Predicting Renewable Energy Growth Using Machine Learning

Edgar Venegas
Specialization: Sustainable Systems
University of Michigan, SEAS
440 Church St, Ann Arbor, MI 48109
Email: evenegas@umich.edu

***Abstract*—The global transition to renewable energy presents a complex set of challenges and opportunities. This project explores the use of machine learning models to predict the share of renewable energy across countries based on infrastructure investments, capacity, production, and environmental metrics. Using a data set from 2010 to 2023, the study applies Linear, Ridge, Lasso, Random Forest, and XGBoost models to evaluate the performance and abilities of renewable forecasting efforts.**

Github Repo Link: Edgar's Final STATS 507 Project

I. Introduction

Nations globally are shifting towards using renewable energy to achieve climate goals and reduce fossil fuel reliance. This shift is supported by investments in solar, wind, hydroelectric, geothermal, and biomass energy sources, guided by policy frameworks and financial incentives.

The rising energy demand has made it essential for system operators and energy planners to anticipate renewable energy growth. This project uses linear and regularized regressions as a baseline to compare with other models. Then, the machine learning models are leveraged—XGBoost, and Random Forest to forecast the share of renewable energy of countries using publicly available data from Kaggle's *Renewable Energy Adoption & Climate Change Response (2010–2023)* dataset.

Several recent studies have utilized statistical and deep learning techniques for similar forecasting problems. This approach integrates both feature engineering and model optimization to address the gaps in model performance and versatility.

II. Method

This project frames the renewable energy forecast problem as a supervised learning task, where the objective is to predict the annual share of renewable energy (%). The data set comprises multiple variables relevant to the development of renewable energy and the response to climate change. It was cleaned and pre-processed to build supervised learning models. Hyperparameter tuning was also used on both the baseline and advanced models to improve the models performance and enhance model generalization.

Categorical variables such as Country/Region and Energy Source were one-hot encoded. Quantitative features were normalized and standardized. These variables were then visualized to view outliers and potential trends throughout the timeline that the features can explain. Figure 1 shows the Renewable Share of all countries / regions with 5 data points outside of the central 90 %. The rest of the features were visualized across the time period to identify potential patterns and trends. Feature engineering was then used to include relationships that could potentially predict future renewable shares.

TABLE I
Renewable Energy Adoption Dataset Features

Feature	Description
Year	Data record year
Country/Region	Geographic identifier
Energy Source	Type of renewable energy
Energy Production	MWh generated
Policy Changes	Key legislation or incentives
Infrastructure Investment	Investment in USD
CO_2 Reduction	Environmental impact metrics
Installed Capacity	Megawatt (MW) capacity
Renewable Share	Share of renewable in energy mix (%)
Jobs Created	Employment impact from renewable projects

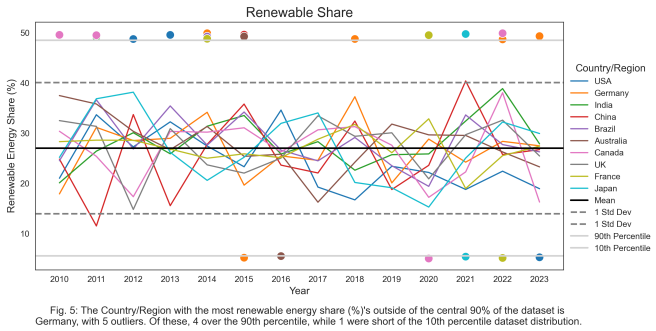

Fig. 5: The Country/Region with the most renewable energy share (%)'s outside of the central 90% of the dataset is Germany, with 5 outliers. Of these, 4 over the 90th percentile, while 1 were short of the 10th percentile dataset distribution.

Fig. 1. Renewables Share.

- Investment per MW and per ton of CO2 reduced
- Capacity Factor
- Lagging features such as prior year's investment and renewable share
- Log and quadratic transformations for variance and normalization

The data was split into five training-test scenarios (10%, 25%, 50%, 75%, 90% train). Models were developed using:

Baseline models included:

TABLE II
R^2 SCORES ACROSS MODELS AND TRAINING SET SIZES

Training %	Linear R^2	Ridge R^2	Lasso R^2	RF R^2	XGB R^2
10%	-1.17	0.00	0.00	-0.04	-0.10
50%	-0.14	0.00	0.00	-0.02	-0.02
90%	-0.10	0.01	-0.02	0.04	-0.02

- Linear Regression
- Ridge Regression (alpha optimized via GridSearchCV)
- Lasso Regression (with extended max iterations)

Advanced models included:

- Random Forest Regressor (100 trees)
- XGBoost (with learning rate tuning and early stopping)

The performance of each model was assessed using RMSE, R^2, MAE, and MAPE.

III. RESULTS

The modeling pipeline was implemented in Python using scikit-learn, XGBoost, matplotlib, and custom utilities for visualization, data summary, and evaluation. Figures were generated for data distribution, feature correlation, and model diagnostics.

Data Pipeline & Figures:

- Visual EDA included time-series line plots of features and trends, alongside heatmaps showing feature correlations and R^2 scores.
- Feature selection heatmaps revealed that there was a weak relationship between most engineered variables and the target.
- Model Predictions were visualized using actual vs predicted plots and residual scatter plots in Appendix B.

Random Forest showed the best R^2 value of 0.04 with 90% training, outperforming other models in capturing complex, non-linear relationships. However, all models struggled to generalize with low R^2 scores across the board, indicating potential data limitations.

IV. CONCLUSION

Despite using a range of machine learning techniques linear regression, regularized regressions, and ensemble learners predictive performance remained weak, with R^2 values near or below zero across multiple different training splits. This project evaluated five machine learning models to forecast renewable energy share using preexisting set of infrastructure and environmental indicators. The findings of this report suggest that:

- Current features fail to capture the policy, geopolitical, or grid-integration complexities that factor into renewable energy adoption.
- Non-linear models like Random Forest and XGBoost slightly outperformed linear regressions, but not significantly.
- Further work needs to be incorporated to encompass macroeconomic, regulatory, and time-series dependencies to improve forecasting accuracy.

This analysis reinforces the importance of high-quality, diverse data in machine learning for energy systems planning and calls for more granular temporal and spacial datasets to conduct a better analysis.

APPENDIX A
EXPLORATORY DATA ANALYSIS

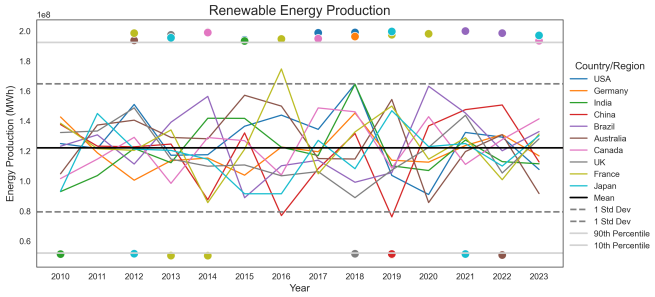

Fig. 1: The Country/Region with the most energy production (mwh)'s outside of the central 90% of the dataset is France, with 6 outliers. Of these, 4 over the 90th percentile, while 2 were short of the 10th percentile dataset distribution.

Fig. 2. Energy Production by Country.

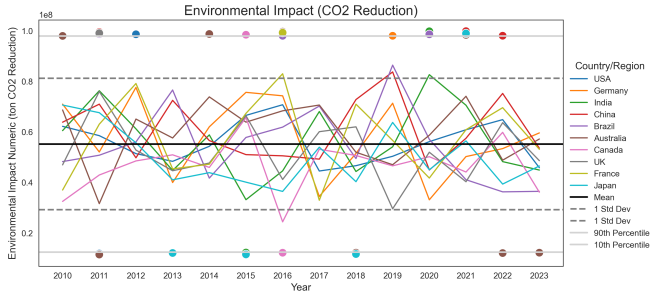

Fig. 3: The Country/Region with the most environmental impact numeric (ton co2 reduction)'s outside of the central 90% of the dataset is Australia, with 9 outliers. Of these, 4 over the 90th percentile, while 5 were short of the 10th percentile dataset distribution.

Fig. 3. Environmental Impact by Country.

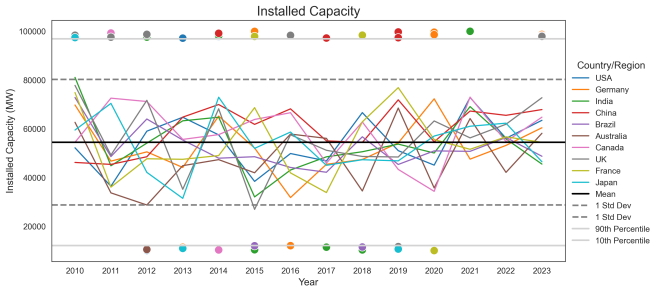

Fig. 4: The Country/Region with the most installed capacity (mw)'s outside of the central 90% of the dataset is India, with 6 outliers. Of these, 5 over the 90th percentile, while 3 were short of the 10th percentile dataset distribution.

Fig. 4. Installed Capacity by Country.

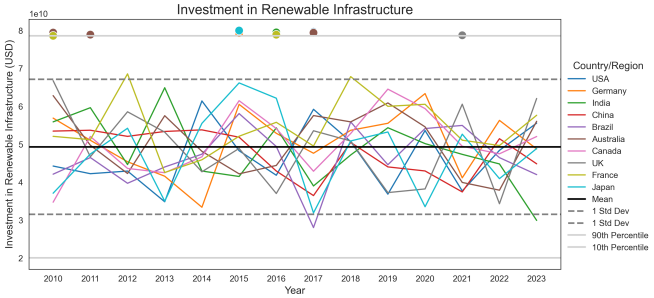

Fig. 2: No significant groups of outliers detected for investment in renewable infrastructure (usd)'s beyond the central 90% of the dataset grouped by Country/Region.

Fig. 5. Investment in Renewable Infrastructure by Country.

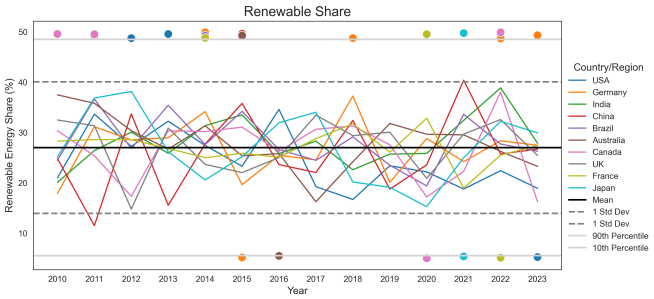

Fig. 5: The Country/Region with the most renewable energy share (%)'s outside of the central 90% of the dataset is Germany, with 5 outliers. Of these, 4 over the 90th percentile, while 1 were short of the 10th percentile dataset distribution.

Fig. 6. Renewables Share.

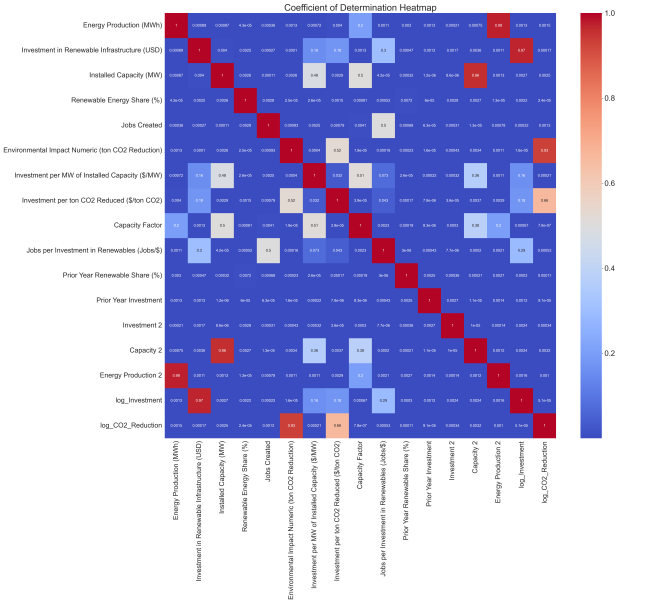

Fig. 7. R^2 Heatmap.

APPENDIX B
MODEL VISUALIZATIONS

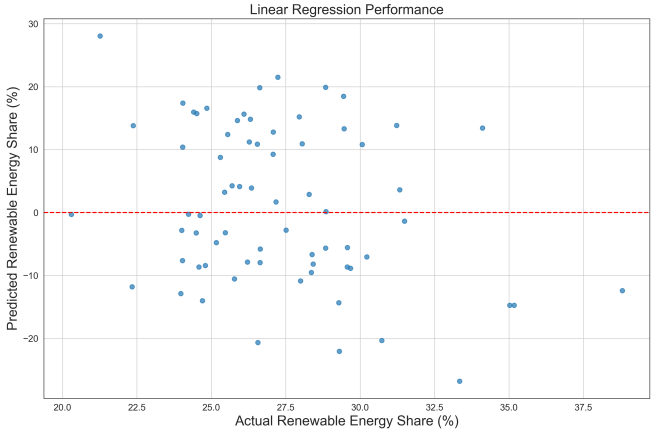

Fig. 8. Linear Regression.

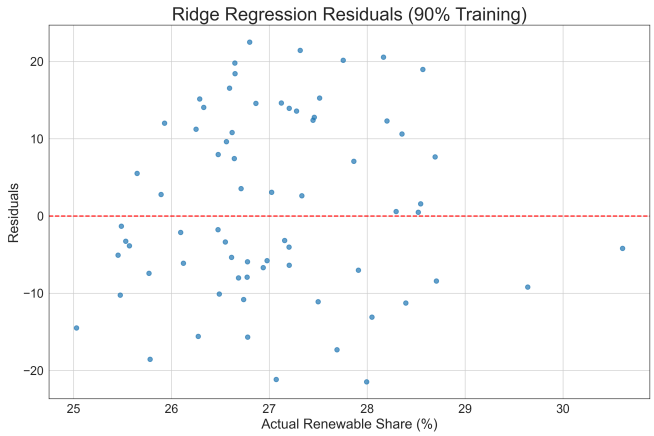

Fig. 9. R^2 Heatmap.

Fig. 10. Lasso Regression.

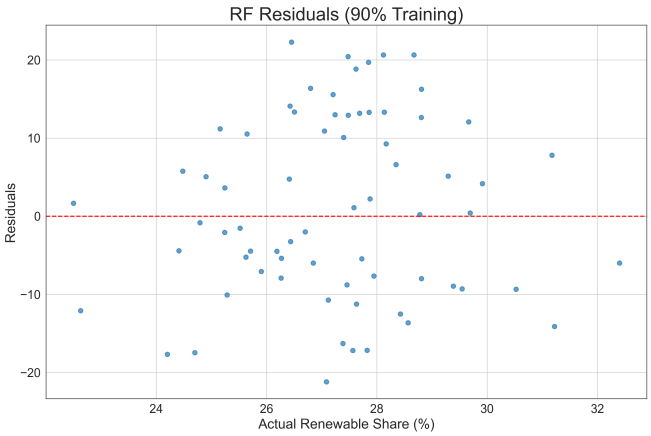

Fig. 11. Random Forest.

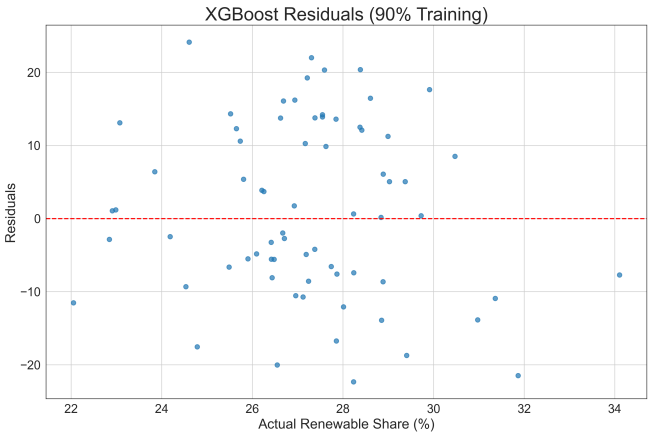

Fig. 12. XGBoost.