Predicting Chronic Hunger

Alex Anguiano, October 2018

Executive Summary

This document presents an analysis of the Capstone challenge including an exploration of the data and a description of the predictive model used. The goal of the challenge is to predict the prevalence of undernourishment for each row of the test data, using a training set of historical data. The data consists of 46 elements, including the target (prevalence of undernourishment), and 45 associated variables.

The variables are as follows:

ID

- country_code - Unique identifier for each country.
- year

AGRICULTURE

- agricultural_land_area - Land area in square kilometers that is suitable or used for growing crops or as pastures.
- percentage_of_arable_land_equipped_for_irrigation - Percent of total arable land that is equipped for irrigation.
- cereal_yield - Average yield in kg/hectare of wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains.
- droughts_floods_extreme_temps - Annual average percent of the population that is affected by droughts, floods, or extreme temperature events (average 1990-2009).
- forest_area - Land area in square kilometers that is under natural or planted stands of trees in their original location. Excludes tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens.
- total_land_area - Total land area of a country in square kilometers.

DEMOGRAPHICS

- fertility_rate - Number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year. Measured in births per woman.
- life_expectancy - Number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
- rural_population - Number of people living in rural areas.

- total_population - Number of people living in a country.
- urban_population - Number of people living in urban areas.
- population_growth - Annual population growth rate (%).

ECONOMICS

- avg_value_of_food_production - Estimated food net production value of a country expressed in per capita terms. Measured in constant 2004-06 international dollars per person.
- cereal_import_dependency_ratio - The cereal imports dependency ratio tells how much of the available domestic food supply of cereals has been imported and how much comes from the country's own production. It is computed as (cereal imports - cereal exports)/(cereal production + cereal imports - cereal exports) * 100. Negative values indicate that the country is a net exporter of cereals.
- food_imports_as_share_of_merch_exports - Value of food imports expressed as a percent of total merchandise exports.
- gross_domestic_product_per_capita_ppp - Gross domestic product (GDP) is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is divided by the total population to be expressed in per capita terms. 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. Measured in constant 2011 international dollars per person.
- imports_of_goods_and_services - Value of all goods and other market services received from the rest of the world expressed as a percent of GDP.
- inequality_index - The Gini index measures how equal the income distribution is in a country. A Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.
- net_oda_received_percent_gni - Net official development assistance received expressed as a share of gross national income (GNI). The ratio of aid to GNI provides a measure of recipient country's dependency on aid, where higher values indicate a greater dependency.
- net_oda_received_per_capita - Net official development assistance received divided by the total population. Measured in current US$ per capita.
- tax_revenue_share_gdp - Tax revenue (value of all taxes collected) expressed as a percent of GDP.
- trade_in_services - Trade in services (sum of service exports and imports) expressed as a percent of GDP.
- per_capita_food_production_variability - Variability of food production value (avg_value_of_food_production). Measured in constant 2004-06 international dollars per person.

- per_capita_food_supply_variability - Variability of the food supply in per capita terms. Measured in kcal/capita/day per person.

EDUCATION

- adult_literacy_rate - Percent of people ages 15 and above who can both read, write and understand a short simple statement about their everyday life.
- school_enrollment_rate_female - Percent of female primary education-aged children enrolled in school.
- school_enrollment_rate_total - Percent of all primary education-aged children enrolled in school.

FOOD SECURITY

- avg_supply_of_protein_of_animal_origin - Average protein supply expressed in grams per capita per day. It includes protein from meat, milk, eggs, fish, seafood, and other animal products.
- caloric_energy_from_cereals_roots_tubers - Percent of total dietary energy supply coming from cereals, roots and tubers.

HEALTH

- access_to_improved_sanitation - Percent of the population with at least adequate access to excreta disposal facilities that can effectively prevent human, animal, and insect contact with excreta. Improved facilities range from simple but protected pit latrines to flush toilets with a sewerage connection.
- access_to_improved_water_sources - Percent of the population with reasonable access to an adequate amount of water from an improved source, such as a household connection, public standpipe, borehole, protected well or spring, and rainwater collection.
- anemia_prevalence - Percent of women of reproductive age (15-49 years) who meet the clinical definition of anemia.
- obesity_prevalence - Percent of adults ages 18 and over whose Body Mass Index is more than 30 kg/m2.
- open_defecation - Percent of the population defecating in the open, such as in fields, forest, bushes, open bodies of water, or on beaches.
- hiv_incidence - Number of new HIV infections among uninfected populations ages 15-49 expressed per 100 people in the uninfected population in the previous year.

INFRASTRUCTURE

- rail_lines_density - Ratio between the length of railway route available for train service and the area of the country (per 100 sq km of land area).

- access_to_electricity - Percent of population with access to electricity.
- co2_emissions - Carbon dioxide emissions in kt (thousands of metric tons).

LABOR

- unemployment_rate - Percent of the labor force that is without work but available for and seeking employment.
- total_labor_force - Total number of people who are currently employed, people who are unemployed but seeking work, and first-time job-seekers.

POLITICS

- military_expenditure_share_gdp - Spending on the armed forces and defense ministries expressed as a percent of the country's GDP.
- proportion_of_seats_held_by_women_in_gov - Percent of seats in national parliaments held by women.
- political_stability - Index of the perceived likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically-motivated violence and terrorism.

Data Exploration

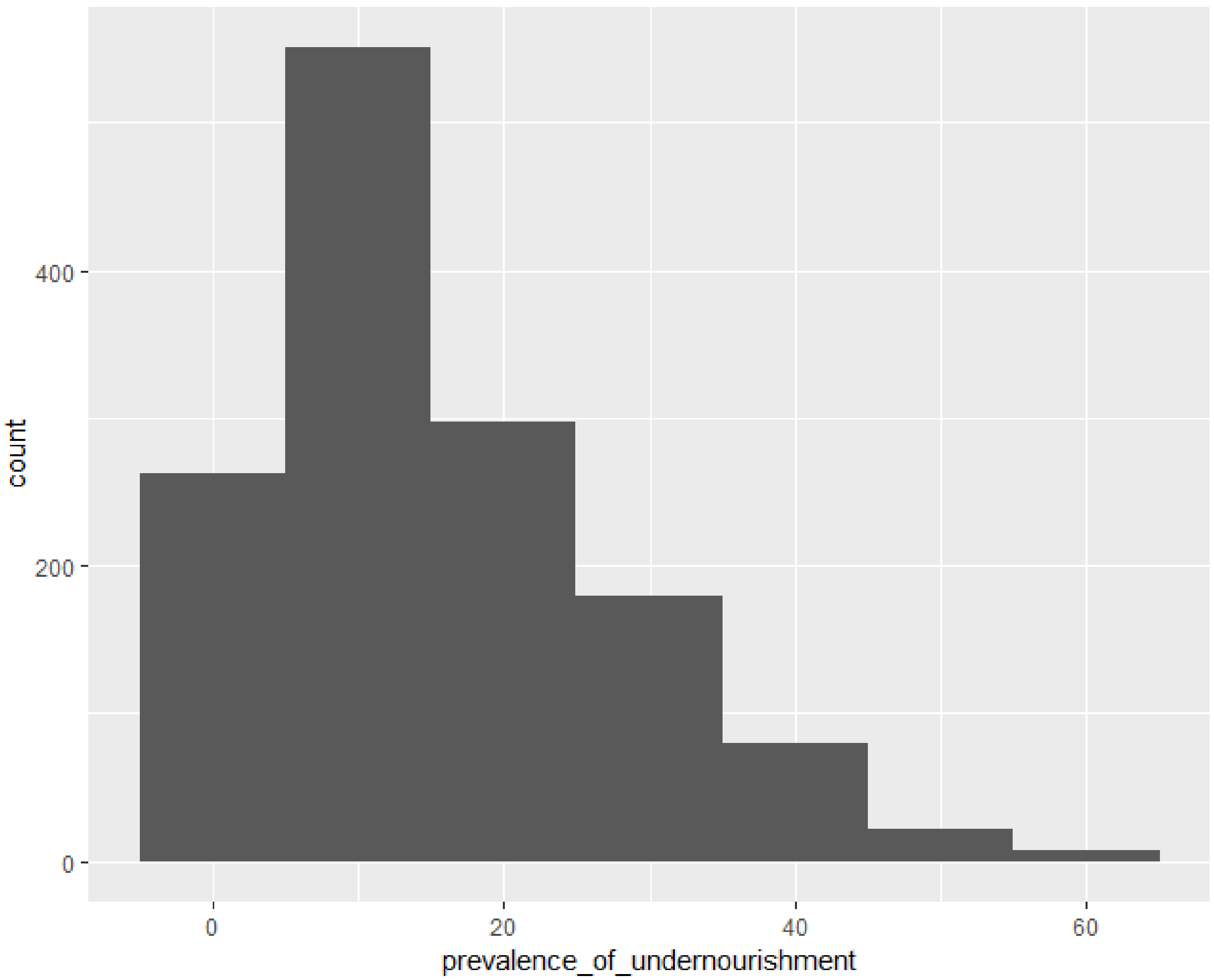
Initial preparation was done to the original data by adding the Training_Labels prevalence_of_undernourishment column to the Training_Values table.

Data exploration began with a summary of the target and variables, as seen in the examples below.

prevalence_of_undernourishment	cereal_yield	forest_area
Min. : 2.493	Min. : 179.3	Min. : 10
1st Qu.: 5.711	1st Qu.: 1424.5	1st Qu.: 4159
Median :12.119	Median : 2221.9	Median : 22242
Mean :15.511	Mean : 2753.2	Mean : 232945
3rd Qu.:22.447	3rd Qu.: 3296.5	3rd Qu.: 125596
Max. :59.090	Max. :27978.3	Max. :8243222
NA	NA's :64	NA's :16

Since Prevalence of Undernourishment is of interest in this analysis, a histogram was created for further exploration. The graphic below shows that the data is right skewed, or in other words, most instances are on the lower end of the range.

Histogram of Prevalence of Undernourishment



Data Preparation

Also noted in the exploration of the data, was that there were a number of variables with issues that might prove to be challenging for a predictive model. First, as can be seen below, there are variables which contain entries of “0”, and have NA’s or missing data. The concern here is whether a zero is a true entry or a representation of a missing value. For instances where 0 was the Min. value and NA’s were present, the zero was turned into a missing value. If the Min. value was a negative number, the zero was left unchanged as the possibility of this being a true entry was more likely. Second, there are 1401 rows of training data. As can also be seen below, some variables are missing more than half of the 1401 observations, and in some instances quite a bit more. The Droughts/Floods/Extreme Temps variable for example is missing 1326 entries, which is 94% of the total observations for this column. In these cases, the entire column of variables was removed from consideration. The remaining missing values had to be resolved. To continue to remove all rows of missing values is an option, but it would remove a considerable amount of training data, potentially weakening the predictive strength of any model. The alternative was to run a Pre-Processing function which uses a bagged tree model for imputation, or filling in, of missing data. Finally, once the imputation process was completed, the countries code column was removed as well.

droughts_floods_extreme_temps	**adult_literacy_rate**	**rail_lines_density**
Min. :0.0000	Min. : 24.14	Min. :0.0000
1st Qu.:0.0974	1st Qu.: 66.77	1st Qu.:0.2977
Median :0.6614	Median : 87.03	Median :0.6082
Mean :1.2368	Mean : 79.63	Mean :1.1831
3rd Qu.:1.3183	3rd Qu.: 93.80	3rd Qu.:1.8692
Max. :9.1773	Max. :100.46	Max. :4.8672
NA's :1326	NA's :1116	NA's :944

Prediction Model

Since the target variable is a numeric quantity, the training data will be processed through 7 regression models, with the best scoring being used for prediction. The models used were:

Linear Models

- Linear Regression
- Partial Least Squares (PLS)
- Penalized Lasso

Nonlinear Models

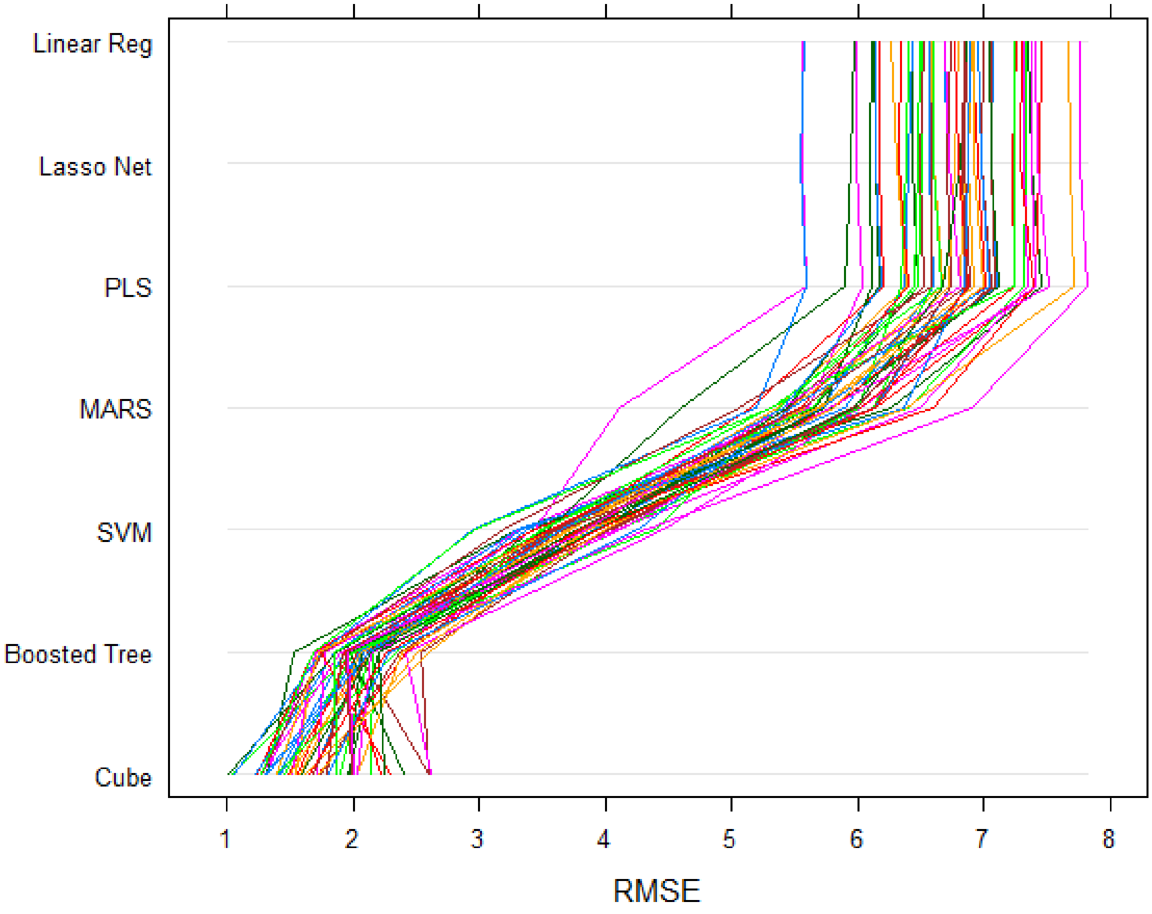
- Multivariate Adaptive Regression Splines (MARS)
- Support Vector Machine (SVM)

Tree and Rule Based Models

- Boosted Tree
- Cubist (Cube)

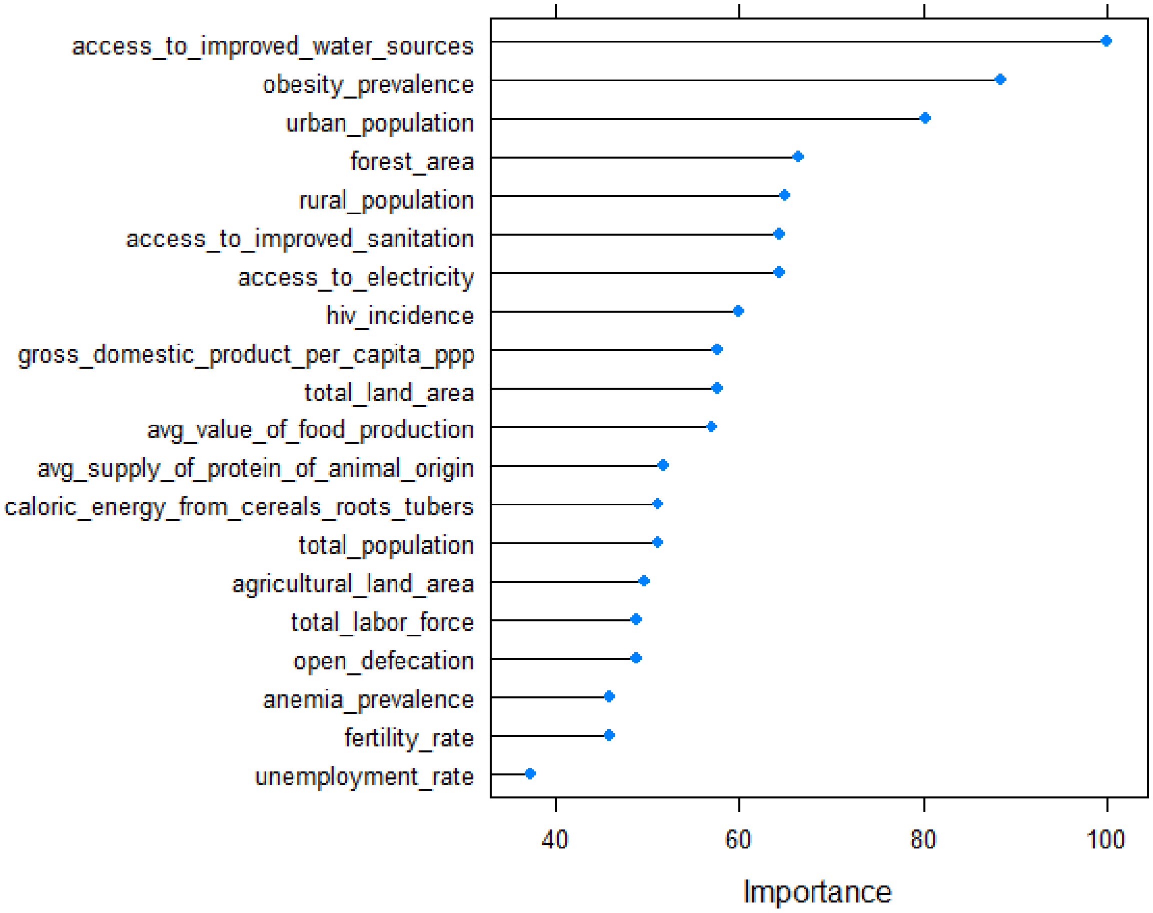
All models were tuned using 6 clusters of 10-fold cross validation repeated 5 times. Once the models were finished, the RMSE of the hold out sets from the tuning process were visualized as can be seen below. According to the visualization, the Cubist model performed the best, with RMSE scores between 1 – 2.75. The Test data was prepared in the same fashion as the training data above, and a prediction was made against it using the trained Cubist model. The prediction against the test data was submitted and returned a RMSE score of 9.1638, achieving a rank above the Benchmark grade of 90+.

RMSE of Cross Validated Holdout Sets

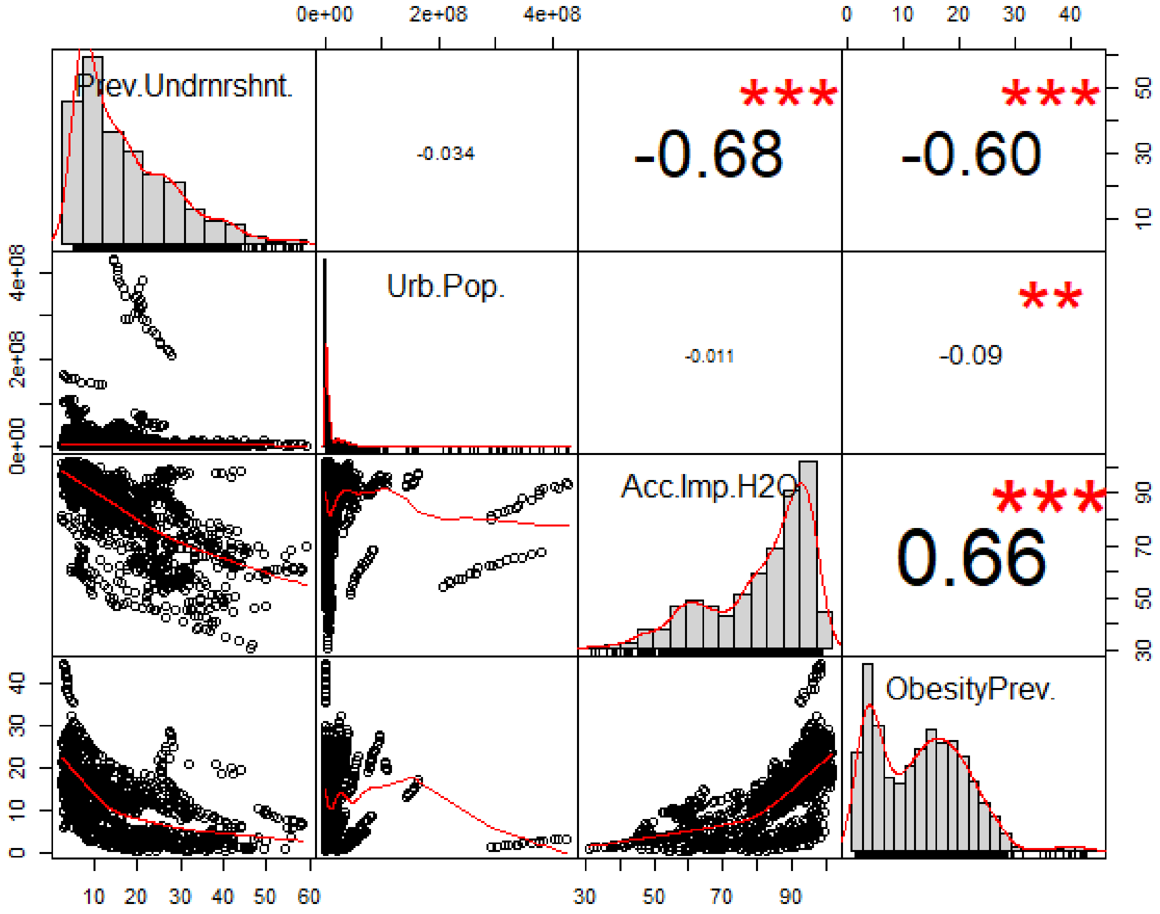


Below is a visualization which shows Variable Importance for the Top 20 predictors used in the Cubist model for predicting prevalence of undernourishment.

Variable Importance Scores



A review of the top 3 Variable Importance Scores from the unprepared training data shows strong negative correlations between prevalence of undernourishment and access to improved water sources, as well as prevalence of undernourishment and obesity prevalence. There is also a weak negative correlation between prevalence of undernourishment and urban population. This can be seen in the correlation matrix below.



Conclusion

This analysis has shown that the prevalence of undernourishment at the country level can be predicted based on certain socioeconomic indicators. In particular, access to improved water sources, the prevalence of obesity, and urban population. A more in-depth examination of the variables and feature selection could provide further insight as to how other indicators contribute to, and help predict, chronic hunger in a given country.