CO_2 Emissions Simulator

Objective

It's a global paradox: the smallest greenhouse gas emitters frequently endure the most severe consequences of climate change. This dashboard was developed to highlight the profound disparity between top CO_2 contributors and the nations that are most vulnerable to climate change's effects. The data used for this project was compiled from Our World in Data, the ND-GAIN Index, and a public global GeoJson file on GitHub.

Data Collection

Data was collected from the U.S. Census Bureau, the U.S. Energy Information Administration, and the National Conference of State Legislatures. The following features were collected:

- State: Abbreviated U.S. state name.
- Year: Calendar year of observation.
- electricity_sales: Total electricity sold to end users (in billion kWh).
- coal_use: Energy consumption from coal (in billion Btu).
- natural_gas_use: Energy consumption from natural gas (in billion Btu).
- petroleum_use: Energy consumption from petroleum products (in billion Btu).
- nuclear_use: Energy generated from nuclear sources (in billion Btu).
- renewables_use: Total renewable energy consumption (in billion Btu).
- biomass_use: Renewable energy from organic materials such as wood and waste (in billion Btu).
- geothermal_use: Renewable energy from geothermal sources (in billion Btu).
- hydro_use: Renewable energy from hydroelectric power (in billion Btu).
- solar_use: Renewable energy from solar power (in billion Btu).
- wind_use: Renewable energy from wind power (in billion Btu).
- residential_use: Total residential-sector energy consumption (in billion Btu).
- commercial_use: Total commercial-sector energy consumption (in billion Btu).
- industrial_use: Total industrial-sector energy consumption (in billion Btu).
- transportation_use: Total transportation-sector energy consumption (in billion Btu).
- total_consumption: Total primary energy consumption across all sectors (in billion Btu).
- real_gdp: Inflation-adjusted gross domestic product of the state (in millions of chained dollars).
- total_co2: Total CO_2 emissions from energy consumption (in million metric tons).
- carbon_intensity: CO_2 emissions / total energy consumption (w/o interstate flow of electricity, metric tons CO_2 per billion Btu)
- rps: Renewable Portfolio Standard (% of electricity that must come from renewable sources).
- population: Total state population for the given year.
- total_energy: Total energy production or availability measure (billion Btu)

Data Cleaning

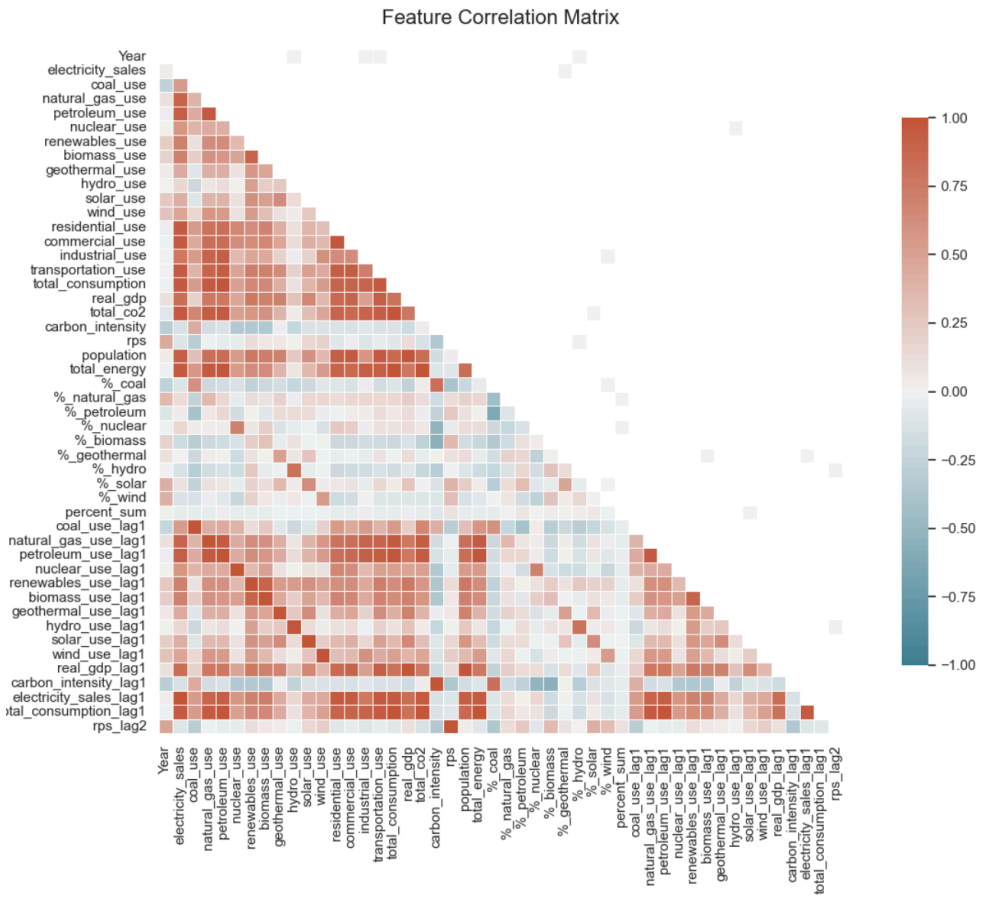
All the features were found in existing excel files, except for rps (this dataset required research to manually create). For all excel files, unnecessary columns were dropped and the dataset was formatted so that for each year (2000-2023), there was one row for each of the 50 states.

Through feature engineering, I created columns to represent the percent breakdown of each energy (energy source use / total energy) and lagged columns for each energy source, GDP, carbon intensity, total consumption, electricity sales, and rps.

An inner merge was performed on the State column to join all the individual dataframes, resulting in a dataframe containing 48 columns and 1200. Only the lag columns contained NaN entries.

Exploratory Data Analysis

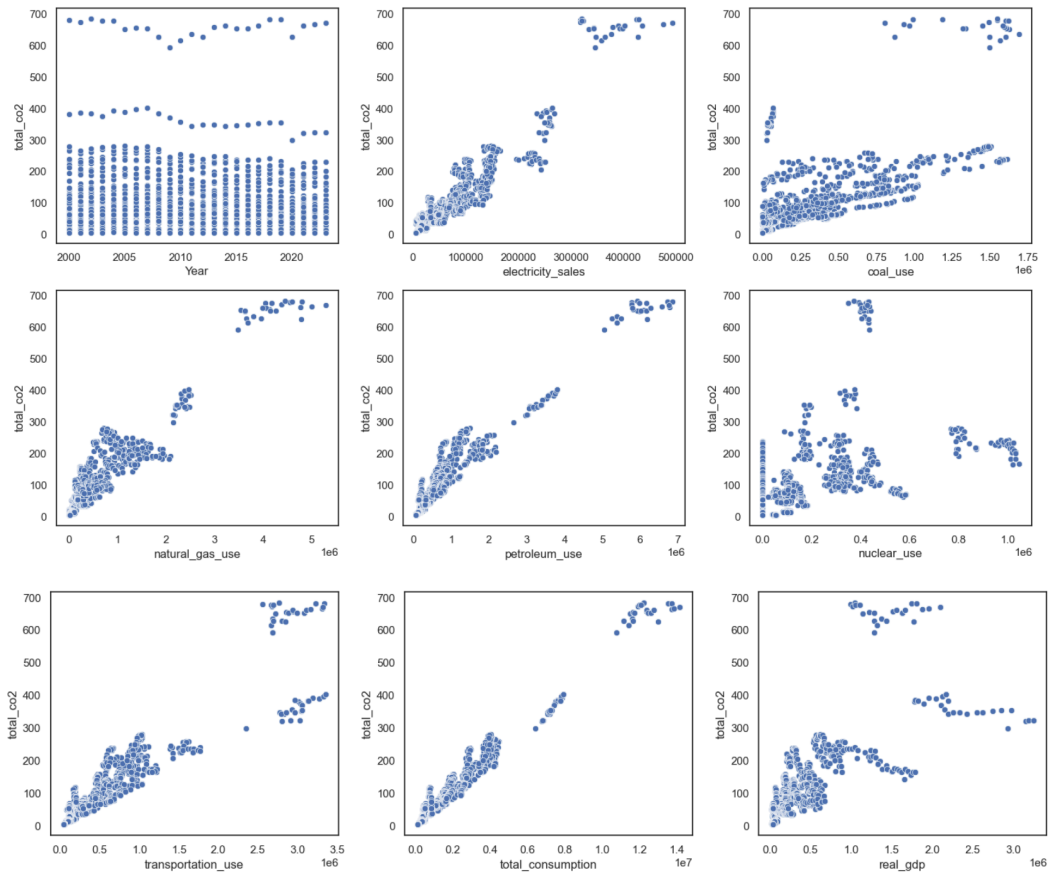
Based on the Figure 1), it is evident that there is a wide range of correlation values. Because it is possible for a feature to have a strong non-linear relationship with co2 emissions, correlation is not a reliable metric for identifying the features with the most predictive power. This analysis, however, is useful in helping determine which features are most correlated with one another and to assess the risk of multicollinearity, as highly correlated features may increase variance in the predictions.

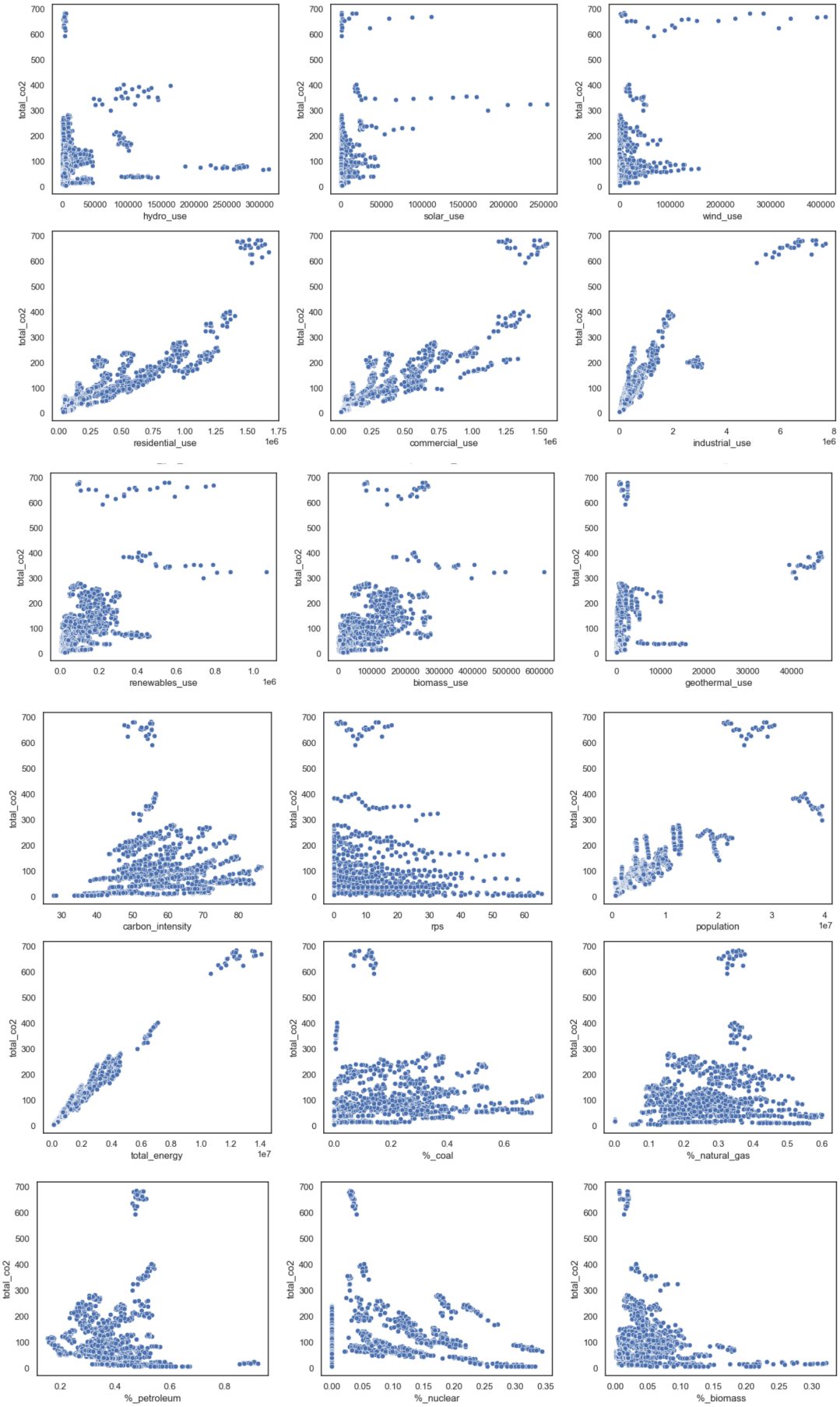


	Feature1	Feature2	Correlation	Category
52	electricity_sales	petroleum_use	0.92	Strong Pos
60	electricity_sales	residential_use	0.95	Strong Pos
63	electricity_sales	transportation_use	0.95	Strong Pos
64	electricity_sales	total_consumption	0.96	Strong Pos
66	electricity_sales	total_co2	0.94	Strong Pos
...	...	...	...	...
2182	electricity_sales_lag1	total_energy	0.95	Strong Pos
2195	electricity_sales_lag1	petroleum_use_lag1	0.91	Strong Pos
2206	electricity_sales_lag1	total_consumption_lag1	0.96	Strong Pos
2223	total_consumption_lag1	transportation_use	0.94	Strong Pos
2230	total_consumption_lag1	total_energy	0.99	Strong Pos

I used a for loop to create a dataframe that lists every highly positively / negatively correlated feature pair. The dataframe of strong positive correlations was 150 rows long, so out of 1600 possible feature pairs, there were 150 pairings with r > 0.9. There were not any feature pairs with a strong negative correlation. Upon a quick examination, some of these correlations make sense. For instance, we would expect that as energy use increases, that electricity sales may also increase, as some of the energy is being used for electricity.

To further understand the relationship between some of the predictors and the target, I created scatterplots for 27 of the features.




total_co2
hydro_use
solar_use
wind_use
residential_use
commercial_use
industrial_use
renewables_use
biomass_use
geothermal_use
carbon_intensity
rps
population
total_energy
%_coal
%_natural_gas
%_petroleum
%_nuclear
%_biomass

The scatterplot of electricity sales, coal use, transportation use, total consumption, residential use, commercial use, and industrial use all have a strong positive relationship with the target variable. All the renewable energy predictors follow the trend where the majority of the data points are clustered in the bottom left corner of the scatterplot (equates to low use and low emissions), with some points sparsely distributed in other locations of the plot. For instance, in geomass use and biomass use, there are some points that correspond to low energy generation and high emissions, which is contradictory to the result that may be expected. This observation can potentially be explained by some plants being more inefficient, in the case of geothermal, or burning plant material that is especially high in CO2 in the case of biomass use. Both solar and wind use have a horizontal line of data points that are sparsely distributed at a high benchmark of CO2 emissions, regardless of the energy generation. The scatterplot of industrial use appears slightly nonlinear; as energy production increases, the resulting co2 emissions appear to taper off. This can be attributed to facilities becoming more efficient in their energy use, or large facilities being more likely to incorporate renewable sources.

Model Building

XG Boost is appropriate for predicting CO2 emissions because:

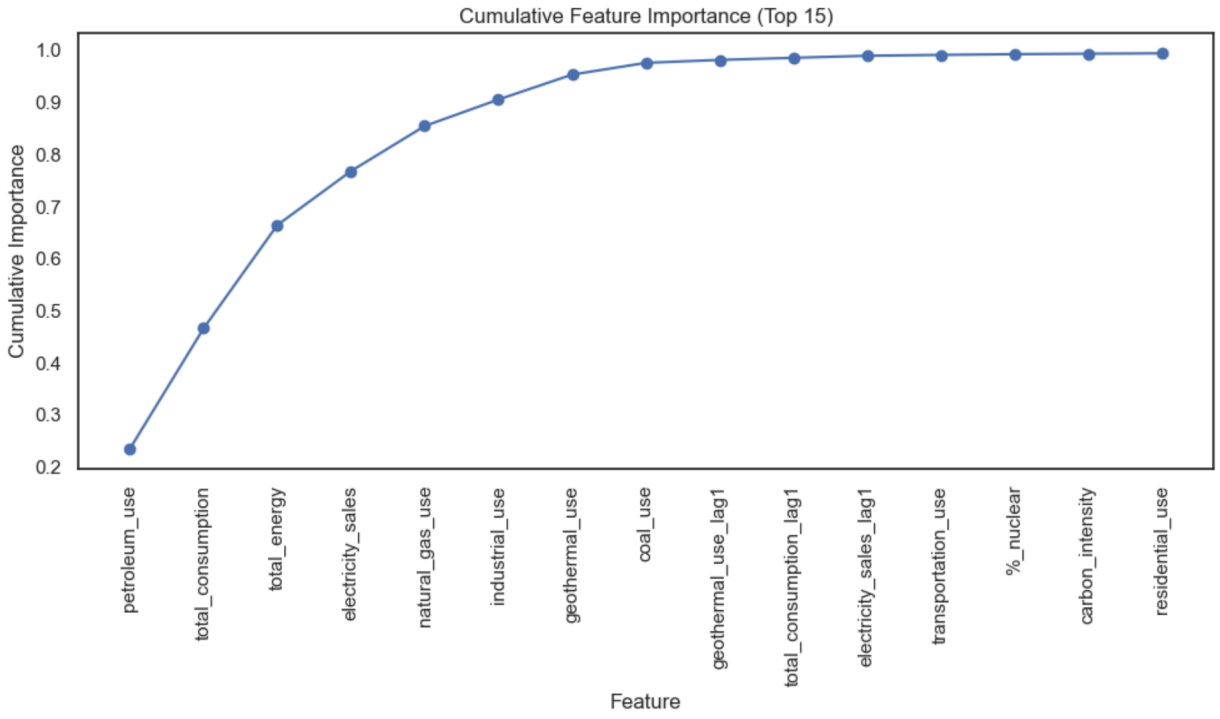
- It handles nonlinear relationships.
- It is robust to missing data: The lagged columns had a missing entry in 2000 for all states.
- It provides insight about feature importance: This information can be used for improved features selection later.

XGBoost Trial 1: All Features

The model was trained using a 80/20 train test split, so the first 80% of the data was used as training data. Because XG Boost doesn't require scaling, the data was left unscaled. To determine the optimal parameters, I employed three cross validation folds and used GridSearchCV to iterate over the following parameter combinations:

```
param_grid = {
        'n_estimators': [100, 200],              # number of trees
        'max_depth': [3, 5, 7],                  # number of splits
        'learning_rate': [0.01, 0.1, 0.2],       # parameter adjustment sizing
        'subsample': [0.8, 1],                   # fraction of training samples used per tree
        'colsample_bytree': [0.8, 1]             # fraction of features used per tree
}
```

From this iteration of GridSearchCV, the best hyperparameters were {colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.8} and the MSE was 36.85. XGBoost models have a built-in feature importances report. Plotting the contribution of the top 15 features to the model:



The top 15 features are able to explain a large proportion of the variance in the response. Given that a small subset of features hold the majority of the predictive power, what if we were to run GridSearchCV on only the top 25 most predictive features?

XGBoost Trial 2: Top 25 Features

From this trial, we obtained an MSE of 22.26. This led to a small improvement in the model, suggesting that by only using the most informative features, we are able to generate better predictions because we are supplying our model with less noise.

XGBoost Trial 3: Top 25 Features and Dropping 1 Feature from each Correlated Pair

This made the MSE significantly worse (MSE: 78.0). Dropping one feature from each correlated pair results in too many features being removed. Because XGBoost is resilient to correlated features and because our MSE is better when keeping all features in the model, retaining all numerical features may be the best option.

XGBoost Trial 4: Exclude the energy % columns, Retain All other Numerical Features

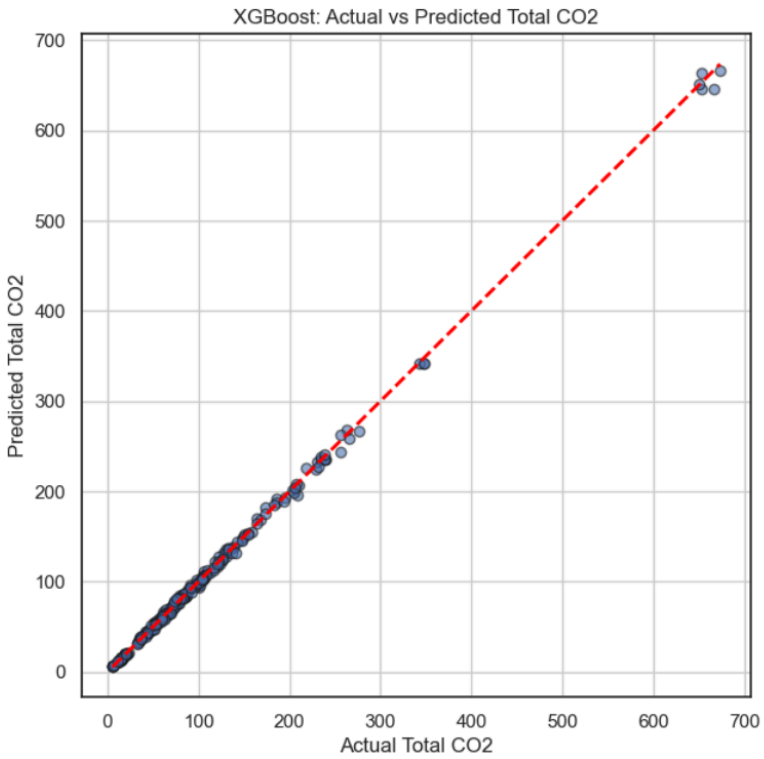
Based on the scatterplots of the energy % columns, these features don't have a strong relationship with the target variable and they are similar to the raw energy columns. What if we rerun GridSearchCV and exclude these predictors? This test resulted in a model with an MSE of 36.85. The result was the same as our first model. From GridSearchCV, the model that uses only the top 25 predictors performed the best.

Comparing all Models Using Test Data

GridSearchCV uses cross validation on the training data. Each fold contains training datapoints and a small validation set on which predictions are made, then the predictions are averaged across all validation sets for a given model. Prior to any rounds of GridSearchCV, we split the data into training and testing data. Examining the testing error on all models yields the following result:

Model	MSE	RMSE	MAE	R2
Full XGB (incl %)	14.493875	3.807082	2.482009	0.998702
Top 25 Features XGB (From full df)	13.325801	3.650452	2.500461	0.998806
Uncorrelated Features XGB (From full df)	30.842883	5.553637	3.875535	0.997237
Mini Feature XGB (No %)	11.781830	3.432467	2.387682	0.998945

Based on the test error, the model that excludes the percentage columns performed the best, followed by the model that only used the top 25 most predictive features. Once again, the model that removed one feature from each correlated pair had the highest error. Looking at error metrics of Mini Feature XGB, our test R^2 of 0.998 implies that 99.8% of the variance in the target variable can be explained by the predictors when using our features to generate predictions on unseen data. The test RMSE of 3.43 means that, on average, the model's predictions are about 3.43 units away from the true values. In our case, this means that the model is off by ~3.43 MMT on average for each prediction. We can put that value into scale by considering the relative root mean squared error, or rather the RMSE as a percentage of the mean of the target. The Relative RMSE 3.31% (model's typical error of the average target value). We will choose the Mini Feature XGB model for the next steps.



Dashboard

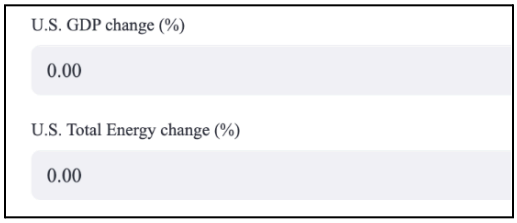
Objective

The objective of the dashboard is to allow users to simulate how changing the U.S's energy mix affects CO_2 emissions by creating a wrapper that harnesses the XGBoost model trained on historical U.S. energy, population, GDP, and policy data. Rather than predicting emissions for future years, the simulator runs a counterfactual analysis: it compares the actual CO2 emissions in 2023 to the user's provided scenario. It answers the question: How would our emissions have changed if we had, for instance, used 10% more energy or had 20% of our energy come from renewable sources?

How it Works

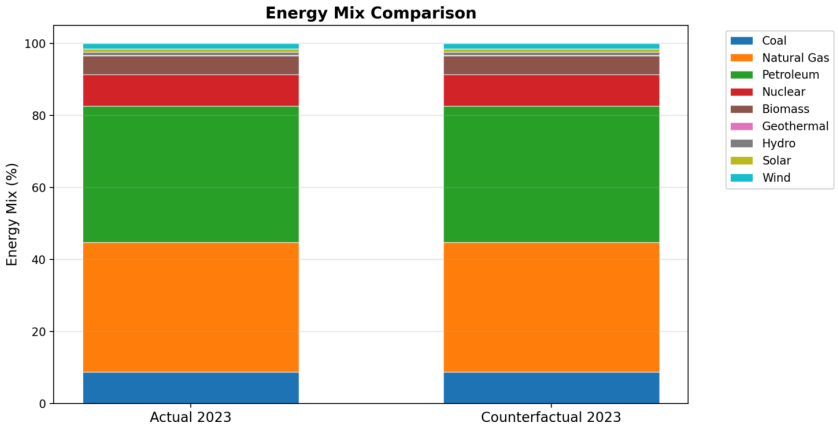
1. User specifies the % increase / decrease for feature(s)
2. Dashboard loads latest historical state data (2023)
3. The % change is used to scale the rows corresponding to 2023 for all states
4. For every state, the model predicts the emissions in 2023
5. State emissions are aggregated and compared to actual 2023 emissions

Without adjusting any features, the scenario is equivalent to the baseline 2023 data.



Energy Mix

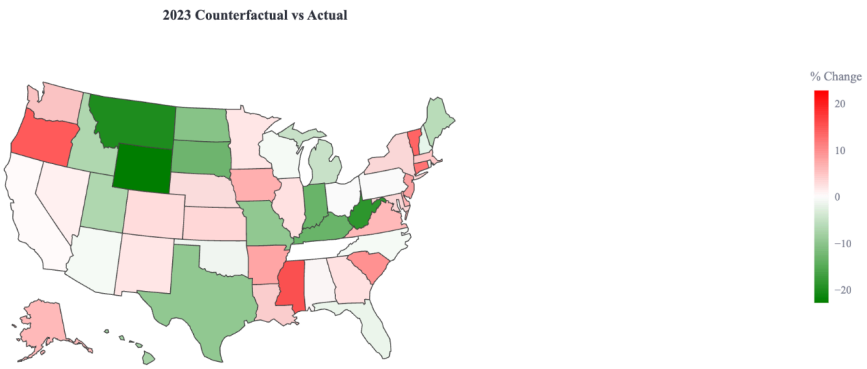
	Energy Source	Original (%)	Scenario (%)
0	Coal	8.74	8.74
1	Natural Gas	35.94	35.94
2	Petroleum	37.93	37.93
3	Nuclear	8.67	8.67
4	Biomass	5.22	5.22
5	Geothermal	0.13	0.13
6	Hydro	0.89	0.89
7	Solar	0.94	0.94
8	Wind	1.54	1.54



Emission Impact (USA)

Actual CO_2 (2023)	Counterfactual CO_2 (2023)	What-If Change
4,783 MMT	4,706 MMT	-76 MMT
	↓ -1.6%	

State CO_2 Emissions (Percent Change)



When the features are not altered, the model is aggregating predictions on the raw data for each state. The coloration that is seen in the map represents the model error. Even though a couple of states have a notable error, the predicted national emissions are similar to the actual emissions (only 1.6% error). Red implies that the predicted emissions exceed the actual emissions.

Challenges

Tools

- Python
  - Plotly
  - Pandas
  - NumPy
  - Matplotlib
  - Sklearn
- Excel
- Streamlit
- Jupyter Notebook