Alex Lark and Lindsey Lee

Professor Wimberley

CPSC 393

December 12, 2025

Driving Global Sustainability: Forecasting the Renewable Energy Impact

Introduction

The transition to renewable energy has become the key global strategy for reducing carbon emissions and the effects of climate change. The goal is for carbon emissions to be reduced by almost 50% by 2030 for the world to avoid the worst impacts (United Nations). In order to power a safer future, we must “invest in alternative sources of energy that are clean, accessible, affordable, sustainable, and reliable” (United Nations). While many countries continue to expand their capacity for renewable energy, their status in making a positive reduction varies immensely. Understanding this relationship is critical for evaluating the true impact of adopting renewable energy and identifying where progress is most effective across the world. This is why we would like to address which countries will reach 50% of electricity produced by renewables by 2030 and which countries will benefit from renewable energy, where they are most effective in cutting down CO_2 emissions.

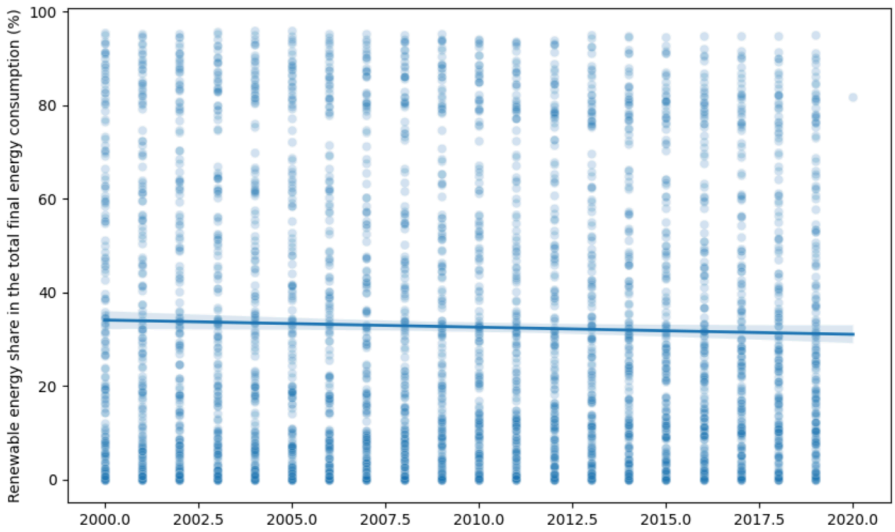
Background

This project examines the impact of adopting renewable energy across countries using a Global Data on Sustainable Energy (2000-2020) dataset from Kaggle. The dataset showcases several sustainable energy indicators and other financial data across 176 unique countries, ranging from developed and developing. It has 21

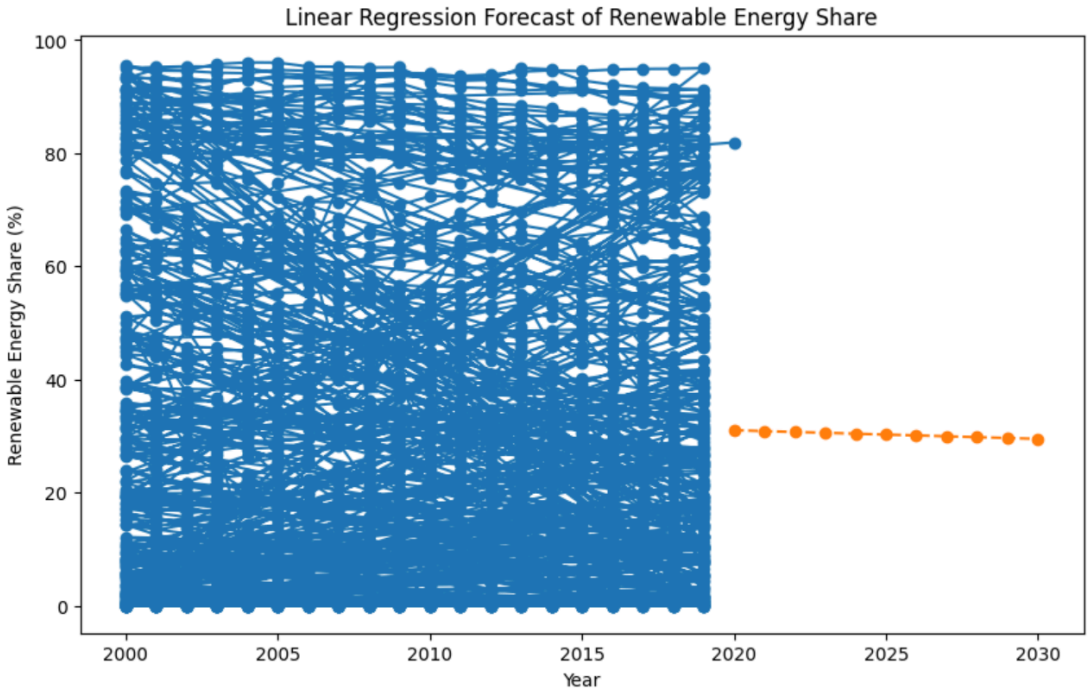
features, key variables being entity (country), year, electricity access (%), renewable energy share in total final energy consumption(%), CO_2 emissions (kt), energy intensity level, types of electricity (from fossil fuels, nuclear, or renewables), financial flows, and GDP per capita. The combination of all these indicators provide us with both environmental outcomes and economic context for making forecasted predictions. Although, after exploring the data, we noticed that there were a lot of missing data values scattered throughout the dataset randomly, as this was a real-world dataset. We discovered that there was one particular country, French Guiana, that was missing almost all features, including latitude and longitude. As a result, we decided to only remove that row. This brought our initial dataset from 3,649 rows to 3,648 rows. This now sets us up to begin our analysis.

Question 1

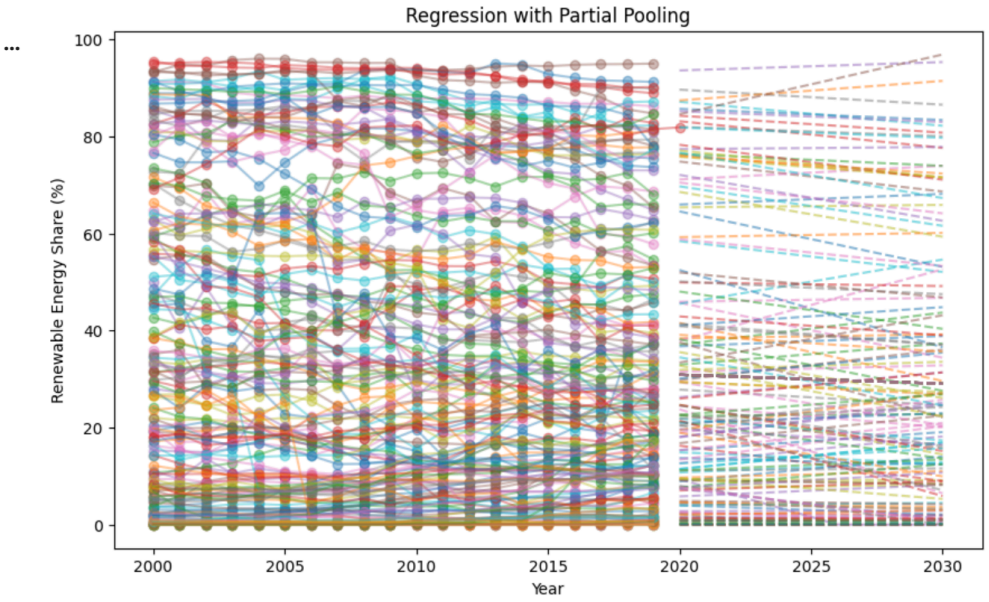
The first question we wanted to answer was: which countries will reach 50% of electricity produced by renewables by 2030? We decided that the best way to answer this was using regression, with Entity and Year as the predictors and renewable energy share as the predicted value, and adding partial pooling to get a by-country model.



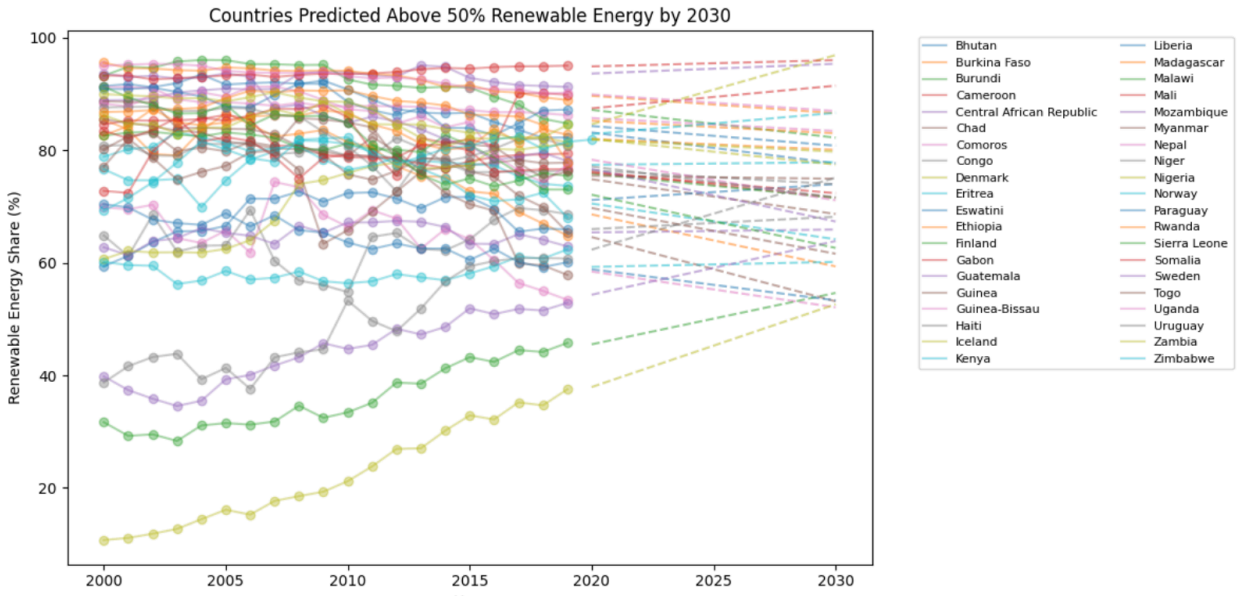
We began by exploring the data a bit, creating a scatter plot of year vs renewable energy and adding a line of best fit. This scatter plot actually indicated a slight negative trend in renewable energy percentage which we found interesting. Looking at it, we thought some possible explanations might be because all countries were treated equally in the plot, regardless of size or development status, and developing countries may be seeing a decline in renewable percentage as their need for electricity outpaces their access to renewables, contributing to this overall negative trend.



After that, we dropped approximately 200 rows of data where renewable energy percent was NA, and ran a basic linear regression model with a time series split. The time series split allowed me to train the model by using years such as 2005-2010 to predict the year 2011, aiding the model since it's a time based model. After running, we found the model had an MAE of 25.5, but because it plotted a linear regression line on all data it didn’t tell us much.



Next, we moved on to partial pooling. We ended up scaling the year for this as otherwise the model was not working well, but once we scaled the year that fixed the problems we were having. Adding partial pooling, the model was able to plot a trend line for each individual country while still keeping the overall average in mind, so countries with less data or inconclusive results would be pulled towards the global mean. This model with partial pooling ended up having an MAE of 25.25.



Looking at the results from partial pooling, we found that 30 countries were expected to exceed 50% renewable energy by 2030. Of those, Denmark and Finland were both expected to pass 50% in the 2020-2030 time range, while Angola, Cambodia, Nicaragua, Papua New Guinea and Sudan were expected to fall below 50%. This supports our earlier hypothesis regarding developing countries and highlights the divide between developed vs developing countries.

Denoising Autoencoder (DAE)

To move any further with our analysis, we needed to address a big issue: missing data. While the data was, for the most part complete, with so many features most of the rows had just 1 or 2 missing values. Not a lot compared to the amount of data we did have, but a problem when it came to dropping rows. Simply dropping rows with NA would bring our data from over 3000 rows down to just over 300, which would not work well for our analysis.

We decided that the best way to address this problem was to impute our data, or to essentially create artificial data points matching the overall trends to fill in the missing values. To do this, we used a Denoising Autoencoder, or a DAE. A DAE is a type of neural network generally used to clean noisy data, or often used to impute missing data. In our case, we would be using it to impute missing data.

To run the DAE, we first started by filling in the NaN values with the mean for that column and country so we could at least run the data, and creating a mask layer that labeled that data as artificially created. Next, we added noise to the original (non NaN) data and scaled all the data so it could be better processed by the DAE. After constructing the model itself, with layers forming an Encoder, Bottleneck, and Decoder,

we ran the model with an exponential decay learning rate, as we found that significantly improved model performance and metrics. While training the data, we specifically only trained on non-masked data, not the NaNs we artificially filled in. Finally, once trained, we ran the model on the masked data to essentially denoise it, creating reasonable data points out of our artificially filled in NaNs.

The model ended up with a loss of .15, which was pretty good considering how complicated the data was. Examining the artificially created data, most of it looked good, naturally fitting into the overall patterns of the data. However, the DAE specifically seemed to struggle with gdp_per_capita and consistently came up with values that were too large to make sense, and we were unable to solve this issue even after significant tweaking of the model.

Question 2

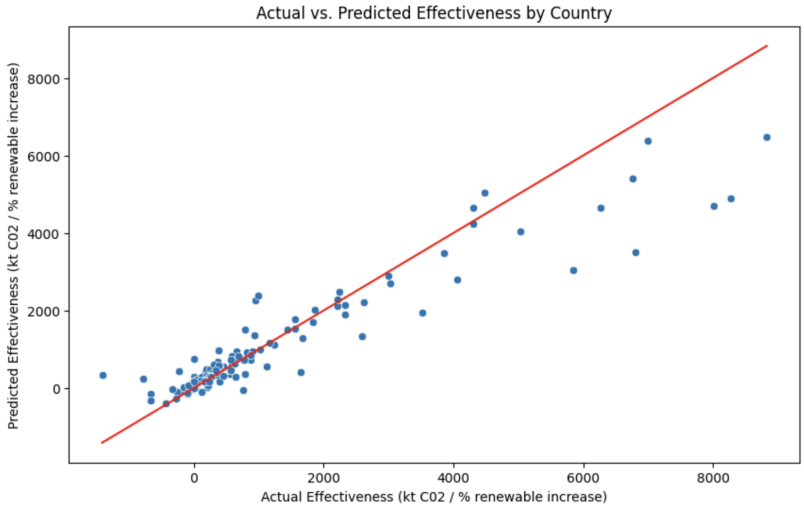
The second question we wanted to address was: which countries will benefit from renewable energy, where they are most effective in cutting down CO_2 emissions? We began by preparing the data for the model with feature engineering. In order to capture the true impact of renewable energy adoption, we decided to calculate the year-to-year changes to get the annual changes in percentage change in renewable energy share and in CO_2 emissions per country. We used the results of these two values to calculate our own effectiveness metric to quantify how much CO_2 emissions changed for unit increase in renewable energy adoption. Effectiveness = -(CO_2 emissions)/(change in renewable energy share). This formula will indicate to us that higher values mean larger reductions in CO_2 emissions after increases in renewable energy. Extreme values in the effectiveness metric were filtered out using the 1st and

99th percentiles to reduce the influence of outliers (this was something we went back to add after receiving weak model results). In addition, because there were heavy year-to-year fluctuations for some countries, we needed to make the effectiveness metric more stable. A solution was to apply a 3-year rolling average which smoothed the effectiveness value over a combined average of the current year and two years before. Smoothing allowed for the effectiveness to better sustain renewable energy adoption trends rather than face random spikes (this was also something added to improve model results).

We explored using a multilayer perceptron (MLP) neural network and gradient boosting methods. However, these models seemed to be sensitive to noise, outliers, and had scaling issues in the effectiveness metric, which resulted in a much weaker performance with accuracy scores below 0. We then tested and decided that a Random Forest regression model would best fit the relationship between all chosen predictors and the level of effectiveness. It proved to be less sensitive to extreme outliers and better handled nonlinear relationships in the data, working well with real-world data. The predictors used were GDP per capita, energy intensity, energy consumption per capita, population density, access to electricity, renewables capacity, electricity from renewables, low-carbon %, and renewable energy share. The dataset was split into training (80%) and testing (20%) sets, with predictor variables (X) standardized prior to model fitting. The model was built with 1000 trees, a minimum 5 samples per leaf, and a random_state of 42.

Model performance was then evaluated using Mean Squared Error (MSE) and

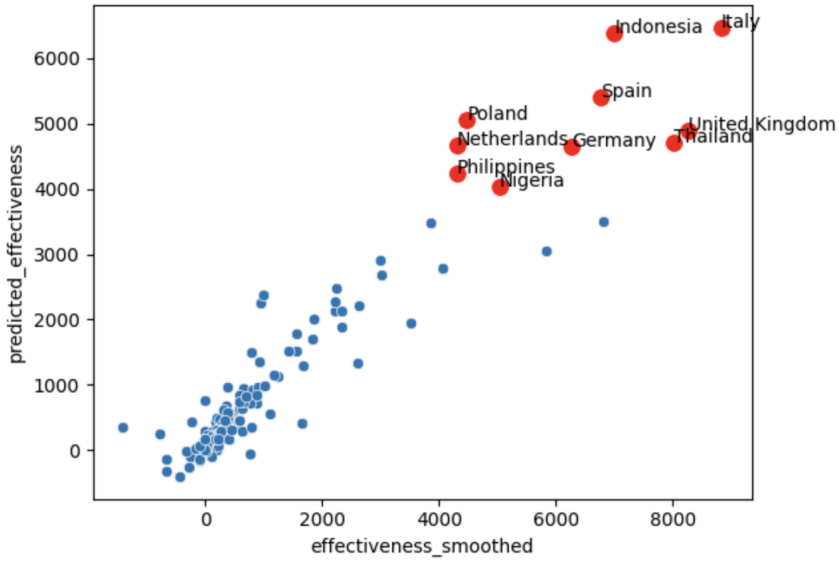
the coefficient of determination (R^2). The model was able to achieve an MSE of 2,249,979.08 and an R^2 score of 0.483. This means that 48.7% of the variation in renewable energy effectiveness across country year observations was explained by the model. While this level of accuracy is not in the higher range, this level of performance is quite reasonable given the complexity of national energy systems. There are also so many more external factors that influence CO_2 emissions reduction that aren't captured in the data, as countries also have unique policies, technologies, and geographic factors.



From the scatterplot, we can see that the majority of countries have low effectiveness. The model then becomes less accurate for countries with very high effectiveness as variation increases and underpredicts how effective they actually are. This suggests that large emissions reductions are more difficult to predict and may depend on country specific policies, infrastructure investments, or technological advances not seen by the data.

The top 10 performing countries consisted of Italy, Indonesia, Spain, Poland, the UK, Thailand, Netherlands, Germany, the Philippines, and Nigeria. These countries are

predicted to be most effective where renewable energy adoption will reduce CO_2. These countries tended to have more consistent renewable energy consumption to achieve a more sustained CO_2 emissions decline. Feature importance analysis also revealed that the top predictors were electricity from renewables and low-carbon electricity (%). This indicates to us that long-term policies will be more associated with shaping effectiveness in CO_2 emissions reductions rather than demographic factors.



Conclusion

Our analysis of the dataset provides us with some important insights regarding renewable energy use, its global impacts and progress towards global goals. It is important to note that despite having 21 features, there are so many things that contribute to renewable energy use, such as policy decisions and energy infrastructure in a given country that are not captured in this dataset, meaning our models can only be so accurate. It's also important to note that countries have significant year-to-year fluctuations and nonlinearity, decreasing the accuracy of the models. Because of that,

we are using the models as a way to gauge global trends and large-scale insights, rather than focusing on specific data that is only as good as the model.

From the analysis of question 1, a few important findings stand out. The first is the distinction between developed and developing countries. Developed countries tend to have an overall positive trend, increasing renewable energy use over time. However, as we can see from the results of question 1, developing countries actually tend to have a declining use of renewable energy as their energy needs expand faster than renewable energy access. This has significant implications on the global level, and indicates more effort should be made towards adopting renewables, especially in developing countries.

From the analysis of question 2, we can see that the expansion of renewable energy does successfully lead to the decline of CO2 emissions. Specifically, countries that focused on consistent renewable energy adoption saw the most reliable declines. This indicates that countries should focus on longer-term, consistent renewable energy adoption to continue to decrease CO2 emissions.

For the future, there are a few ways we could improve our work. For the first question, one possible solution would be to test non-linear models, supporting the non-linearity of the data and possibly giving us more accurate results. It's also important that we further refine the DAE, as, while mostly accurate, it did have some issues, and improving the model would give us more confidence in models using the data it imputes. Beyond that, it would be important to see if we could gain more insightful information by expanding our dataset. This could be finding more recent data that goes up to 2025, or finding data with even more features that could further improve our analysis.

Works Cited

Hasan, Syed. “AutoEncoders: Theory + PyTorch Implementation - Syed Hasan - Medium.” *Medium*, 24 Feb. 2024, medium.com/@syed_hasan/autoencoders-theory-pytorch-implementation-a2e72f6f7cb7

Tanwar, Ansh. “Global Data on Sustainable Energy (2000-2020).” *Kaggle*, 2023, www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy.

United Nations. “Renewable Energy – Powering a Safer Future.” *United Nations*, 2025, www.un.org/en/climatechange/raising-ambition/renewable-energy..