Analysis of Chronic Hunger

Leon Kotze, October 2018

Executive Summary

This document presents an analysis of data concerning Chronic Hunger, or more specifically the Prevalence of Undernourishment (PoU). The data is provided by the Food and Agriculture Organization of the United Nations (FAO) and consists of various socioeconomic statistics by country.

The FAO defines PoU as:

> *An estimation of the probability that a randomly chosen individual in the population regularly consumes less food than his dietary energy required to live a healthy active life.*

PoU is a statistical model and represents an estimation of how likely individuals are to suffer from chronic hunger. It is a complex metric and as such difficult to calculate, hence the desire to estimate PoU from more readily available statistics.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between socioeconomic indicators and PoU were identified. After exploring the data, a regression model to predict PoU from other socioeconomic statistics was created.

While many features help to indicate PoU, significant statistics found in this analysis were:

- *access_to_improved_water_sources* - Percent of the population with reasonable access to an adequate amount of water from an improved source. From the initial analysis, lack of access to improved water sources was the strongest indicator or chronic hunger.
- *gross_domestic_product_per_capita_ppp* - GDP per capita is often used as proxy for average income levels in a country and having it at purchasing power parity (PPP) allows is to be comparable across countries. Stands to reason that higher GDP per capita should indicate lower levels of chronic hunger.
- *access_to_electricity* - Percent of population with access to electricity.
- *caloric_energy_from_cereals_roots_tubers* - Percent of total dietary energy supply coming from cereals, roots and tubers. Strong indicator of the availability of food.
- *obesity_prevalence* - Percent of adults ages 18 and over whose Body Mass Index is more than 30 kg/m2. Countries with high levels of obesity are unlikely to have a high level of chronic hunger.
- *avg_supply_of_protein_of_animal_origin* - Average protein supply expressed in grams per capita per day. Strong indicator of the availability of food.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

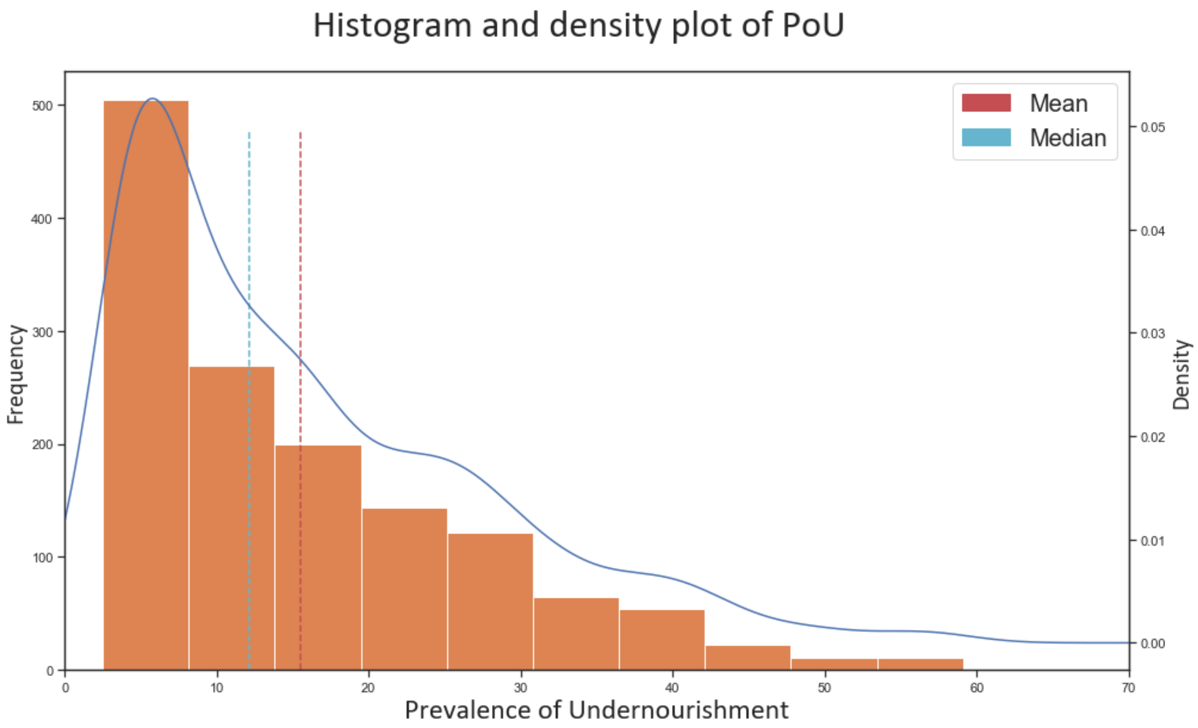
Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, distinct count, and % missing values were created for numeric features, and the result from the 1401 observations are shown here:

Feature	Min	Max	Mean	Median	Std Dev	Count	% NaN
access_to_electricity	0.01	102.00	73.80	89.16	31.28	1397	0.29
access_to_improved_sanitation	10.34	101.75	65.05	73.47	28.42	1327	5.28
access_to_improved_water_sources	30.78	101.97	83.30	88.44	15.28	1339	4.43
adult_literacy_rate	24.14	100.46	79.63	87.03	18.23	285	79.66
agricultural_land_area	2.94	10.46 M	353.96 k	47.02 k	1.17 M	1385	1.14
anemia_prevalence	12.57	69.62	32.78	30.11	12.00	1321	5.71
avg_supply_of_protein_of_animal_origin	2.96	83.21	27.96	25.15	15.98	1149	17.99
avg_value_of_food_production	3.95	1.04 k	229.47	205.29	149.06	1234	11.92
caloric_energy_from_cereals_roots_tubers	22.59	84.39	50.89	50.31	13.93	1149	17.99
cereal_import_dependency_ratio	-228.30	101.98	34.37	35.04	51.94	1084	22.63
cereal_yield	179.26	27.98 k	2.75 k	2.22 k	2.78 k	1337	4.57
co2_emissions	100.83	2.27 M	83.05 k	7.64 k	224.84 k	1317	6.00
droughts_floods_extreme_temps	0.00	9.18	1.24	0.66	1.88	75	94.65
fertility_rate	0.84	7.54	3.25	2.75	1.47	1387	1.00
food_imports_as_share_of_merch_exports	0.99	763.38	37.15	16.18	66.56	1148	18.06
forest_area	9.81	8.24 M	232.95 k	22.24 k	926.63 k	1385	1.14
gross_domestic_product_per_capita_ppp	573.17	137.95 k	10.84 k	6.96 k	15.28 k	1362	2.78
hiv_incidence	0.01	4.27	0.22	0.04	0.52	1030	26.48
imports_of_goods_and_services	0.07	237.30	45.48	42.39	22.84	1324	5.50
inequality_index	16.24	64.49	42.77	43.09	9.28	429	69.38
life_expectancy	38.20	84.77	67.11	69.86	8.79	1386	1.07
military_expenditure_share_gdp	0.00	13.33	1.92	1.54	1.48	1113	19.49
net_oda_received_per_capita	-49.36	807.19	63.06	34.54	89.16	1239	11.56
net_oda_received_percent_gni	-0.67	189.13	6.11	2.17	12.02	1237	11.71
obesity_prevalence	0.70	44.45	12.77	12.83	8.36	1244	11.21
open_defecation	0.00	66.69	11.70	4.77	15.13	1244	1.43
per_capita_food_production_variability	0.30	106.02	10.57	7.02	12.30	1314	6.21
per_capita_food_supply_variability	2.02	141.28	37.96	31.07	23.66	1229	12.28
percentage_of_arable_land_equipped_for_irrigation	0.00	101.91	27.89	18.85	28.58	1152	17.70
political_stability	-2.78	1.38	-0.38	-0.29	0.86	1261	9.64
population_growth	-2.87	14.22	1.64	1.55	1.30	1399	0.07
prevalence_of_undernourishment	**2.49**	**59.09**	**15.51**	**12.12**	**11.61**	**1401**	**0.00**

Feature	Min	Max	Mean	Median	Std Dev	Count	% NaN
proportion_of_seats_held_by_women_in_gov	0.00	64.77	15.62	13.09	10.32	1228	10.21
rail_lines_density	0.00	4.87	1.18	0.61	1.17	449	67.38
rural_population	0.00	894.73 M	26.58 M	3.37 M	105.24 M	1370	0.00
school_enrollment_rate_female	35.62	101.62	88.67	93.57	12.86	795	43.25
school_enrollment_rate_total	35.34	101.78	90.25	94.64	11.17	897	35.97
tax_revenue_share_gdp	0.06	58.76	16.41	15.19	7.86	856	38.90
total_labor_force	0.03 M	498.58 M	18.71 M	3.41 M	61.12 M	1337	4.57
total_land_area	20.18	24.03 M	0.82 M	0.13 M	2.79 M	1401	0.00
total_population	61.72 k	1.31 B	44.99 M	7.38 M	154.67 M	1401	0.00
trade_in_services	2.31	269.98	23.04	17.31	21.66	1236	11.78
unemployment_rate	0.49	37.98	8.58	6.63	6.65	1337	4.57
urban_population	24.14 k	430.22 M	18.40 M	3.51 M	51.51 M	1401	0.00

Since **prevalence_of_undernourishment** (PoU) is of interest in this analysis, it was noted that the mean and median are significantly different, and the comparatively large standard deviation indicates a considerable variance in the PoU for countries. A histogram of the PoU shows its values are right-skewed – in other words, most countries in the data provided have a low PoU, as shown here:



In addition to the numeric values, the FAO statistics include categorical features, including:

- *country_code* – Anonymised country identifier.
- *year* – For the purposes of this analysis year is defined as categorical and ranges from 2000 to 2015.

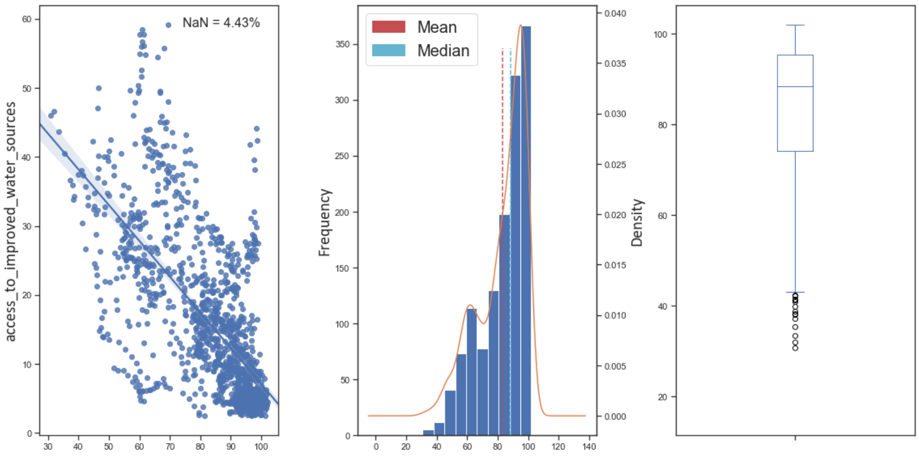
One key observation is that the number of observations per country is not constant and many counties are missing data for some years. This was a key factor in deciding how to deal with missing data.

Correlation and Apparent Relationships

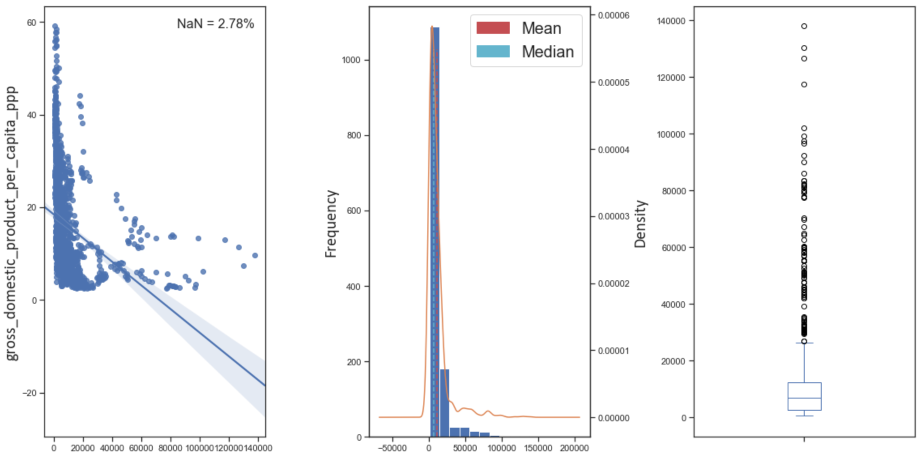
While exploring the individual features, an attempt was made to identify relationships between the features in the data – specifically between PoU and other features.

The following charts show the most promising features based on availability and correlation with PoU:

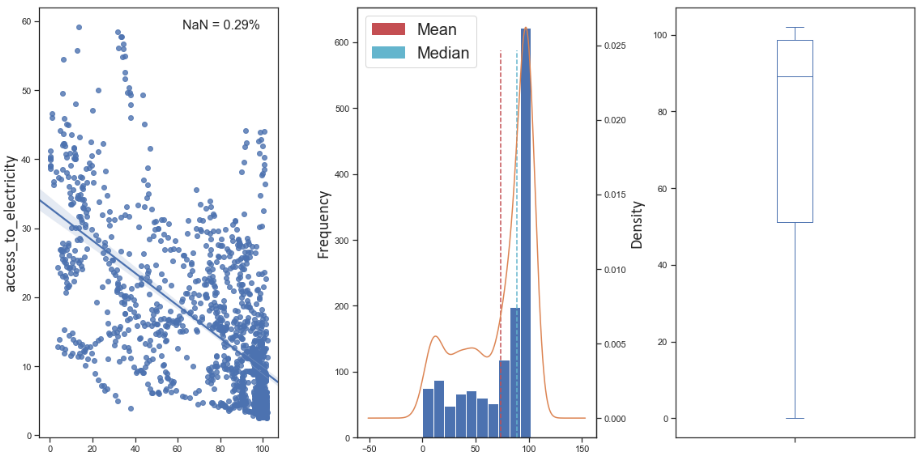
access_to_improved_water_sources



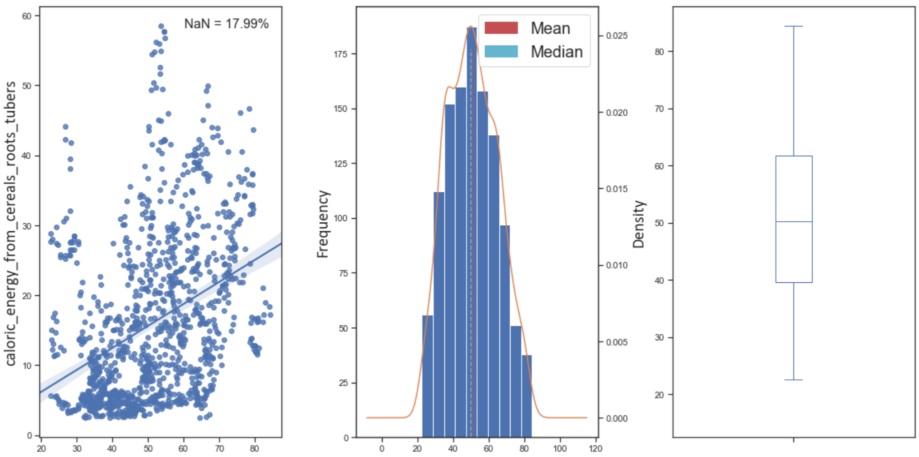
gross_domestic_product_per_capita_ppp



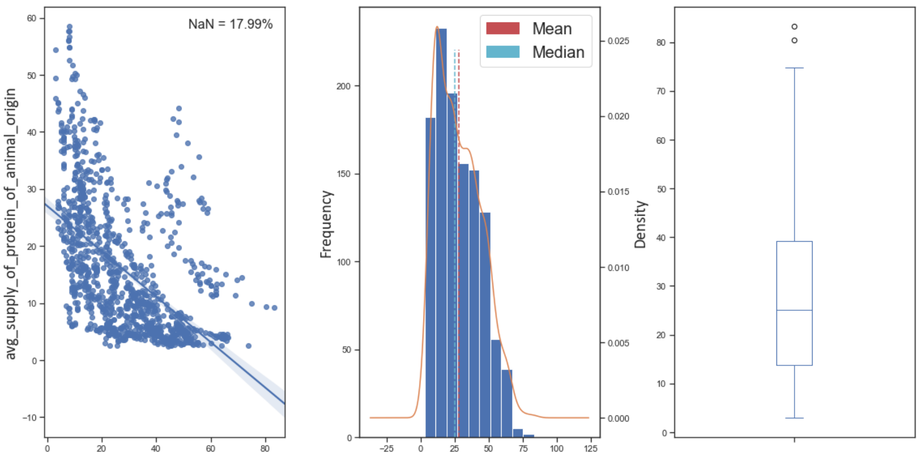
access_to_electricity



caloric_energy_from_cereals_roots_tubers



avg_supply_of_protein_of_animal_origin

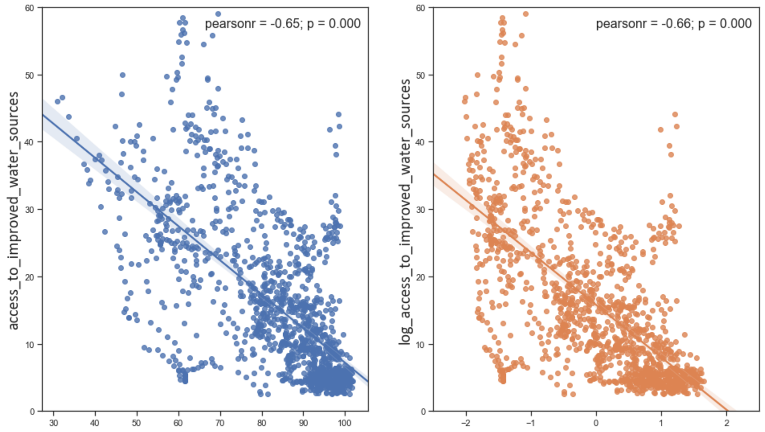


Apart from caloric_energy_from_cereals_roots_tubers, all the features have negative correlations. All the features have missing data and after reviewing this it was decided to inter- and extrapolate the missing data per country linearly. Given the value are over time and show trends it is better to do this than substitute with mean or median values.

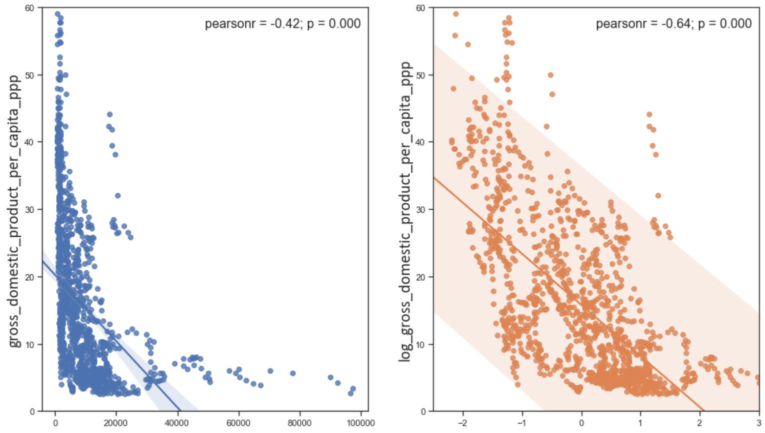
Based on the relationship between prevalence_of_undernourishment and: access_to_improved_water_sources, gross_domestic_product_per_capita_ppp, access_to_electricity, and avg_supply_of_protein_of_animal_origin, as well as their distribution it was felt that transforming them could improve correlation.

Making these changes produced the following correlations:

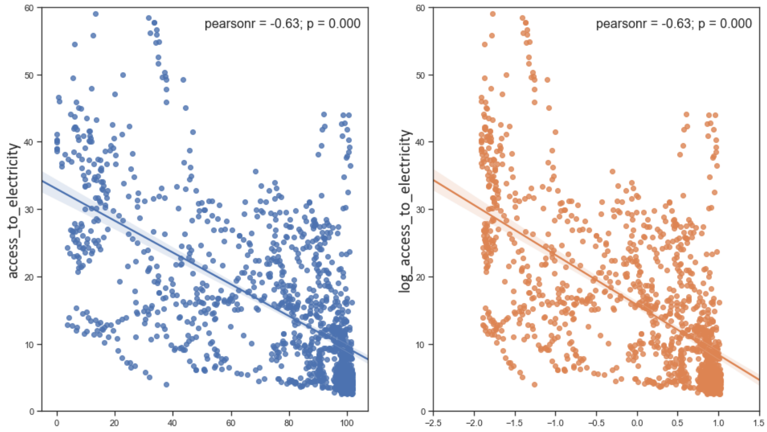
access_to_improved_water_sources

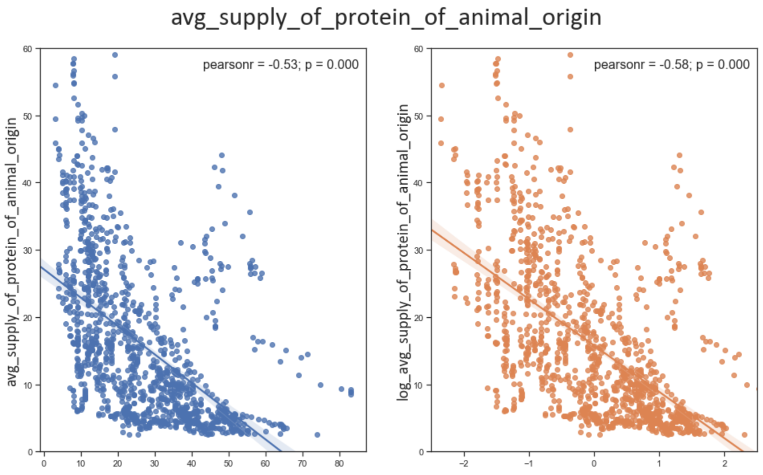


gross_domestic_product_per_capita_ppp



access_to_electricity





From the graphs we can see the Pearson Correlation factor improved and these moderately strong log features were used in building the final regression model.

Regression – Prediction of PoU

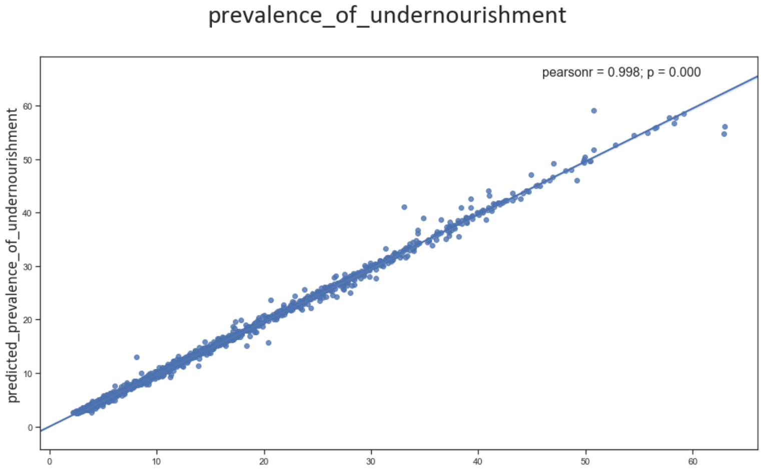
After this initial analysis and data cleansing a Probability Principal Component Analysis was performed. PCA is a technique to reduce the dimensionality of the features. This is a long way of saying it identifies the features with highest impact on the prediction and thus the best features to utilise. The PCA confirmed the scatterplot analysis and gave the following probabilities:

- country_code – 1
    - This is not particularly useful as the challenge is to predict PoU for a new set of countries, but it is an indication that the model should cater for this requirement.
- access_to_improved_water_sources – 0.65
- gross_domestic_product_per_capita_ppp – 0.64
- access_to_electricity – 0.63
- caloric_energy_from_cereals_roots_tubers – 0.62
- obesity_prevalence – 0.61
- avg_supply_of_protein_of_animal_origin – 0.60

These features were used to train a Neural Network based prediction model. Additionally, a 4-fold stratified split was made. The stratified split was based on the country_code in order to ensure the model worked by country rather than generically. This consisted of training 4 models on 75% of the data and testing on the remaining 25%, in a rolling manner thereby utilising all the available data.

All 4 folds were cross-validated to ensure the optimal model hyper-parameters were chosen. In short, cross validation uses testing to verify model is a good as can be given the selected features. The result is a single optimised Neural Network prediction model.

A scatter plot showing the predicted prevalence_of_undernourishment vs. the given prevalence_of_undernourishment is shown below:



This plot shows a clear linear relationship between the predicted and actual values in the test dataset. The Root Mean Square Error (RMSE) for the test results in 0.7555. The actual and predicted summary statistics are:

PoU	Min	Max	Mean	Median	Std Dev
Given	2.49	59.09	15.78	12.19	11.81
Predicted	2.18	62.98	15.94	12.36	11.88

These figures suggest the model performs reasonably well.

Conclusion

This analysis shows that the Prevalence of Undernourishment can be predicted from the PAO statistics provided. Specifically, access to improved water sources; GDP per capita adjusted for pricing parity; access to electricity; calorific energy from cereals, roots, and tubers; prevalence of obesity; and average supply of proteins from animal sources.

Additionally, some opportunities for feature engineering exists as there are strong correlations between, for example, access to electricity and improved water, or calorific energy and average supply of proteins.