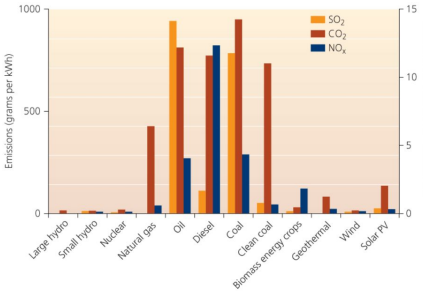
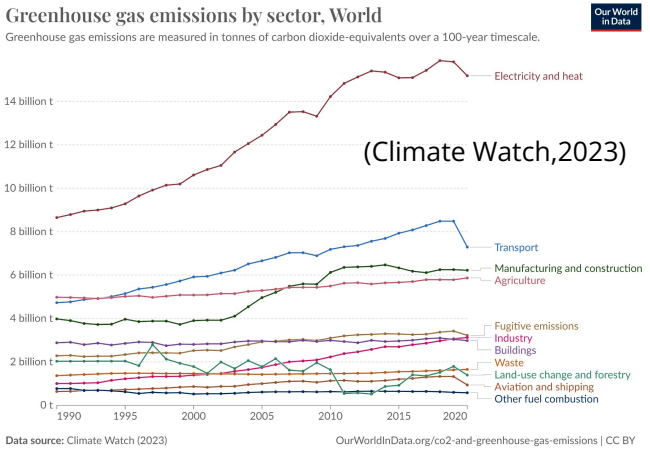
Predicting Green Energy Usage

Agishan Thaya, Abhishek Srikaran, Johnny Cao,
David Ly, Thomas Oliver

(Withgott & Laposata, 2021)



Motivations & Goals



- Energy industry is the largest contributor of greenhouse gas emissions by a significant margin (Climate Watch,2023)
- Green energy sources produce minimal greenhouse gases during operation (Withgott & Laposata, 2021)
- Forecasting green energy usage by country can inform us which countries are lagging behind on reaching emission goals
- By finding correlations between the features and the output, we can determine which factors have the greatest effect on green energy adoption which we can inform countries on what actions can be taken to meet their goals
- Utilized linear regression, lasso regression, random forest, and xgboost to identify the most effective model with the correct predictors
- Our plan was to clean our data by dropping any missing information or filling it in, then develop insights on which factors are important towards our goal, and finally train, test and develop different models to identify which would be the best in helping us identify the needs for our goals
- The reason why we chose our dataset is because it presented all of the correct factors that people would look for when looking for clues to increase green energy production and lowering greenhouse gas emissions
    - The dataset included a variety of countries across a span of 20 years which showcases the data that would be needed to determine if a country is improving or failing

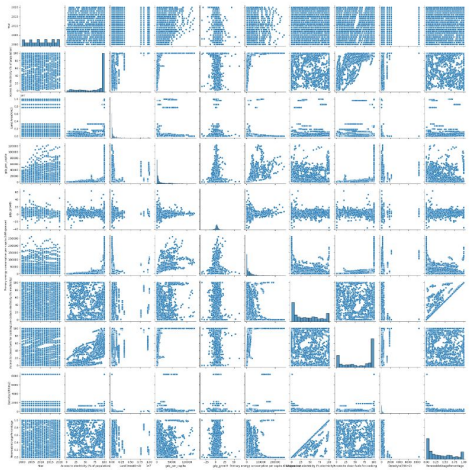
Dataset Insights

- Sourced from kaggle
- Self described as a dataset containing sustainable energy indicators (Tanwar,2023)
- 3649 observations initially
- 176 countries
- 21 features
- All columns are numeric except country
- Calculated outcome label using the formula:

Green Energy % = (Total Green Energy)/∑(Energy From Each Source)

- Represents the amount of green energy produced compared to all forms of energy
- Dataset is useful as it contains several features which are related to the viability of green energy such as land area, density, and gdp (Gross, 2020).

```
RangeIndex: 3649 entries, 0 to 3648
Data columns (total 21 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Entity                                                  3649 non-null   object
 1   Year                                                    3649 non-null   int64
 2   Access to electricity (% of population)                 3639 non-null   float64
 3   Access to clean fuels for cooking                       3480 non-null   float64
 4   Renewable-electricity-generating-capacity-per-capita    2718 non-null   float64
 5   Financial flows to developing countries (US $)          1560 non-null   float64
 6   Renewable energy share in the total final energy consumption (%)  3455 non-null   float64
 7   Electricity from fossil fuels (TWh)                     3628 non-null   float64
 8   Electricity from nuclear (TWh)                          3523 non-null   float64
 9   Electricity from renewables (TWh)                       3628 non-null   float64
 10  Low-carbon electricity (% electricity)                  3607 non-null   float64
 11  Primary energy consumption per capita (kWh/person)      3649 non-null   float64
 12  Energy intensity level of primary energy (MJ/$2017 PPP GDP)  3442 non-null   float64
 13  Value_co2_emissions_kt_by_country                       3221 non-null   float64
 14  Renewables (% equivalent primary energy)                1512 non-null   float64
 15  gdp_growth                                              3332 non-null   float64
 16  gdp_per_capita                                          3367 non-null   float64
 17  Density\n(P/Km2)                                        3648 non-null   object
 18  Land Area(Km2)                                          3648 non-null   float64
 19  Latitude                                                3648 non-null   float64
 20  Longitude                                               3648 non-null   float64
dtypes: float64(18), int64(1), object(2)
memory usage: 598.8+ KB
```



Data Cleaning

- Dropped columns missing over 150 observations and dropped countries missing over half of its data for any given column.
- Filled remaining nulls with mean for that country
- One hot encoded all countries for linear and lasso regression while used nominal encoding for other methods to minimize features for selection (optimized for training speed).
- After cleaning:
  - 3081 observations
  - 148 countries

Before Cleaning

```
df.isna().sum()

Entity                                                          0
Year                                                            0
Access to electricity (% of population)                        10
Access to clean fuels for cooking                             169
Renewable-electricity-generating-capacity-per-capita          931
Financial flows to developing countries (US $)               2089
Renewable energy share in the total final energy consumption (%)  194
Electricity from fossil fuels (TWh)                            21
Electricity from nuclear (TWh)                                126
Electricity from renewables (TWh)                              21
Low-carbon electricity (% electricity)                         42
Primary energy consumption per capita (kWh/person)              0
Energy intensity level of primary energy (MJ/$2017 PPP GDP)   207
Value_co2_emissions_kt_by_country                             428
Renewables (% equivalent primary energy)                     2137
gdp_growth                                                    317
gdp_per_capita                                                282
Density\n(P/Km2)                                                1
Land Area(Km2)                                                  1
Latitude                                                        1
Longitude                                                       1
dtype: int64
```

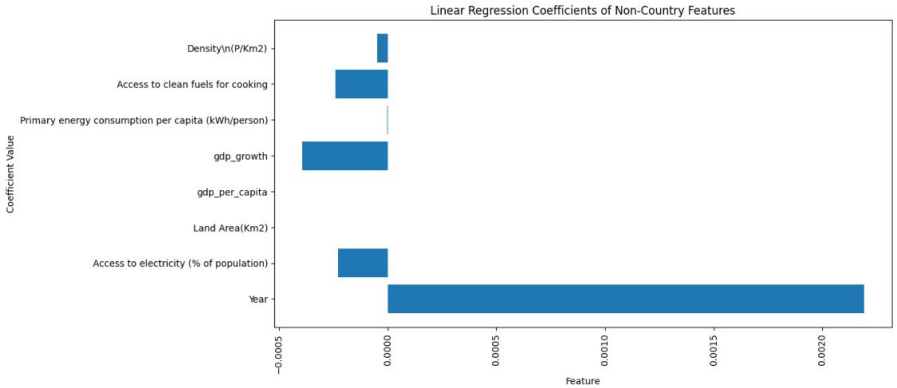
After Cleaning

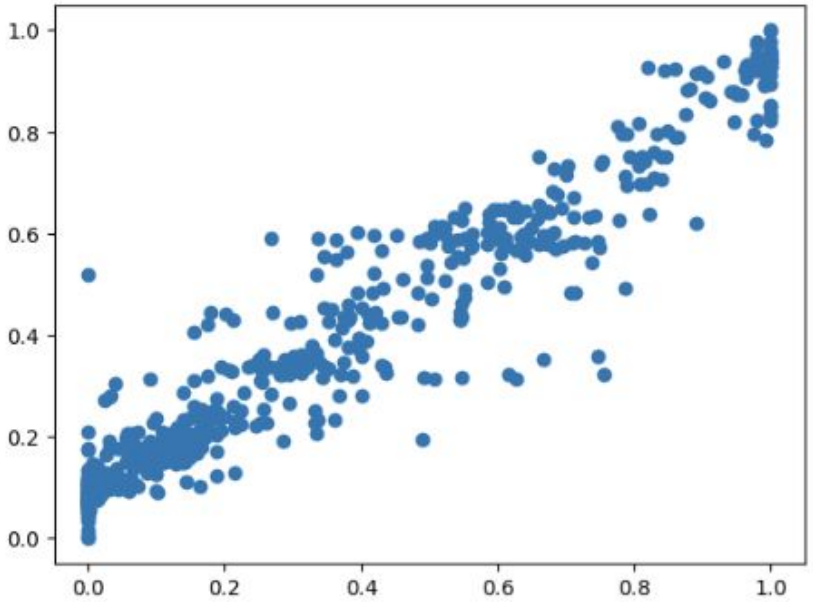
```
df.isna().sum()

Entity                                                          0
Year                                                            0
Access to electricity (% of population)                         0
Land Area(Km2)                                                  0
gdp_per_capita                                                  0
gdp_growth                                                      0
Primary energy consumption per capita (kWh/person)              0
Access to clean fuels for cooking                               0
Density\n(P/Km2)                                                0
RenewableUsagePercentage                                        0
dtype: int64
```

Linear Regression

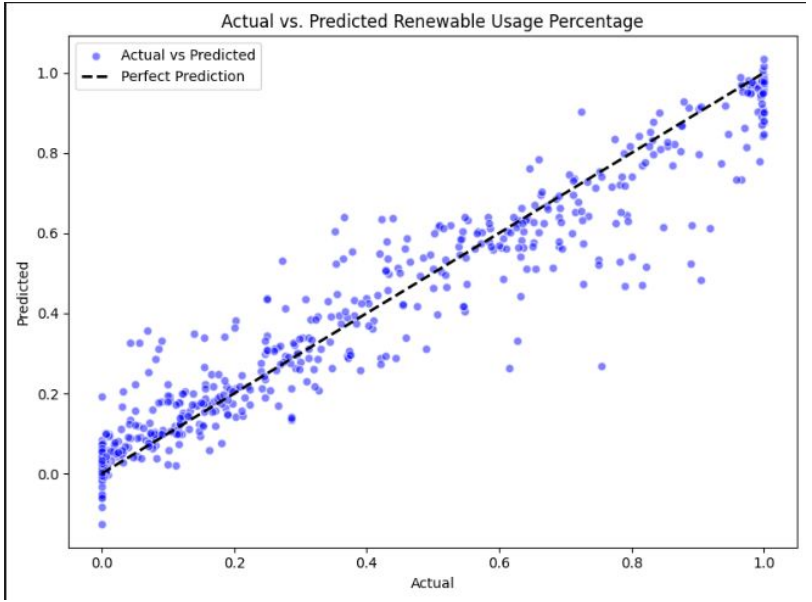
- Linear regression was used initially as a benchmark for other models due to its simplicity
- Total of 156 features split into 148 one-hot encoded country columns and 8 other columns
- Year plays a significant role in the influence of the result
- R2: 0.9130, MAE: 0.0798, MSE: 0.0099

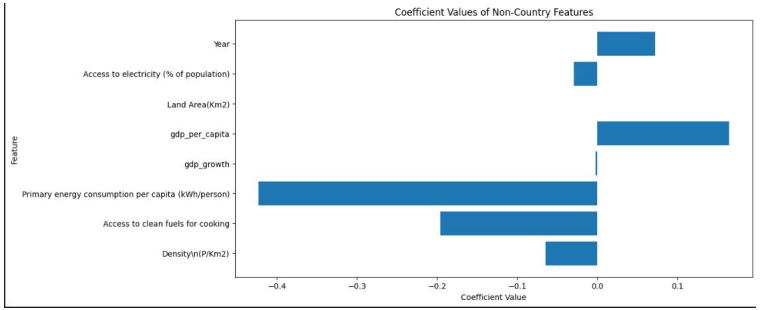




Lasso Regression

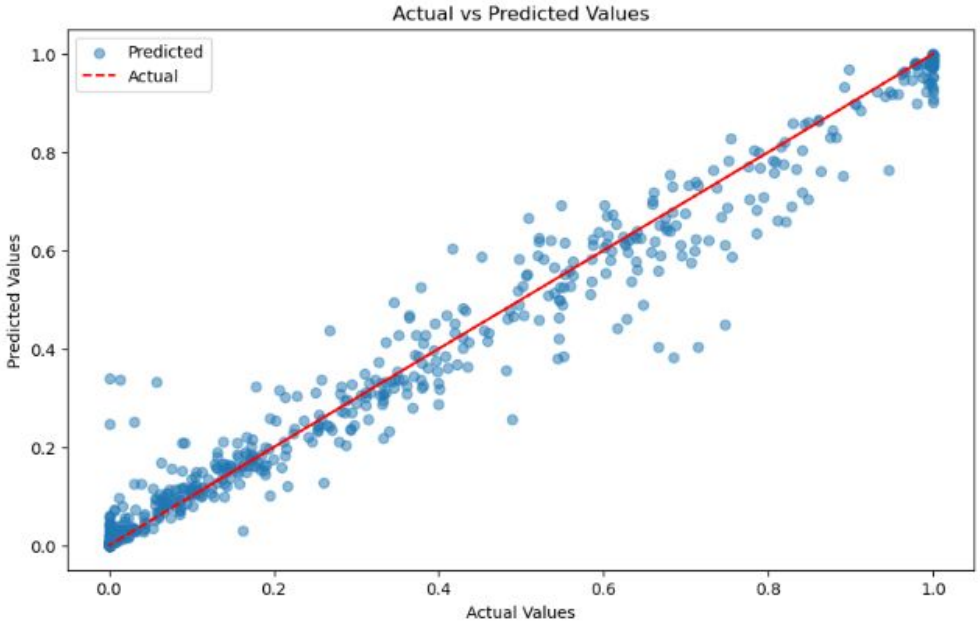
- Linear regression resulted in a base model that we can iterate and improve on. We can use lasso to drop less significant columns to try and improve the score.
- Using a LassoCV to calculate best alpha and cross validation of 5. Dropped 6 columns total.
- R2: 0.9326, lasso slightly improves on linear regression and changes the coefficient values drastically.
- Regression models are a decent fit and starter for our analysis.





Random Forest

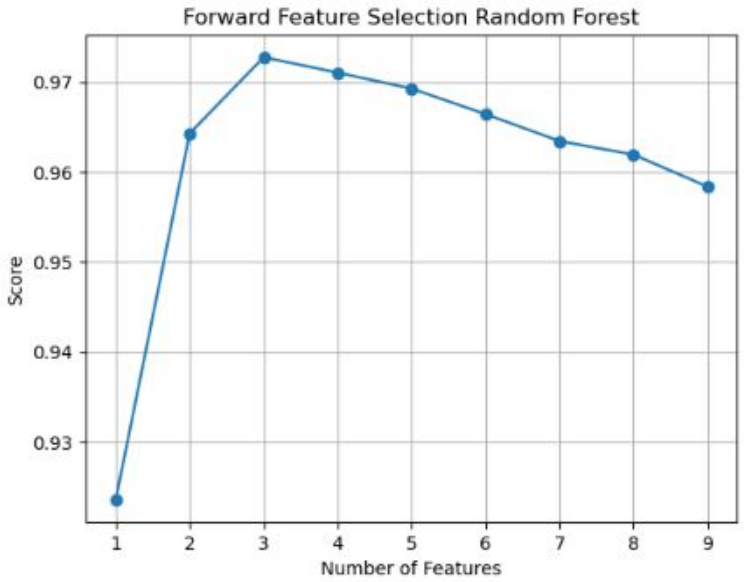
- RF is good for both classification and regression
- Great against overfitting since it takes the average across all trees
- RF is able to capture a lot of non linear relationships which were present when looking at the pairplot
- Utilized parameter tuning:
  - Max depth: 20
  - Min simples leaf: 1
  - Min samples split: 2
  - N estimators: 500
- MSE: 0.0037
- MAE: 0.0362
- R^2: 0.9678




```
Fitting 3 folds for each of 108 candidates, totalling 324 fits
Best parameters found:  {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}
Mean Squared Error: 0.0036686482264094906
Mean Absolute Error: 0.0362271351210359
R-squared: 0.967843504789073
```

Forward Selection RF

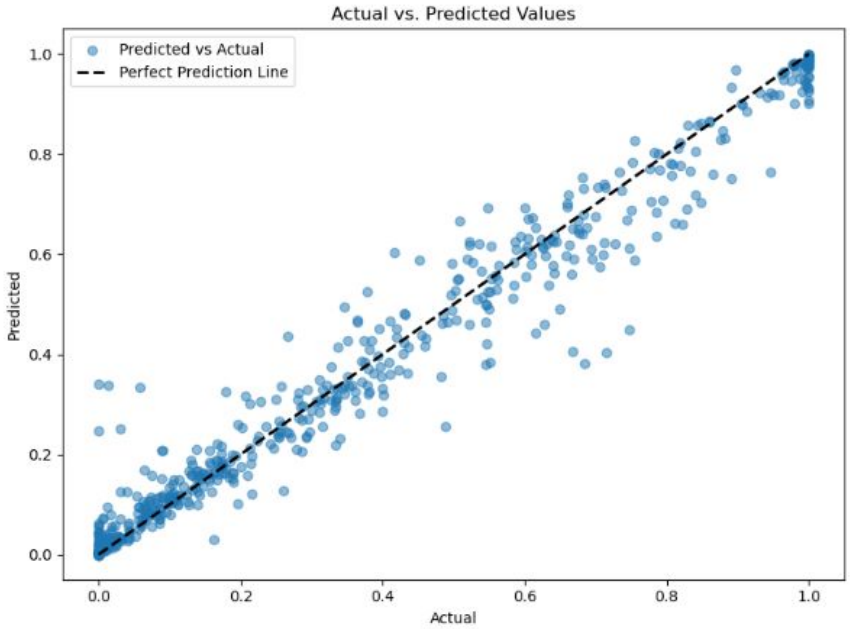
- Model with 3 features produce the highest score (0.9727)
- Country (Entity), Year, and Land Area
- The biggest factor is the Year, increases the score by 0.0407, all other factors make incremental increases
- Random Forest is a lot less prone to overfitting since it combines the predictions of multiple trees
- RF is a very simple model and allows for easy integration in both regression and classification tasks
- Goal: identify the countries that might need attention and investments
    - Easier to use regression to identify the renewable energy percentage the country can have in the near future to identify when issues might arise



1	(3,)	[0.9138853383435531, 0.9294270330554776, 0.920...	0.923559
2	(1, 3)	[0.9708646448822328, 0.968492240588324, 0.9721...	0.964298
3	(0, 1, 3)	[0.9750063277000633, 0.9774048944570429, 0.973...	0.972758
4	(0, 1, 3, 8)	[0.9723702945072267, 0.9752480479142669, 0.968...	0.971049
5	(0, 1, 3, 7, 8)	[0.9692109742249295, 0.9778956605464739, 0.966...	0.969276
6	(0, 1, 2, 3, 7, 8)	[0.9674953217639596, 0.973745637999459, 0.9647...	0.966435
7	(0, 1, 2, 3, 6, 7, 8)	[0.9607170643875428, 0.9718914856473395, 0.964...	0.96344
8	(0, 1, 2, 3, 4, 6, 7, 8)	[0.950527020274173, 0.9705355840200673, 0.9630...	0.961934
9	(0, 1, 2, 3, 4, 5, 6, 7, 8)	[0.948567884046858, 0.9696276313271378, 0.9569...	0.958376

XGBoost

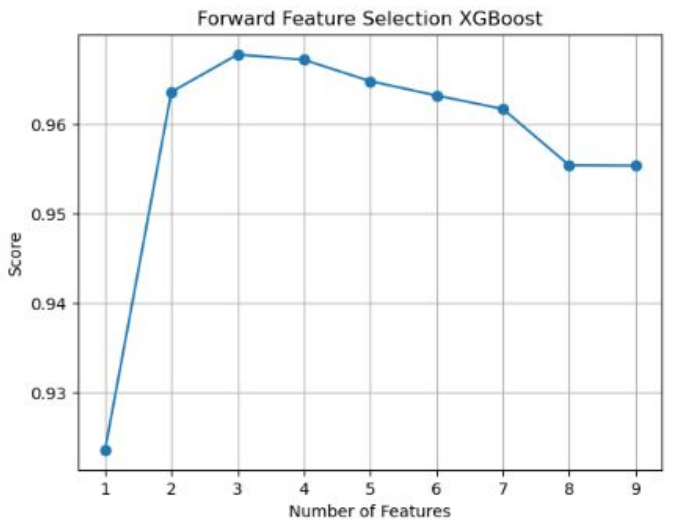
- Determined optimal hyperparameters for the XGBoost model:
- colsample_bytree: 0.7 (subsample ratio of features when constructing each tree)
- learning_rate: 0.2 (step size shrinkage used to prevent overfitting)
- max_depth: 7 (maximum depth of a tree)
- n_estimators: 500 (number of gradient boosted trees)
- subsample: 0.7 (subsample ratio of the training instances)
- Achieved a Mean Squared Error (MSE) of approximately 0.004 on the test set.
- Recorded a Mean Absolute Error (MAE) of approximately 0.0409 on the test set.
- Attained an R-squared of approximately 0.9636 on the test set
- Includes built-in L1 and L2 regularization which helps prevent overfitting
- Uses a depth-first approach which allows it to stop splitting a node when it encounters a negative loss




```
Fitting 3 folds for each of 108 candidates, totalling 324 fits
Best parameters found: {'colsample_bytree': 0.7, 'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 500, 'subsample': 0.7}
Best mean squared error from GridSearch: 0.005343264588401563
Mean Squared Error on test set: 0.004143015279657452
Mean Absolute Error: 0.04094202637027843
R-squared on test set: 0.9636855749646268
```

Forward Selection XGBoost

- XGBoost is great at handling datasets with a lot of zeros or missing values
- Label Encoding to include Countries
- Model with 3 features produce the highest score (0.967782)
- Country (Entity), Year, Density
- Year is the biggest factor, increasing the score by 0.04
- By having forward selection used by both XGBoost and Random Forest it provides a more diverse perspective on the data allowing to find different reason for why different factors affect the model



	feature_idx	cv_scores	avg_score
1	(0,)	[0.9140122037898019, 0.9286670496220254, 0.921...	0.923603
2	(0, 1)	[0.9616322994819942, 0.9692595372026982, 0.964...	0.963593
3	(0, 1, 8)	[0.9754208017270087, 0.9660346751807457, 0.965...	0.967782
4	(0, 1, 3, 8)	[0.9724821210011655, 0.9730535250822893, 0.965...	0.96719
5	(0, 1, 3, 7, 8)	[0.9669868915608232, 0.9750679913671518, 0.955...	0.964799
6	(0, 1, 2, 3, 7, 8)	[0.9617653929053025, 0.973458482142751, 0.9543...	0.963197
7	(0, 1, 2, 3, 6, 7, 8)	[0.9532878362339942, 0.971011262699587, 0.9581...	0.961699
8	(0, 1, 2, 3, 5, 6, 7, 8)	[0.9484484650654235, 0.9657313593223611, 0.953...	0.955408
9	(0, 1, 2, 3, 4, 5, 6, 7, 8)	[0.9530486887043186, 0.959052018456817, 0.9504...	0.955364

Results and Comparative Analysis

- From the models tested, forward selection using random forest is the model with the highest R^2 score
- Utilizing this model we are able to predict future values for renewable energy usage, allowing countries to have a gauge on where they might end up in the future
- This model does have limitations, since the data is only using from 20 years (2000-2020) the data can be very limiting as there are a lot of years that can help in the increase of accuracy, in addition some of the data was missing had to be replaced with the mean or be deleted leading to further limitations in predictions
- Through the usage of this model we are able to predict what changes need to be made to which countries to be able to meet global energy goals
- By 2030, 80% of new power generation are renewable resources (Iea 2023). With our model we will be able to give an accurate representation of the global goal and reasonability

Model	MAE	MSE	R^2
Linear Regression	0.0798	0.0099	0.9130
Lasso Regression	0.0589	0.0078	0.9326
Random Forest	0.0362	0.0037	0.9678
Forward Selection Random Forest	0.0356	0.0032	0.9727
XGBoost	0.0409	0.0041	0.9637
Forward Selection XGBoost	0.0353	0.0030	0.9678

References

Climate Watch. (2023, October 31). Greenhouse gas emissions by Sector. Our World in Data. https://ourworldindata.org/grapher/ghg-emissions-by-sector

GfG. (2023, May 23). Stepwise regression in python. GeeksforGeeks. https://www.geeksforgeeks.org/stepwise-regression-in-python/

Gross, S. (2020, January). Renewables, Land Use, and Local Opposition in the United States. Brookings. https://www.brookings.edu/wp-content/uploads/2020/01/FP_20200113_renewables_land_use_local_opposition_gross.pdf

guest_blog. (2024, January 16). XGBoost: Introduction to xgboost algorithm in Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/#:~:text=High%20accuracy%3A%20XGBoost%20is%20known,millions%20of%20rows%20and%20columns.

Iea. (2023, October 1). The Energy World is set to change significantly by 2030, based on today's policy settings alone - news. IEA. https://www.iea.org/news/the-energy-world-is-set-to-change-significantly-by-2030-based-on-today-s-policy-settings-alone

Tanwar, A. (2023, August 19). Global Data on Sustainable Energy (2000-2020). Kaggle. https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy

Withgott, J., & Laposata, M. (2021). Environment: The science behind the stories. Pearson.