Environmental Performance Index Review

By: Ben, Dan & Ryan

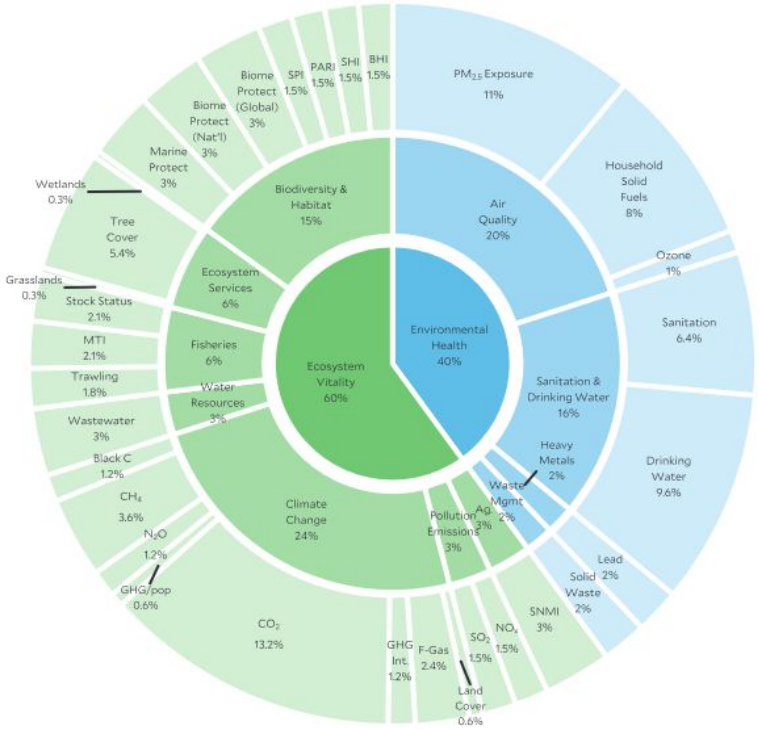
Problem Statement:

The United Nations has become aware of the Environmental Performance Index (EPI), a score calculated by a research team at Yale to rate countries on their perceived environmental performance. They have contracted us to audit the EPI and determine if can be useful in future endeavors, or if it contains meaningful data.

Approach:

We acquired the data that was used to calculate EPI scores. Using this data we attacked the problem from several different angles. Models were trained on EPI data for each year as well as aggregated EPI data based on country group, with the goal of predicting EPI score, based on the same inputs of the original EPI calculation. The models' feature importances and coefficients were compared with the EPI's published feature breakdown.

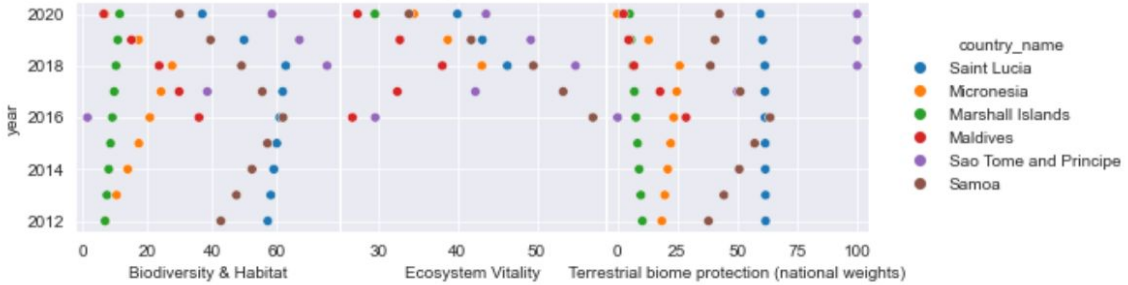
What is the Environmental Performance Index?



The Environmental Performance Index is a method of quantifying and numerically marking the environmental performance of a state's policies. This index was developed from the Pilot Environmental Performance Index, first published in 2002, and designed to supplement the environmental targets set forth in the United Nations Millennium Development Goals.

Data cleaning and imputing:

The data was highly varied and had frequent nulls, so a variety of imputing techniques were used, including:



- Filling nulls with the regional average where no other data was present.
- Filling nulls with a linear step imputation.
- Padding outside nulls with the last known value.

- Dropping a small collection of island nations which had both high null count and high variance.
- Dropping columns that were largely composed of null values.
- Identifying landlocked countries and setting their marine-based features to a fixed value.

EPI Tools

Random Forest Model: Providing measurements of all countries across the globe from the dataset. Showing 10 indicators of environmental public health and ecosystem vitality.

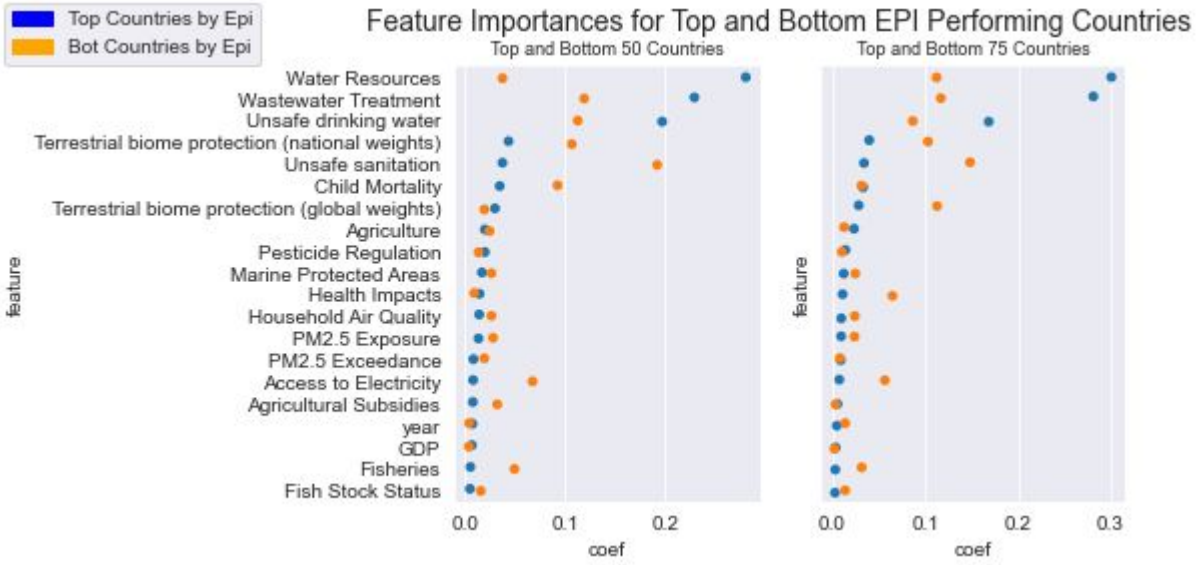
- ★ Test Score: .98
- ★ Train Score: .99
- ★ Best Score: .97

Feature Importances:

According to the model, unsafe drinking water has emerged as the leading component of feature importance.

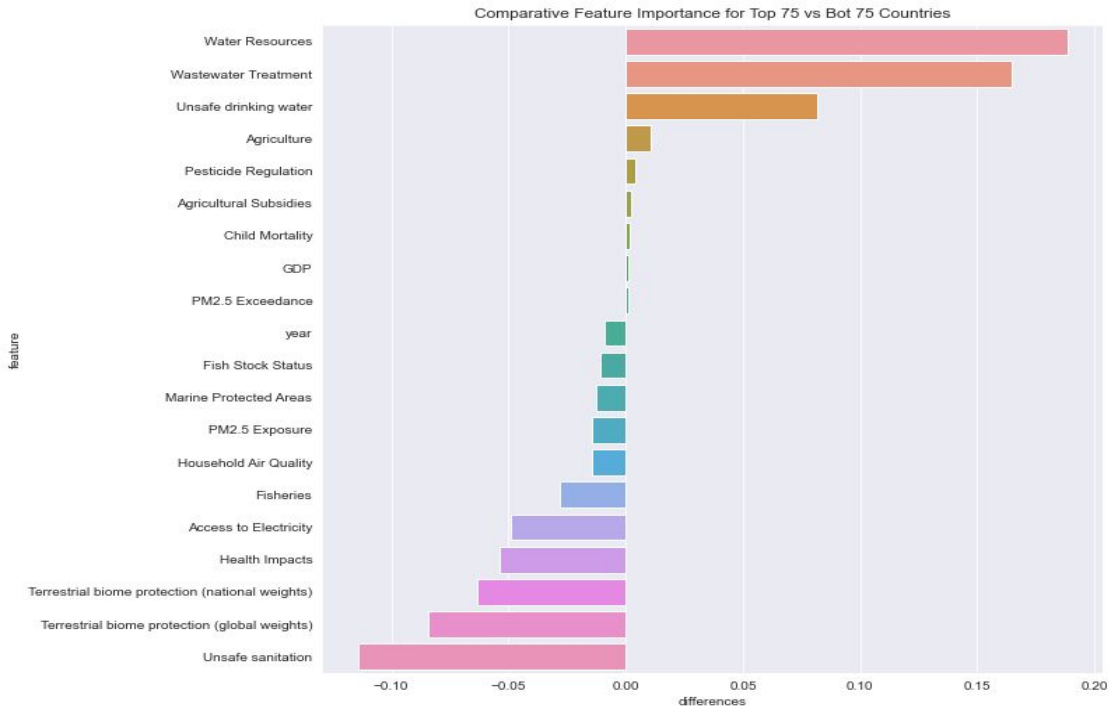
	Column Name	Feature Importance
2	Unsafe drinking water	0.180835
5	Wastewater Treatment	0.165914
6	Water Resources	0.131360
9	Unsafe sanitation	0.092390
16	Health Impacts	0.083702
14	Household Air Quality	0.063345
18	Child Mortality	0.057073
15	Access to Electricity	0.051963
10	Terrestrial biome protection (global weights)	0.032824
8	Terrestrial biome protection (national weights)	0.030143

Difference in Feature Importance Top performers vs bottom performers:



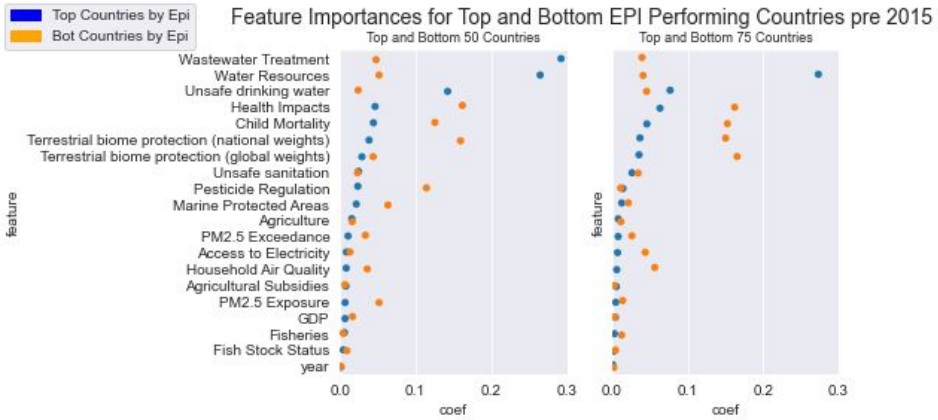
- All variables but GDP are on a relatively even scale.
- Poor Performing Countries feature importance rely on unsafe sanitation and the Terrestrial Biome Protections
- Top Performing Countries feature importances rely more Water resources and Wastewater Treatment

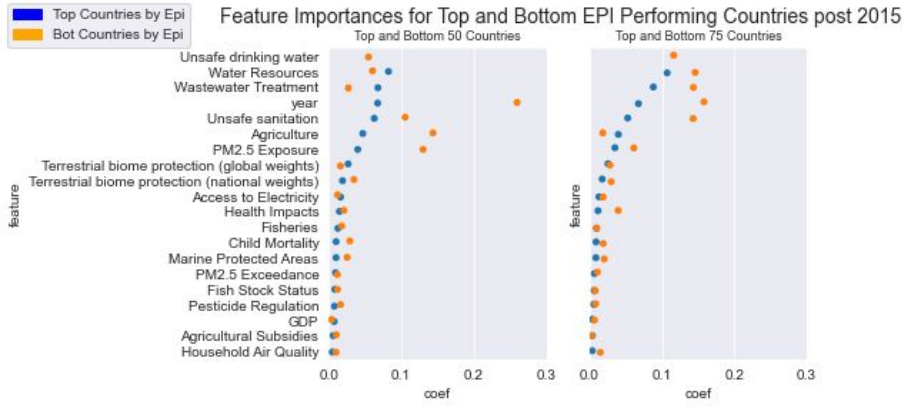
Difference in Feature Importance Top performers vs bottom performers:



- Difference graph confirms the eye test from the previous graph about the differences between High and Low performers.
- GDP is notably unimportant for determining EPI among the two groups.

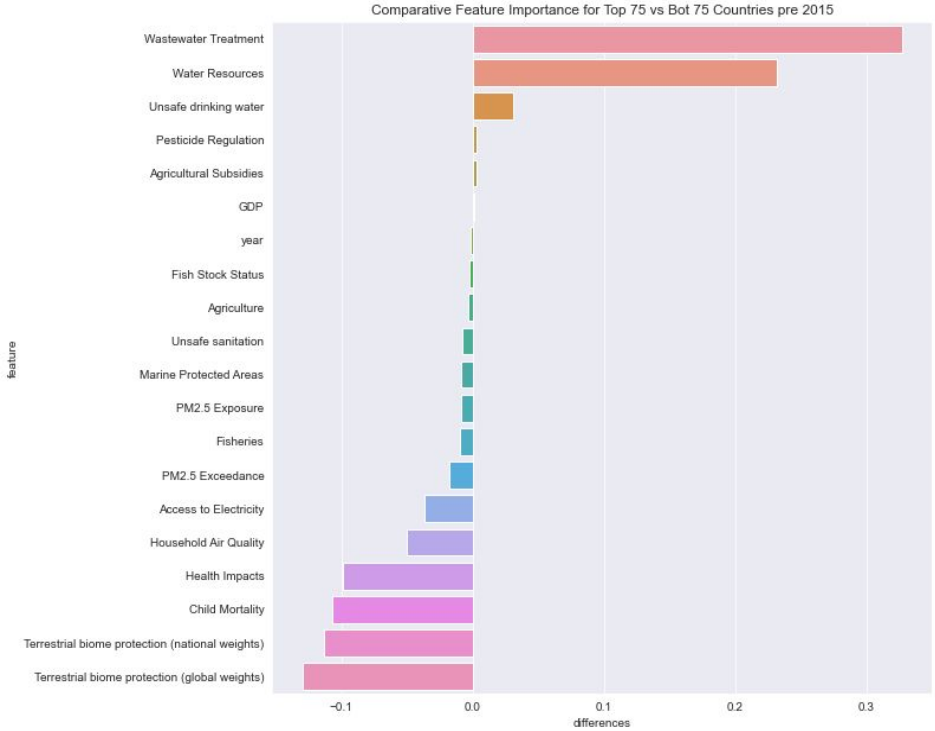
Difference in Feature Importance Pre and Post 2015

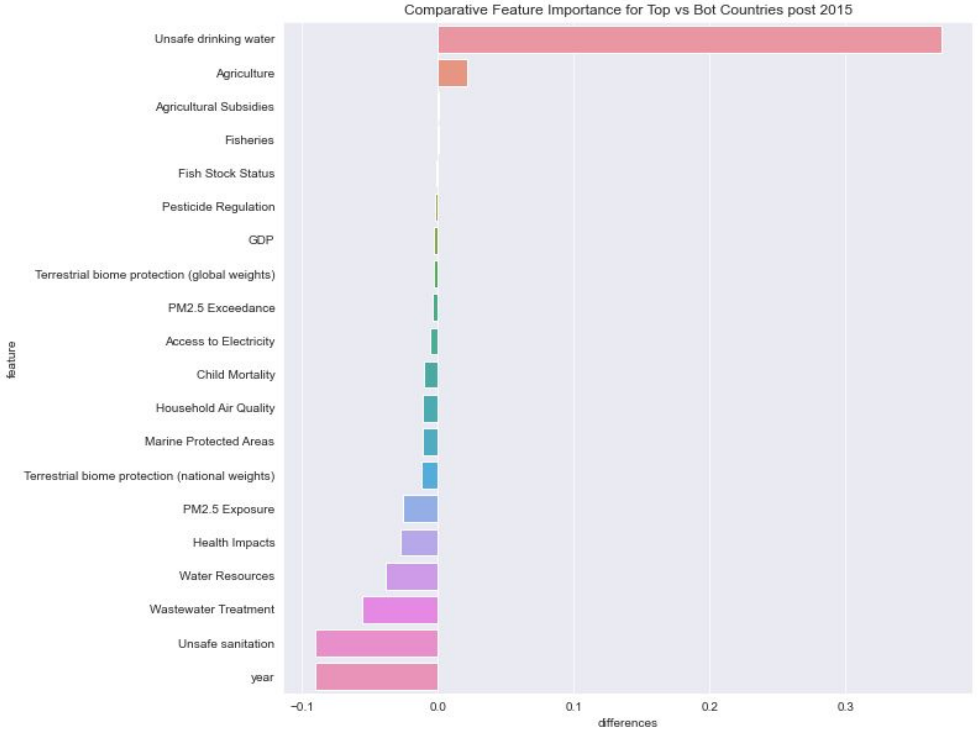




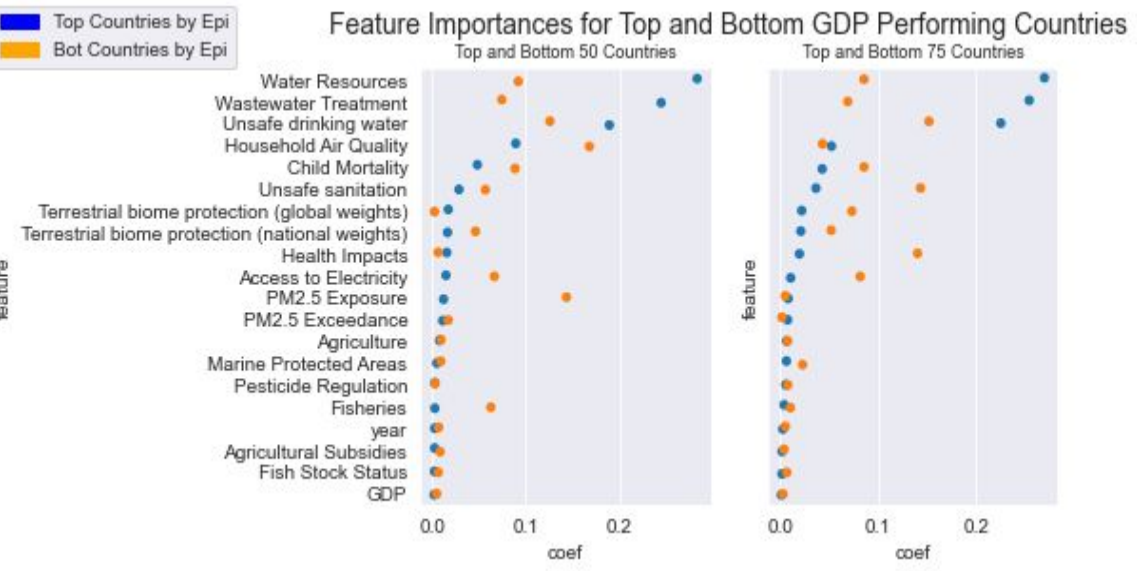
- As you can see on the left, before the big re-definition of the EPI calculation, the features show higher variance.

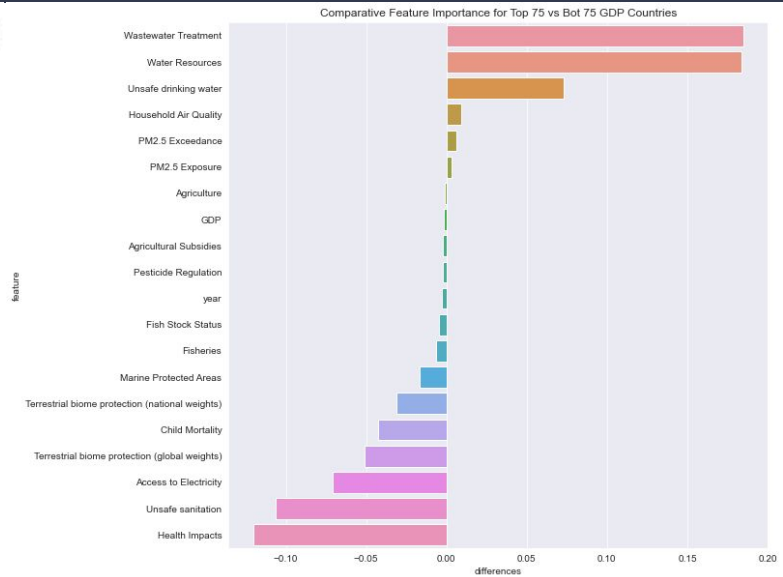
Difference in Feature Importance Pre and Post 2015





Difference in feature importance by top and bottom GDP countries



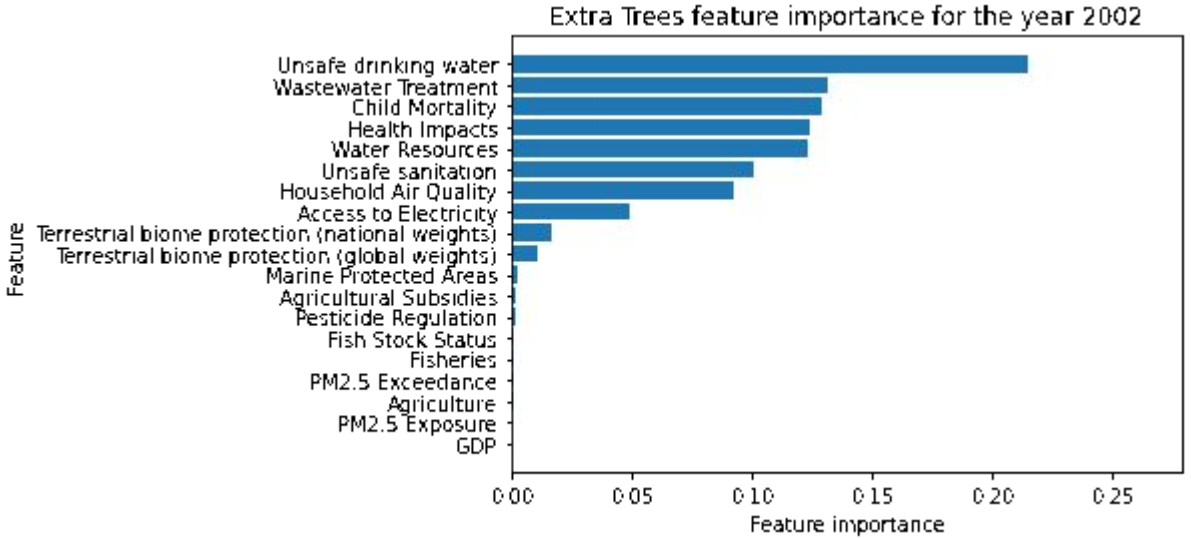


- The same few culprits again find themselves at the top of the feature importance list.

Change of Feature Importance over time.

When predicting the EPI of a country, our Extra Trees model consistently placed high importance on Water, Sanitation, and Health features.

This contrasts heavily with the EPI 2020 calculation, which features a 16% weight for Water & Sanitation an 8% weight for Household Air Quality.



Conclusions

Based on our modeling of the data, we do not believe the U.N. should use EPI as a metric for informing recommendations.

From the year-by-year analysis, we conclude that the current weighting of the EPI is not representative of the factors that can most accurately predict its value. Furthermore, the calculation of the EPI changes every other year, but the features that are determined most important by our models maintain their status over the 19 year period.

From the top country and top GDP analysis there seems to be a need to re-address the value of many of the weights that are used when calculating EPI. Water Resources, Wastewater Treatment, and Unsafe drinking water all seem to have a much bigger impact on top countries EPI scores than they do on bottom countries, whereas bottom countries are identified by the model by their low Child Mortality, Access to electricity, and Household Air Quality.

Next Steps:

- A more in depth analysis on the features used in calculation EPI and decide which features need to have adjusted weights.
- Consider looking into other metrics of environmental performance, to use instead of, or in tandem with the EPI.
- Build a performance index that relies more heavily on environmental data instead of policy.
- Stronger data collection, to avoid having so much missing values in the data.

Possible Improvements:

- Add more external features to the data set in order to look for more confounding, or predictive variable.
- More in depth look into the GDP feature and how it is affected by/effects the data.
- Compare EPI to other similar indexes and see how different types of scoring and weights are used for different metrics.
- Removal of metrics that are bias towards industrialized countries and are not good indicators of environmental health.

Credits

https://epi.yale.edu/

https://en.wikipedia.org/wiki/Random_forest