4

WEEKS

LEFT

DAT102x: Predicting Chronic Hunger

HOSTED BY MICROSOFT

HOME (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/PAGE/28/)

PROBLEM DESCRIPTION (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/PAGE/29/)

EXPORT SCORE FOR EDX (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/PAGE/30/)

LEADERBOARD (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/LEADERBOARD/)

DATA DOWNLOAD (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/DATA/)

SUBMISSIONS (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/SUBMISSIONS/)

OFFICIAL RULES (/COMPETITIONS/9/PREDICTING-CHRONIC-HUNGER/RULES/)

Problem Description

- References

About the Data

Your goal is to predict the annual prevalence of undernourishment at the country level from other socioeconomic indicators. The prevalence of undernourishment expresses "the probability that a randomly selected individual from the population consumes an amount of calories that is insufficient to cover her/his energy requirement for an active and healthy life" (FAOSTAT). It can be understood as the percent of the total population that is facing chronic hunger. Data is compiled from the Food and Agricultural Organization (http://www.fao.org/home/en/) of the United Nations as well as the World Bank (https://www.worldbank.org/).

Target Variable

We're trying to predict the variable `prevalence_of_undernourishment` for each row of the test data set. `prevalence_of_undernourishment` is a positive floating point number (e.g. 1.0) between 0.0 and 100.0 inclusive.

Your job is to:

1. Train a model using the inputs in `train_values.csv` and the labels `train_labels.csv`
2. Predict floats for each row in `test_values.csv` for which you don't know the true prevalence of undernourishment.
3. Output your predictions in a format that matches `submission_format.csv` **exactly**.
4. Upload your predictions to this competition in order to get a score.
5. Export your token (https://datasciencecapstone.org/competitions/9/predicting-chronic-hunger/page/30/) and paste it into the assignment grader on edX to get your course grade.

Submission Format

The format for the submission file is two columns with the `row_id` and the `prevalence_of_undernourishment`. The data type of `prevalence_of_undernourishment` is a float, **so make sure there is a decimal point in your submission**. For example `100.0` is a valid float but `100` is *not*.

The first few lines of the `.csv` file that you submit would look like:

```
row_id,prevalence_of_undernourishment
0,1.0
1,1.0
2,1.0
3,1.0
4,1.0
⋮
```

Performance Metric

We're predicting a numeric quantity, so this is a regression problem. To measure regression, we'll use a metric called root-mean-squared error. It is an error metric, so lower value is better (as opposed to an accuracy metric, where a higher value is better).

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{y}_n - y_n)^2)}{N}}$$

Where $\hat{y}$ is the predicted prevalence of undernourishment and "y" is the actual prevalence of undernourishment. The best possible score is 0, but the worst possible score can be infinite.

Features

There are 45 variables in this dataset. Each row in the dataset represents a country in a given year. Each country has a unique identifier, `country_code`, as country names are not given.

The country codes in the test set **are distinct** from those in the train set. In other words, no country that appears in the train set appears in the test set. Thus, country-specific features (i.e. country dummy variables) will not be an option. However, the countries in the test set still share similar patterns as those in the train set and so other feature engineering will work the same as usual.

When you are doing your local train/test split, it is imperative that you split the data by country so that all years of data for a country appear either in the train set or the test set, but are not split across both. Otherwise, your model will simply interpolate between the known values in your train set rather than learning the underlying patterns in the data. The model will then score poorly when submitted as it will be evaluated on data from entirely different countries.

If your leaderboard score is significantly lower than your local score, it is likely because you have not split your data by country when training your model.

For each country, there are between 5 and 16 years of data in the train set. In the test set, there are at least 10 years of data for each country.

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.

Example Row

Here's an example of one of the rows in the dataset so that you can see the kinds of values you might expect in the dataset. Most are numeric, one is categorical, and there can be missing values.

	0
country_code	889f053
year	2002
agricultural_land_area	235078
percentage_of_arable_land_equipped_for_irrigation	38.5585
cereal_yield	935.754
droughts_floods_extreme_temps	NaN
forest_area	5397.74
total_land_area	537408
fertility_rate	5.92898
life_expectancy	60.4522
rural_population	1.35409e+07
total_population	1.88995e+07
urban_population	5.08645e+06
population_growth	2.86493
avg_value_of_food_production	60.8913
cereal_import_dependency_ratio	78.4528
food_imports_as_share_of_merch_exports	22.0805
gross_domestic_product_per_capita_ppp	3969.52
imports_of_goods_and_services	38.3518
inequality_index	NaN
net_oda_received_percent_gni	2.23356

	0
net_oda_received_per_capita	11.659
tax_revenue_share_gdp	NaN
trade_in_services	NaN
per_capita_food_production_variability	1.96809
per_capita_food_supply_variability	15.6935
adult_literacy_rate	NaN
school_enrollment_rate_female	NaN
school_enrollment_rate_total	NaN
avg_supply_of_protein_of_animal_origin	11.0241
caloric_energy_from_cereals_roots_tubers	64.1296
access_to_improved_sanitation	43.2865
access_to_improved_water_sources	57.9328
anemia_prevalence	59.0614
obesity_prevalence	8.29394
open_defecation	28.9335
hiv_incidence	0.00997508
rail_lines_density	NaN
access_to_electricity	52.4332
co2_emissions	15485.1
unemployment_rate	14.788
total_labor_force	4.35052e+06
military_expenditure_share_gdp	7.02107
proportion_of_seats_held_by_women_in_gov	0.698153
political_stability	-1.3938

References

- FAOSTAT, Food and Agriculture Organization of the United Nations. Definitions and Standards. http://www.fao.org/faostat/en/ ().

Terms of Use (/termsofuse/) · Copyright Policy (/copyrightpolicy/) · Privacy Policy (/privacypolicy/)

Work with our community (https://www.drivendata.org/) · Work with our team (http://www.drivendata.co/)

Hosted by (https://www.drivendata.org/)