Modeling Global Health - Insights into Energy Usage and Management

Table of Contents

1. Key Points

I developed a Bayesian optimized, gradient boosted random forest model for predicting GDP per capita across 71 countries.

Evaluated with a 0.96 adjusted R2, MAE of 2600.

Use cases: scenario modeling, planning, education.

2. Motivation

Sustainability, clean energy, and reduced emissions are all crucial topics in the modern world. We want to be healthier and adopt better means of living.

How can we take a data-centric approach to this task?

To answer this question, I investigated developing a model to predict a population's health based on energy related inputs from the World Bank Group.

3. Data

Data comes from a Kaggle dataset scraped by Ansh Tanwar.

Sourced from the World Bank Group with a CC 4.0 license.

The World Bank Group gets their data through surveys, reports, and statistical systems of countries.

Their mission is to "help countries share and apply innovative knowledge and solutions to the challenges they face."

The original CSV is downloaded with 3,649 entries x 21 columns, with 6,978 total missing values.

Features pertain to energy usage, outputs, co2 emissions, and geographical attributes of 176 countries worldwide.

Examples include:

- Electricity access %
- Renewable Energy Output
- Density
- Land Area

I chose to use **GDP per capita** as the predicator and indication of a country's health. Accounts for country economic prosperity and population size.

= GDP of Country / Total Population

Not a perfect measure, as income distribution, QoL, and other factors aren't included.

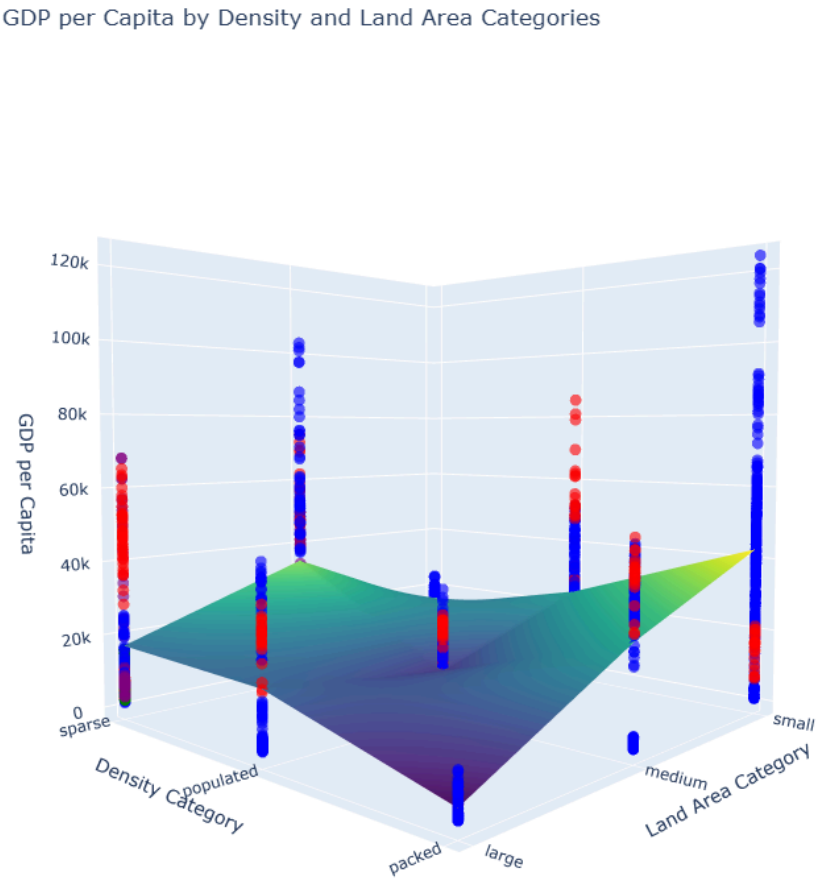
Dataset was transformed from 3,649 x 21 entries with 6,978 missing values to 1,491 x 22 entries with 0 missing values.

Achieve the highest level quality of data possible while still retaining significant volume.

Augmented co2 emissions per capita for Egypt, Slovakia, Turkey from World Bank Database. Listed as Arab Republic, Turkiye, and Slovak Republic.
Imputed with 0's for electricity from nuclear output for Malaysia, Saudi Arabia, Chile, Indonesia, Kazakhstan as these are true 0's without nuclear programs established.
Added land area category using 3 intervals of small, medium, large
Added density category usingf 3 intervals of sparse, populated, and packed
Added quadrants based on lat / lon paring as NW, NE, SE, SW

4. **EDA**

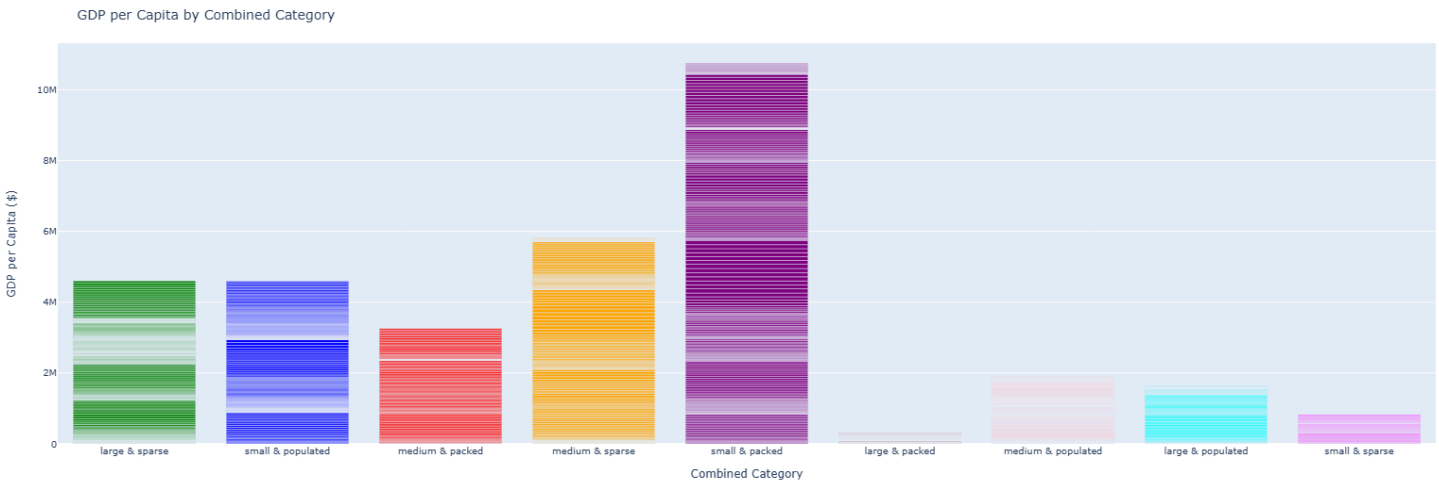
We start with a surface plot to visualize land area, density, and quadrants. The more elevated peaks represent countries with higher GDP per capita, while the valleys represent countries with lower GDP per capita.



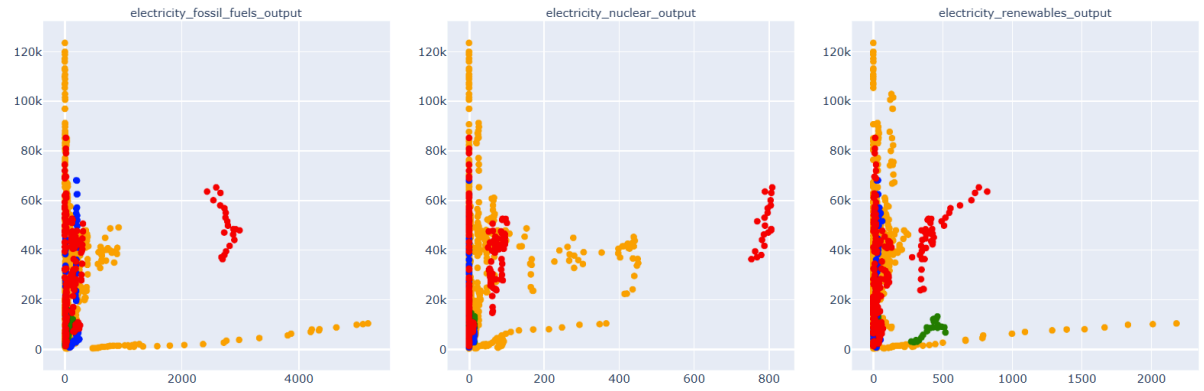
Small and packed countries (Luxembourg, Switzerland, Qatar) are more compact, allowing for more efficient energy delivery and infrastructure. These are found on the highly elevated peak on the far right.

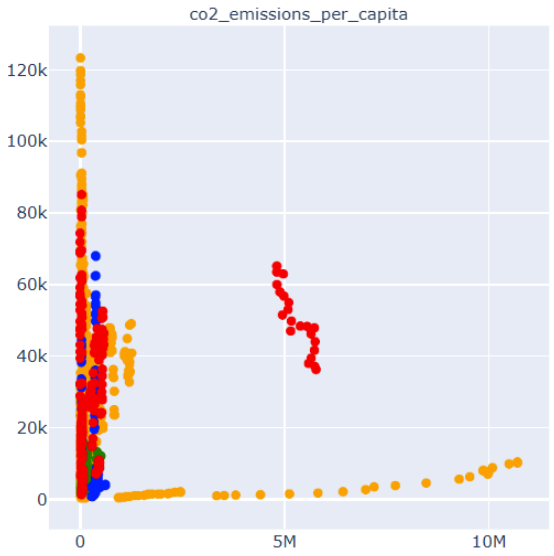
Large and packed countries (China, Pakistan, India, Thailand) are overburdened with energy demands, lacking infrastructure to meet them. These can be seen on the front most valley.

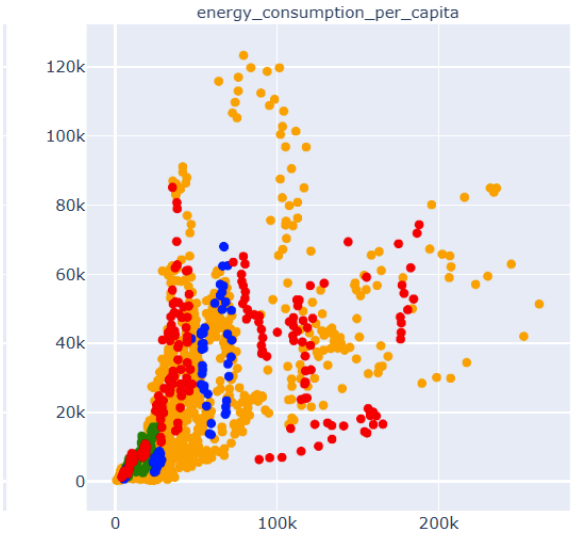
Let's take a closer look at our GDP per capitas based on land area and density.



There's a clear discrepancy between small and packed countries (dark purple) against large and packed countries (very light pink). Let's look at some of our features to try and answer this. Here are some of the most notable features in comparison to GDP per capita.







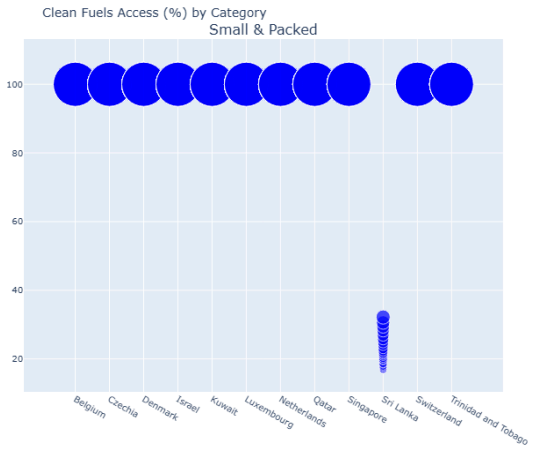
These are mostly nonlinear relationships. Energy consumption per capita was the most linear, and even then relatively weak. The outlier seen are the US and China – we will take a look at these later.

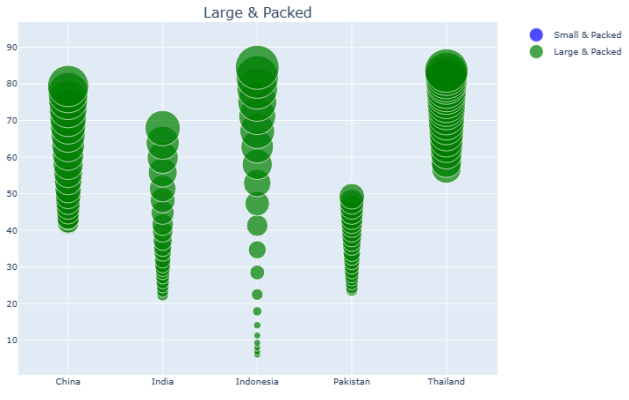
Looking into spearman correlations, the 3 strongest correlating features with GDP per capita are (with next highest score of (0.40).

1. Clean fuels access % (0.81)
2. Energy Consumption per capita (0.81)
3. Electricity Access % (0.71)

How do these features look when visualized for small versus large counties?

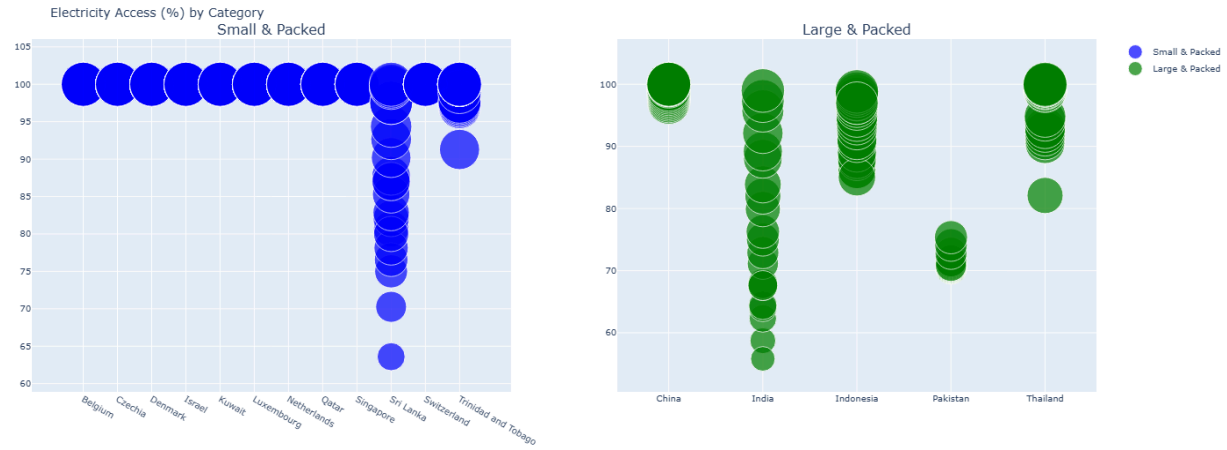
Small, packed, vs large, packed for cleans fuels access %





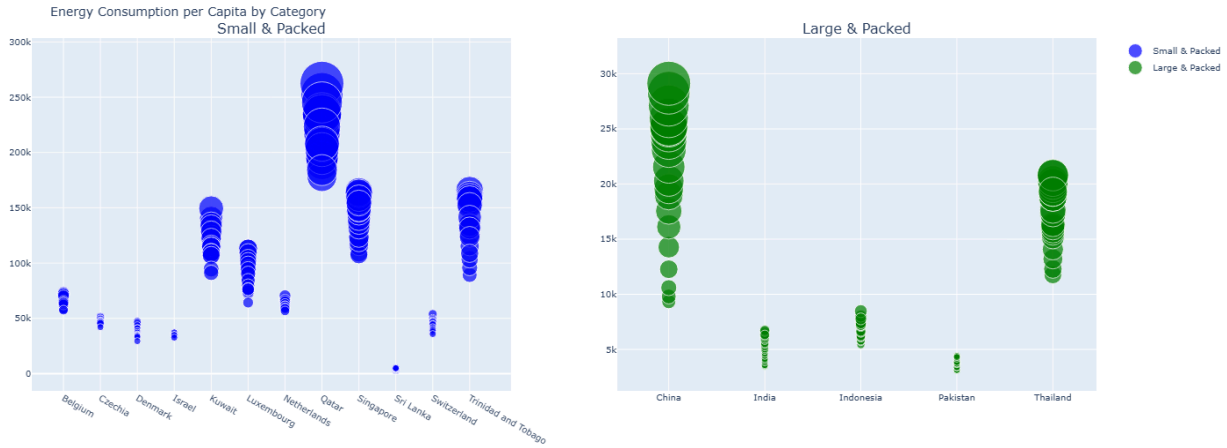
Apart from Sri Lanka, small & packed countries are able to provide clean fuel to essentially 100% of their population, in comparison to large & packed countries who are lagging behind.

Small, Packed vs Large, Packed – Electricity Access %



Again, besides Sri Lanka, small and packed countries provide electricity to almost all of their population. Large and packed countries lag behind here too.

Small, Packed vs Large, Packed – Energy Consumption per Capita



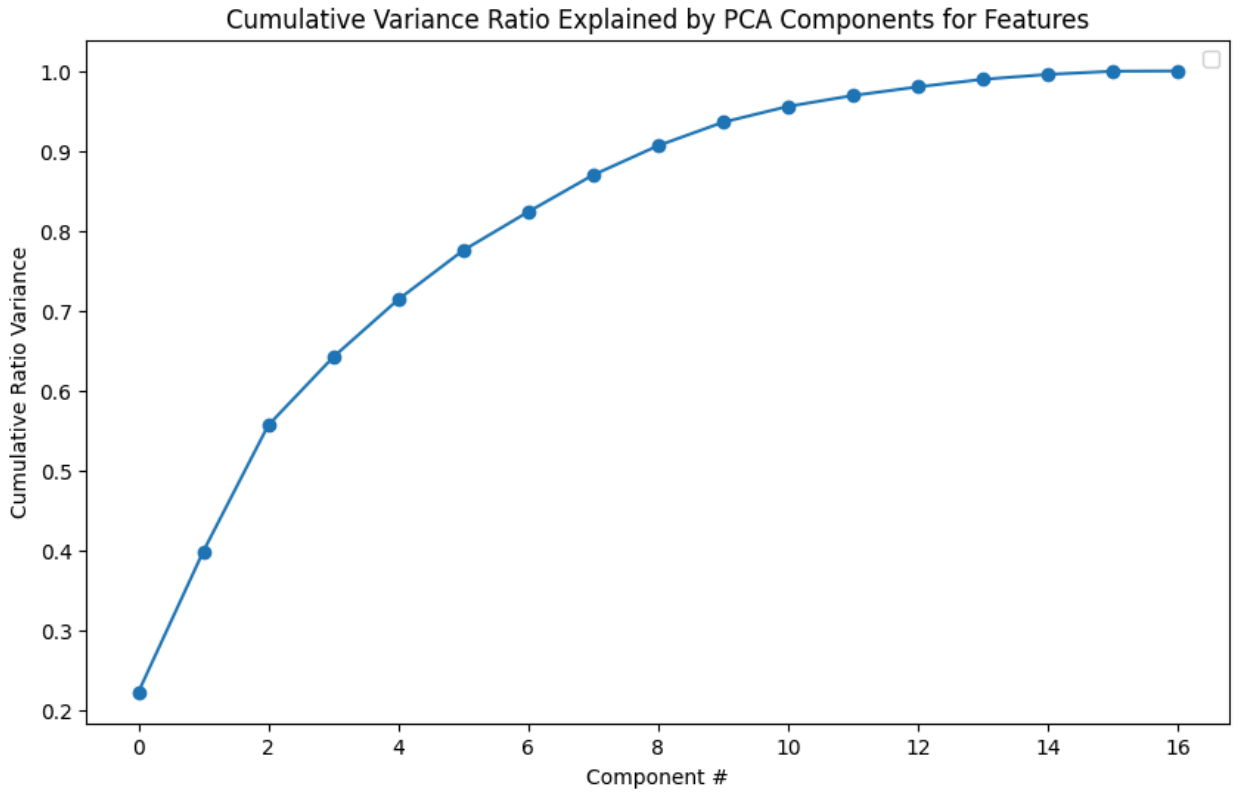
The visual discrepancy is much smaller here.

From looking into these features, we can come up with some takeaways.

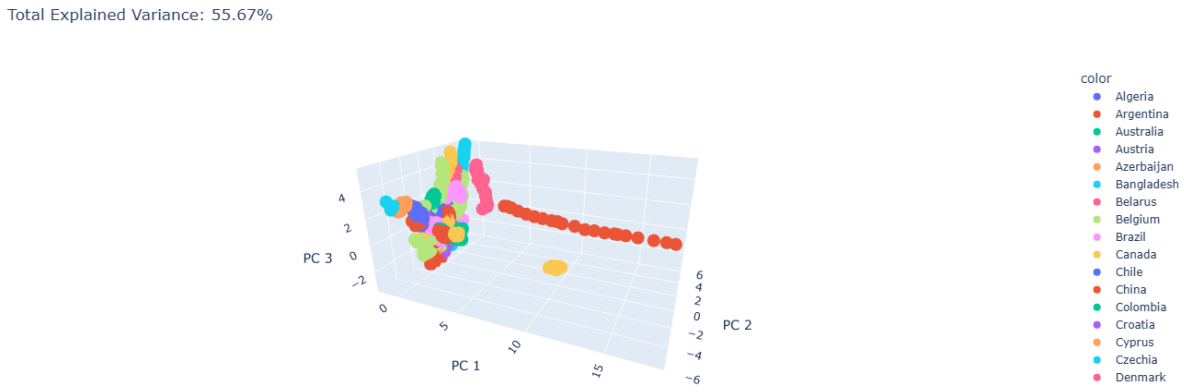
1. Resource distribution grid can be made much more efficient for small, packed, enabling better clean fuel and electricity dist.

2. Strong link for clean fuel and electricity access to gdp per capita – healthy environment for individuals to grow up in, become educated, contribute to GDP.

Moving on to a principal component analysis.



I choose component #2 as my elbow point. I then plot the countries against these 3 components and notice something interesting.



USA (line of red) and China (yellow clump) stand out from the rest of the group, especially in PC1. This is in agreement with what we saw from those feature correlation charts above.

Why could this be? When looking at PC1, our strongest loaders are

1. Co2 emissions per capita (0.48)
2. Electricity fossil fuels output (0.48)
3. Electricity renewables output (0.46)
4. Land area (0.40)
5. Electricity nuclear output (0.36)

USA and China are the 2 world leaders when it comes to manufacturing output in billions but are low to middling when it comes to gdp per capita. Such a high demand on the energy grids of the country separates them visually in our principal component analysis. Perhaps this is due to distribution of manufacturing responsibilities - areas in the country responsible for manufacturing generate jobs and stimulate the economy, but those areas might not be dispersed evenly throughout the country.

```
energy['electricity_fossil_fuels_consumption_ratio'] = energy['electricity_fossil_fuels_output'] / energy['energy_consumption_per_capita']
energy['electricity_renewables_consumption_ratio'] = energy['electricity_renewables_output'] / energy['energy_consumption_per_capita']
energy['co2_emissions_consumption_ratio'] = energy['co2_emissions_per_capita'] / energy['energy_consumption_per_capita']
energy['land_area_with_electricity_access'] = energy['land_area'] * energy['electricity_access_%']
energy['land_area_with_clean_fuels_access'] = energy['land_area'] * energy['clean_fuels_access_%']
```

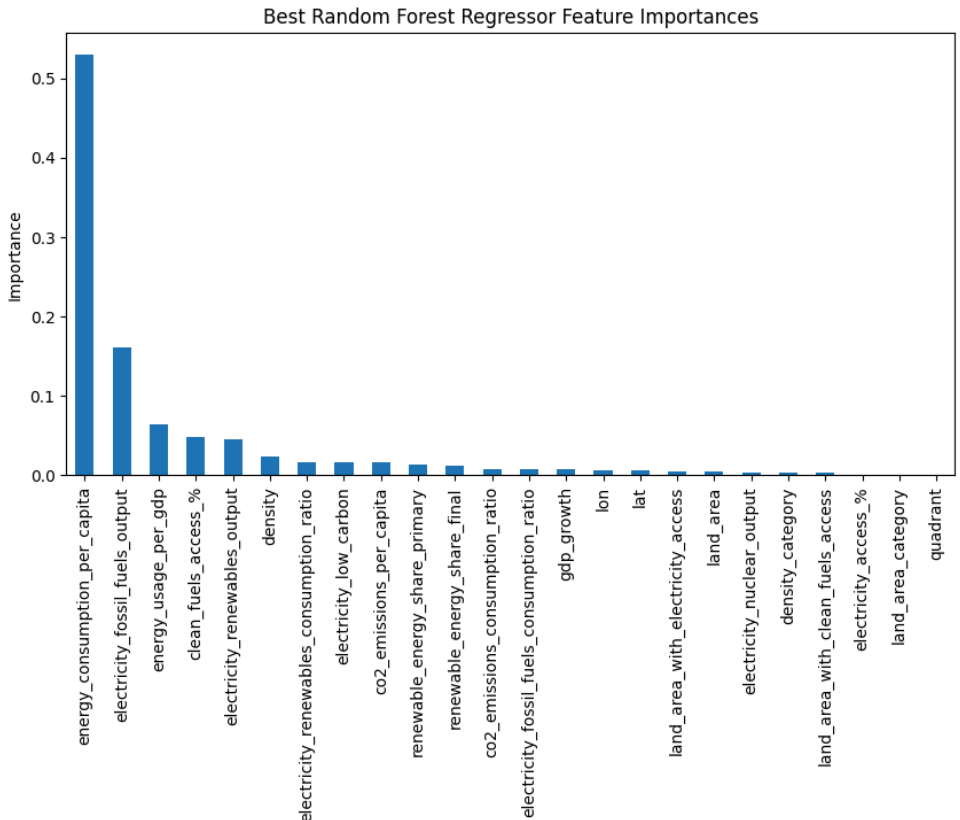
From our 3 strongest correlations with GDP per capita, and our 5 strongest PC1 loaders, I engineer new features to enhance the dataset.

5. Modeling

For preprocessing, I go with ordinal encoding for the non numeric columns and standardize with StdScalar().

For selecting a model, I knew feature relationships were mainly non linear. Therefore, I decided on Random Forest, which is robust to outliers and noise in the data, which we saw in the EDA - clear patterns and distinctions are not always present.

For our base Random Forest with grid CV, Energy consumption per capita shows as the strongest feature, which ties into seeing clean fuels and electricity access as the strongest correlation to gdp per capita.



Adjusted r2 of 0.93 and MAE of 3172.

Best params:

- n_estimators = 33

I then opt for Random Forest with BayesSearch CV, to offer a more effective exploration into the hyperparameter space.

Best params:

- n_estimators = 114
- max_depth = 14
- max_features = 0.362
- min_samples_leaf = 1
- min_samples_split = 2

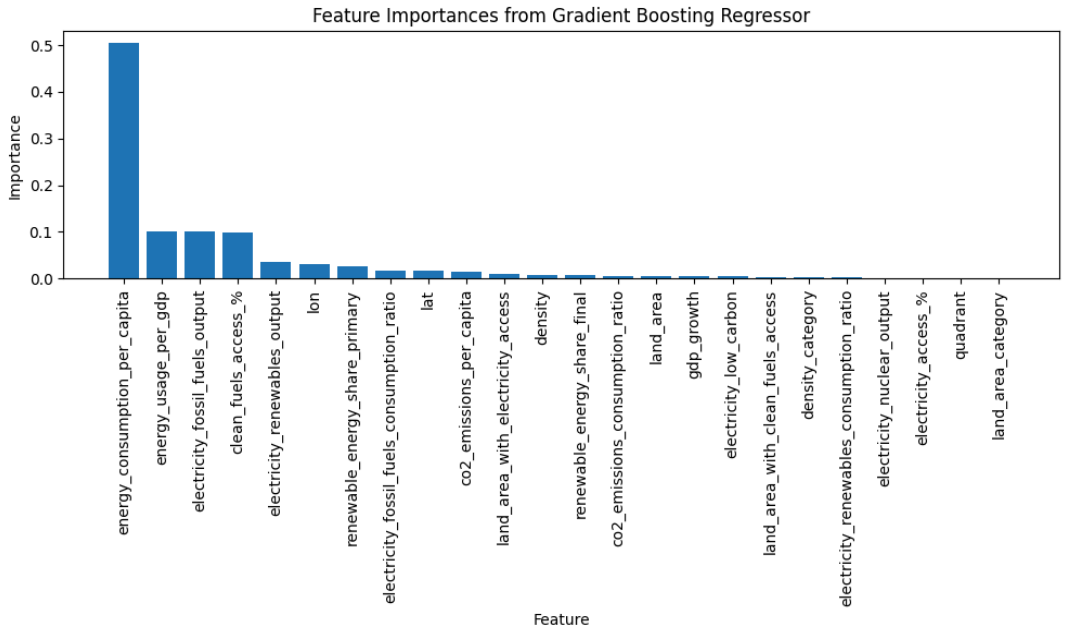
It gives an adjusted r2 of 0.955 and MAE of 2846, improvements from before.

Finally, I enhance the model further by stacking with GB to compliment Random Forest. GB offers bias reduction, and random forest offers variance control.

Again, we see energy consumption per capita as the strongest feature.

We get an adjusted r2 of 0.96, with a MAE of 2600. These are the best performances.

Hyperparameter tuning on the GB model gives diminishing returns, making this a good stopping point.



6. Next Steps

Improve feature engineering around energy consumption per capita – strong importance in both RF and GB. Make sure to not introduce collinearity into the set.

Potential hyperparameter tuning advancements within GB – without overusing computation power for minimal gain.

Productionize the final model into a web app which can allow the user to make changes to energy values and see the effect on the target.

For example – how would GDP per capita for the US change if co2 emissions are reduced by 10%, 15%, 20%, etc?

7. Conclusions

I was able to develop a good fitting model to predict gdp per capita based on our dataset.

Points to several areas of exploration – geophysical vs economic relationships in countries, energy grid management, distribution of wealth and resources.

How to model feature relationships for a simulation? I would need to consider the relationships between features, along with relationships to target.