Nvidia. What is XGBoost? [Internet]. NVIDIA Data Science Glossary. 2024. Available from: https://www.nvidia.com/en-us/glossary/xgboost/
7. MLPRegressor [Internet]. scikit-learn. Available from: https://scikit-learn.org/dev/modules/generated/sklearn.neural_network.MLPRegressor.html