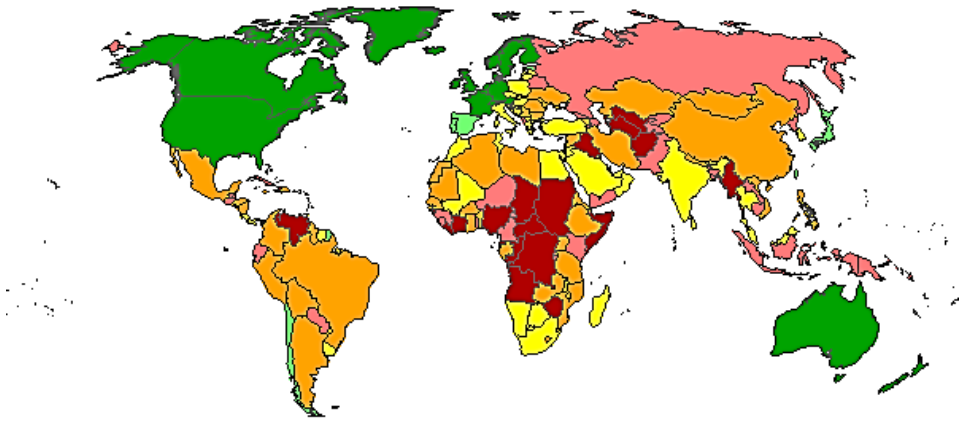
Predicting Chronic Hunger

Oct. 2018,
Shao-Chi Lee, Ting-Yao Chang, Chieh-Sheng Liao, Kai-Yuan Cheng

Executive Summary

This document presents an analysis of 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 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. This dataset is provided by Food and Agricultural Organization of the United Nations. After exploring the data by calculating summary, descriptive statistics, and creating visualizations of the data, several potential relationships between a country's socioeconomic indicators and its prevalence of undernourishment were identified. After exploring the data, a regression model to predict a country's undernourishment from its features was created.



Here are the steps of our research following:

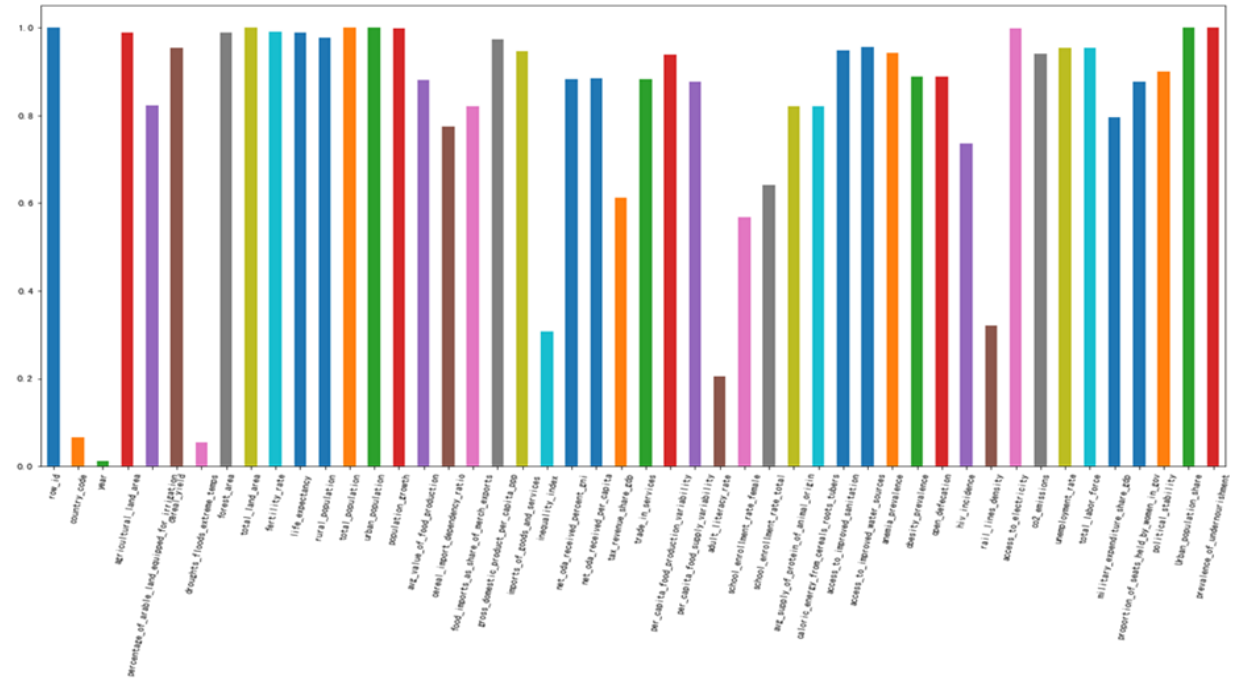
1) Exploratory Data Analysis
2) Data cleaning
3) Correlation and Apparent Relationships
4) Feature selecting
5) Model building and Testing
6) Conclusion

At first, doing data explore by calculating summary and descriptive statistics, and by creating visualizations of the data to find out the data's properties and figure out the way to solve it. Second, start to clean data from fill null numbers. If the features had the too much null number, drop it out. After that, according the correlation between features to do feature selecting. Finally, building a regression model to predict the annual prevalence of undernourishment and test Performance Metric with RMSE.

Exploratory Data Analysis

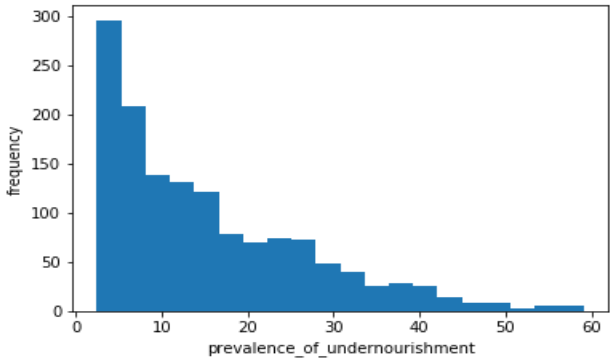
According to the problem description, our target is to predict the prevalence of undernourishment from features provided. At the beginning, it is important to overview the dataset and do EDA (Exploratory Data Analysis). This step can help us to understand the dataset properly and efficiently.

In order to classify numerical features and categorical features, the number of unique values of each features is provided as below. It is recovered that most of features have almost 100% unique values. Therefore, those features should be treated as numerical features. In addition to the numeric values, this dataset include categorical features. Columns “country_code”, “year” can be treated as categorical features at the later analysis.



Target Value Distribution

Since prevalence of undernourishment is of interest in this analysis, it was noted that the mean and median of this value are significantly different and that the comparatively large standard deviation indicates that there is considerable variance in the prevalence of undernourishment among the countries. A histogram of the prevalence of undernourishment column shows that the price values are right-skewed – in other words, most countries have low rate of prevalence of undernourishment, as shown here:



count	1401
mean	15.5107
std	11.6104
min	2.4934
25%	5.7109 8
50%	12.1187
75%	22.4475
max	59.089

Initial Data's Information:

The table below list all features' duplicate numbers, null numbers and the rate of null numbers. It makes a quick check of our data.

	Duplicate	Is null	Null number	Null rate
row_id	1401	False	0	0.0000
country_code	92	False	0	0.0000
year	16	False	0	0.0000
agricultural_land_area	1386	True	16	0.0114
percentage_of_arable_land_equipped_for_irrigation	1153	True	248	0.1770
cereal_yield	1338	True	64	0.0457
droughts_floods_extreme_temps	76	True	1326	0.9465
forest_area	1386	True	16	0.0114
total_land_area	1401	False	0	0.0000
fertility_rate	1388	True	14	0.0100
life_expectancy	1387	True	15	0.0107
rural_population	1370	False	0	0.0000
total_population	1401	False	0	0.0000
urban_population	1401	False	0	0.0000
population_growth	1400	True	1	0.0007
avg_value_of_food_production	1235	True	167	0.1192
cereal_import_dependency_ratio	1085	True	317	0.2263
food_imports_as_share_of_merch_exports	1149	True	253	0.1806
gross_domestic_product_per_capita_ppp	1363	True	39	0.0278
imports_of_goods_and_services	1325	True	77	0.0550
inequality_index	430	True	972	0.6938
net_oda_received_percent_gni	1238	True	164	0.1171
net_oda_received_per_capita	1240	True	162	0.1156
tax_revenue_share_gdp	857	True	545	0.3890
trade_in_services	1237	True	165	0.1178
per_capita_food_production_variability	1315	True	87	0.0621
per_capita_food_supply_variability	1230	True	172	0.1228
adult_literacy_rate	286	True	1116	0.7966
school_enrollment_rate_female	796	True	606	0.4325
school_enrollment_rate_total	898	True	504	0.3597
avg_supply_of_protein_of_animal_origin	1150	True	252	0.1799
caloric_energy_from_cereals_roots_tubers	1150	True	252	0.1799
access_to_improved_sanitation	1328	True	74	0.0528
access_to_improved_water_sources	1340	True	62	0.0443
anemia_prevalence	1322	True	80	0.0571
obesity_prevalence	1245	True	157	0.1121
open_defecation	1245	True	20	0.0143
hiv_incidence	1031	True	371	0.2648
rail_lines_density	450	True	944	0.6738
access_to_electricity	1398	True	4	0.0029
co2_emissions	1318	True	84	0.0600
unemployment_rate	1338	True	64	0.0457
total_labor_force	1338	True	64	0.0457
military_expenditure_share_gdp	1114	True	273	0.1949
proportion_of_seats_held_by_women_in_gov	1229	True	143	0.1021
political_stability	1262	True	135	0.0964

- Green background means that column of data is category type.
- The yellow background columns means that feature has too many null numbers.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns. The statistic results for each features are shown as below.

Feature	count	mean	std	min
row_id	1401	700	404.5782	0
agricultural_land_area	1385	353958.8	1172377	2.944179
percentage_of_arable_land_equipped_for_irrigation	1153	27.89145	28.57762	0
cereal_yield	1337	2753.178	2777.815	179.2589
droughts_floods_extreme_temps	75	1.2368	1.877823	0
forest_area	1385	232945.5	926633.4	9.806688
total_land_area	1401	818114.6	2792117	20.18306
fertility_rate	1387	3.251874	1.471044	0.836053
life_expectancy	1386	67.11405	8.78685	38.20414
rural_population	1401	26580250	1.05E+08	0
total_population	1401	44991055	1.55E+08	61724.55
urban_population	1401	18404858	51507632	24138.44
population_growth	1400	1.636426	1.299897	-2.87225
avg_value_of_food_production	1234	229.4743	149.0591	3.945363
cereal_import_dependency_ratio	1084	34.37284	51.9373	-228.3
food_imports_as_share_of_merch_exports	1148	37.14854	66.5649	0.990945
gross_domestic_product_per_capita_ppp	1362	10843.43	15275.31	573.1677
imports_of_goods_and_services	1324	45.47967	22.84027	0.06506
inequality_index	429	42.76917	9.278521	16.24072
net_oda_received_percent_gni	1237	6.105307	12.02022	-0.66536
net_oda_received_per_capita	1239	63.0577	89.16066	-49.3556
tax_revenue_share_gdp	856	16.40988	7.863698	0.057901
trade_in_services	1236	23.041	21.65651	2.308559
per_capita_food_production_variability	1314	10.57109	12.30408	0.300291
per_capita_food_supply_variability	1229	37.95659	23.65707	2.018557
adult_literacy_rate	285	79.63224	18.22817	24.14042
school_enrollment_rate_female	795	88.6715	12.86126	35.62018
school_enrollment_rate_total	897	90.2537	11.16576	35.33573
avg_supply_of_protein_of_animal_origin	1149	27.96357	15.98439	2.957107
caloric_energy_from_cereals_roots_tubers	1149	50.88803	13.92569	22.58993
access_to_improved_sanitation	1327	65.05176	28.42234	10.33727
access_to_improved_water_sources	1339	83.2994	15.28494	30.7846
anemia_prevalence	1321	32.78167	11.99932	12.57047
obesity_prevalence	1244	12.76597	8.360314	0.699575
open_defecation	1381	11.70486	15.13444	0
hiv_incidence	1030	0.218624	0.52396	0.0098
rail_lines_density	457	1.183129	1.175	0
access_to_electricity	1397	73.79539	31.28031	0.010012
co2_emissions	1317	83046.71	224836	100.8288
unemployment_rate	1337	8.580335	6.645133	0.491115
total_labor_force	1337	18712325	61123469	34906.59
military_expenditure_share_gdp	1128	1.919332	1.480842	0
proportion_of_seats_held_by_women_in_gov	1258	15.61846	10.32428	0
political_stability	1266	-0.37602	0.858888	-2.78126
Urban_population_share	1401	0.49849	0.214448	0.082009

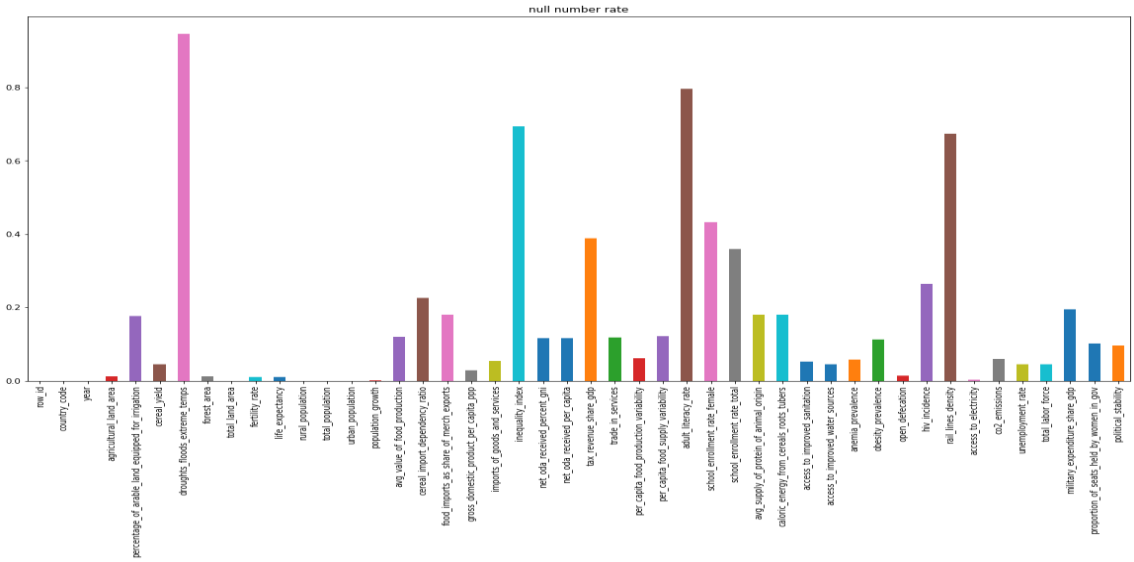
Data Cleaning

In order to train model with more informative features, some methods of data cleaning must be performed to clear the noises of data.

Null Rate over 40%

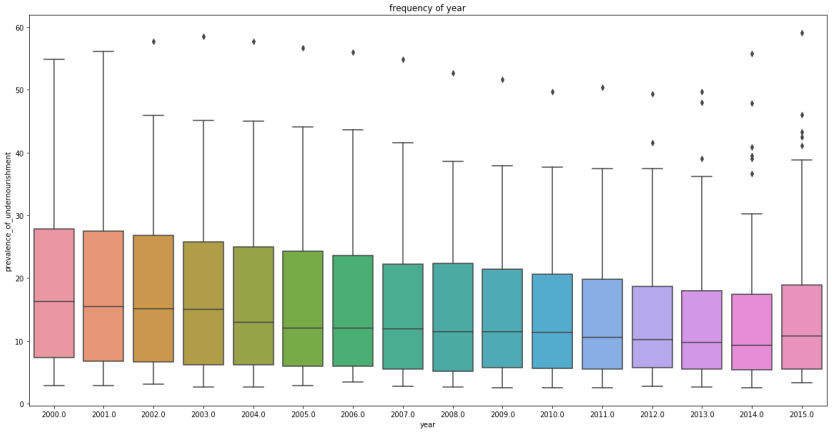
	duplicate	is_null	null_number	null_rate
droughts_floods_extreme_temps	76	True	1326	0.9465
inequality_index	430	True	972	0.6938
adult_literacy_rate	286	True	1116	0.7966
school_enrollment_rate_female	796	True	606	0.4325
rail_lines_density	450	True	944	0.6738

The picture below is the null number's rate. Since the null rate of 'droughts_floods_extreme_temps', 'inequality_index', 'adult_literacy_rate', 'school_enrollment_rate_female' and 'rail_lines_density' are over 40%. Since these columns contains too few information, they are dropped and the rest of columns are used for training the model.



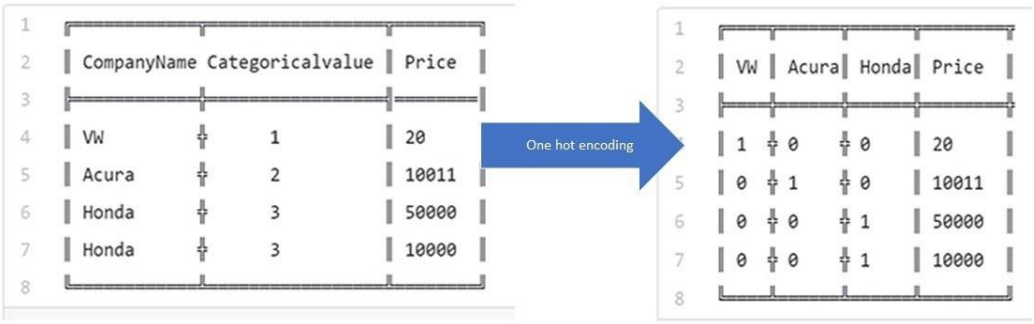
Categorical Data

- **The country codes:** The country codes in the test set **are distinct** from those in the train set. That means it should be drop out the model should not be influenced by it.
- **Year:** As the picture down below, there are exist some relation between 'year' and 'prevalence_of_undernourishment'. To do one-hot encoding to 'year' make it be 16 dummy variables.



One-hot encoding:

One-hot encoding is often used for indicating the state of a state machine. When using binary or Gray code, a decoder is needed to determine the state. A one-hot state machine, however, does not need a decoder as the state machine is in the nth state if and only if the nth bit is high.



Outlier: IQR

In statistics, an outlier is an observation point that is distant from other observations. The above definition suggests that outlier is something which is separate or different from the crowd. The outlier data may have been coded incorrectly or an experiment may not have been run correctly.

The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution.

While calculating the Z-score we re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used. If the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers. According to the characteristic, we can use Z-score to detect the outliers and remove them from our dataset for analysis.

In this project, we apply Tukey's range test to deal with outliers at first. The reason doesn't use 6 sigma is that not every feature are normal distribution or not. If used it to deal with outliers, it would have to normalize data at first. It will probably make the data lose some information. That's the reason why apply Tukey's range test to deal with outliers.

The idea here of Tukey's range test is used 3 times of interquartile range (IQR) to make maximum and minimum estimated value. If the data out of this range, made it be the maximum or minimum estimated value.

$$Maximum\ Estimated\ Value = Q_3 + k \times IQR$$

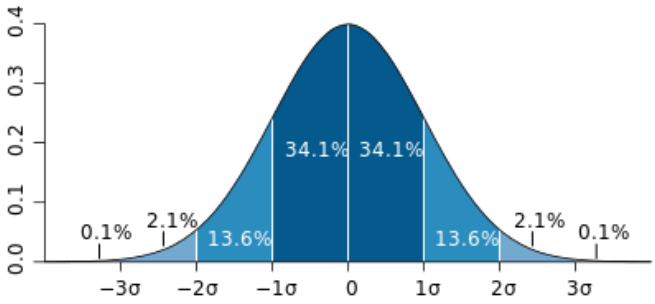
$$Minimum\ Estimated\ Value = Q_1 - k \times IQR$$

interquartile range (IQR) is the difference between upper and lower quartiles i.e.

$$IQR = Q_3 - Q_1$$

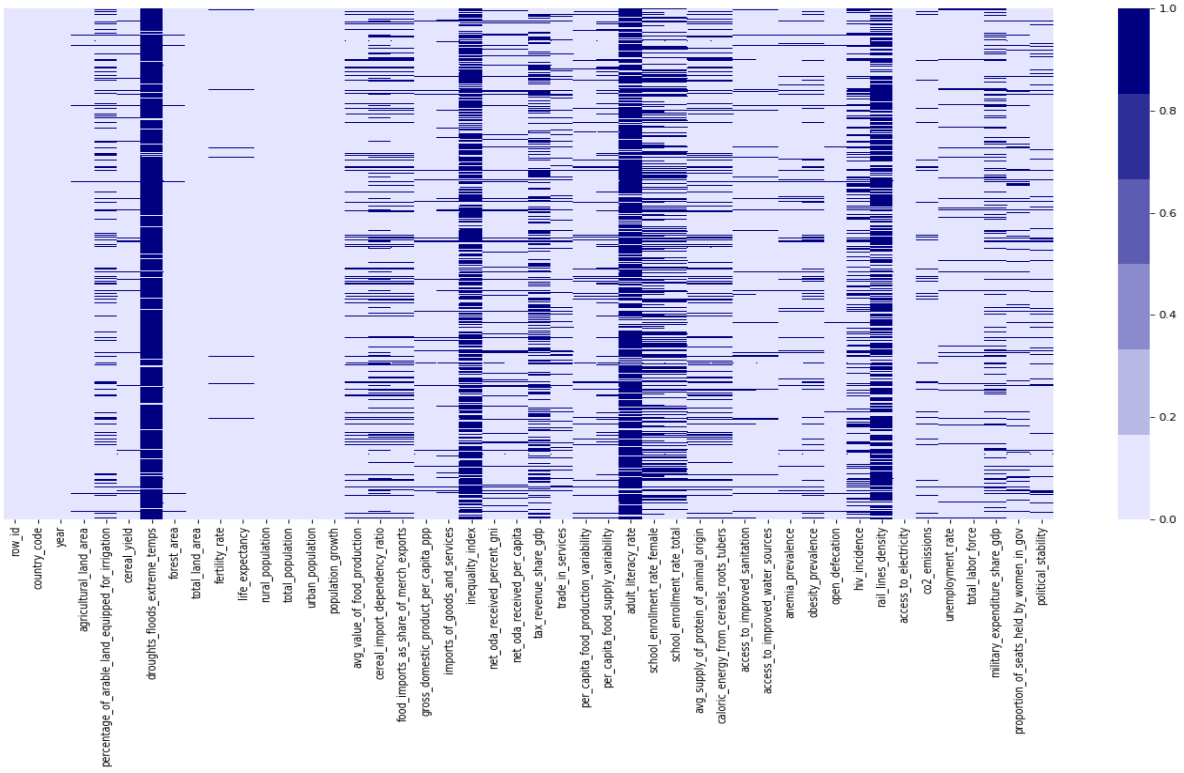
Outlier: Standardize

Besides IQR, the method of standardization was also implemented , which is standardizing all features and change all the outliers to 3 times of standard deviation, then inverse them back. It should have similar effect as IQR method.



Fill Missing Values

The picture below shows the scatter of null numbers. Beside the features we had dropped, still exists many null values have to be filled.



To complete this task, K-Nearest Neighbor Algorithm (KNN) is chosen. Idea of k-NN classification is that the object is classified by a majority vote of its neighbors, with the object being assigned to the class

most common among its k nearest neighbors. It is a simple classification algorithm but it can still give highly competitive results

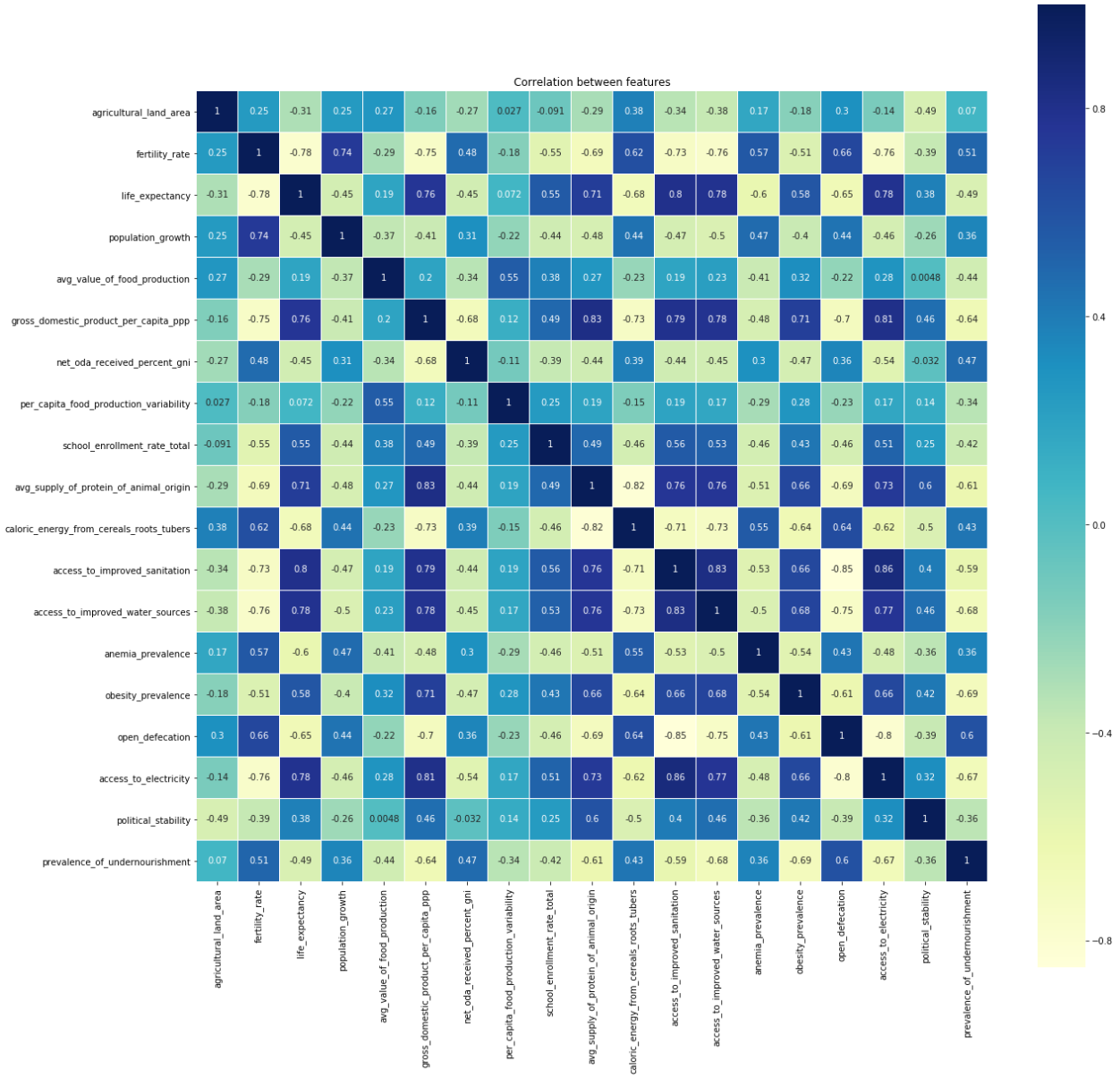
In this stage, the data has no missing value. It's time to build model to predict the annual prevalence of undernourishment.

Correlation and Apparent Relationships

After exploring the individual features and the process of data cleaning, an attempt was made to identify relationships between features in the data – in particular, between prevalence of undernourishment and the other features.

Numeric Relationship

The following heat map was generated initially to compare numeric features with one another. The key features in this matrix are shown here:



Viewing the heat map shows that there are some features highly correlated to each other, this may cause the problem of multicollinearity, which will negatively impact the precision of the model. Therefore, some sort of feature selection must be implemented to erase the effect of multicollinearity.

After studying Feature explanations and combining the correlation table with each other and correlations with target, we decided to adjust these features by method below:

1.Transforming rural population and urban population to population ratio.

2.Using PCA with columns ['avg_supply_of_protein_of_animal_origin', 'caloric_energy_from_cereals_roots_tubers'] as health_food.

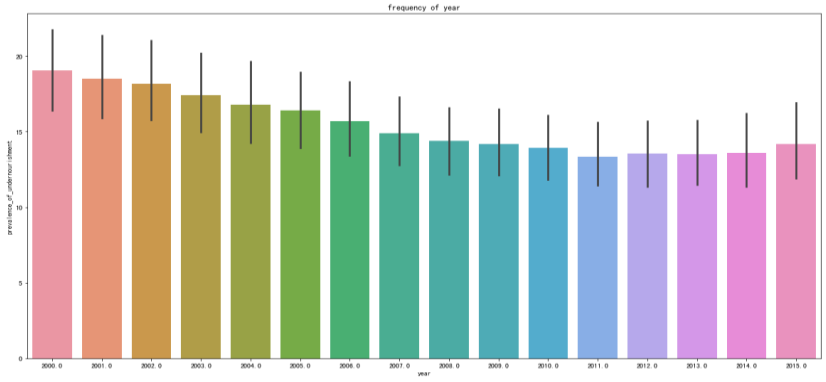
3.Using PCA with columns ['life_expectancy', 'access_to_improved_sanitation', 'access_to_improved_water_sources', 'access_to_electricity', 'fertility_rate'] as heath_facility.

According to this diagram, the correlation coefficient between each feature can be easily observed. Some features like “fertility_rate”, “access_to_electricity”, “access_to_improved_water_sources”, “access_to_improved_sanitation”, “avg_supply_of_protein_of_animal_origin”, “obesity_prevalence” have absolute correlation coefficient higher than 0.5 to the undernourishment prevalence. These features can be considered as significant features to predict the undernourishment.

- ***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.
- ***access_to_electricity:***
  Percent of population with access to electricity.
- ***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.
- ***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.
- ***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.
- ***obesity_prevalence:***
  Percent of adults ages 18 and over whose Body Mass Index is more than 30 kg/m2.

Categorical Relationship

Having explored the relationship between price and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and prevalence of undernourishment. However, in the dataset, there is only one column was treated as categorical data aside from country code column, that is the year column. Therefore, the year column is the only categorical feature be explored. The following boxplots show the year columns. Unfortunately, there appear to be no distinctly difference in prevalence of undernourishment among the distributions of year. Therefore year, the only categorical data, may not be an important feature when predicting the prevalence of undernourishment.



Here comes another problem — How can we judge the features by these levels? If we don't the domain know-how of this field, it will not make sense since we select features in this way. For this reason, we prefer to apply mathematic way to solve it. Here come Stepwise Regression and Lasso.

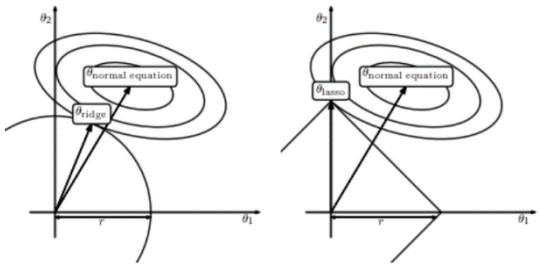
Feature Selection

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modeling are

1. Reduces Overfitting: Less opportunity to make decisions based on noise.
2. Improves Accuracy: Less misleading data means modeling accuracy improves.
3. Reduces Training Time: Less data means that algorithms train faster.

In order to select features, L1-based algorithm (Lasso) is chosen. Linear models penalized with the L1 norm have sparse solutions. Many of their estimated coefficients are zero. Based on this characteristic, it can be used to select the features which coefficients are non-zero. These features are selected to use for modeling.



Lasso

The idea of Lasso is to add a penalty to our loss. It will make our model not learning too perfect then cause overfitting. (Loss + L1)

$$\frac{1}{k}\sum_{j=1}^{k} l(y^{(j)}, \widehat{y^{(j)}}) + \alpha \sum_{j} |w_j|$$

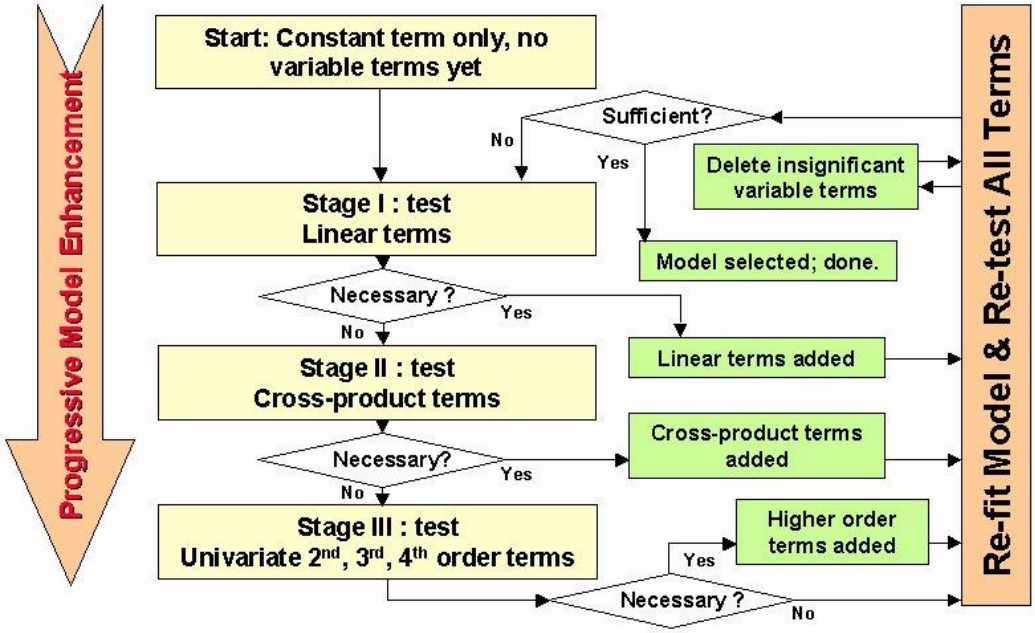
Lasso regression puts constraints on the size of the coefficients associated with each variable. However, this value will depend on the magnitude of each variable. It is, therefore, necessary to center and reduce, or standardize, the variables. Therefore, Normalization is very important for methods with regularization. This is because the scale of the variables affects the how much regularization will be applied to specific variable.

As mentioned earlier, there are problems of multicollinearity existed among features. Therefore, some sort of feature selection method must be implemented to solve this problem. For this particular case, since most of the features are numeric features, it's most suitable to use the LassoCV to select features. After the LassoCV method was implemented, the number of features drop from 47 to 35 features.

Stepwise

stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

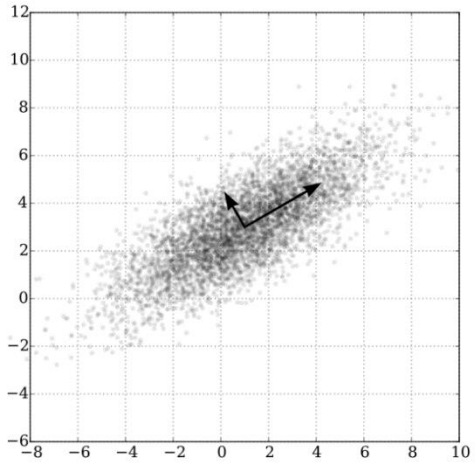
Here chose **Forward selection**, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.



picture from wiki: https://en.wikipedia.org/wiki/Stepwise_regression

PCA

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. If there are n observations with p variables, then the number of distinct principal components is min (n-1, p). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.



In this case, we only use PCA to deal with multicollinearity columns, do not consider about PCA to decrease dimensions before model building. It would make the model lose some information. If the data which is going to analysis is not too huge, we should prevent to use as possible as we can.

Standardization

Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. Standardization assumes that your data has a Gaussian distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression and linear discriminant analysis. Therefore, in order to make gradient descent for the regression analysis smoother, each feature is standardized before modeling. Then the values are rescaled to have a mean value of 0 and a standard deviation of 1.

Model Building and Testing

Split data

- The training data is split to 80% training and 20% validation in order to adjust model parameters in the process of modeling. However, in order to get the best parameter of each model, we used 10 folds cross-validation to train models.

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.

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

Where y^ is the predicted prevalence of undernourishment and "y" is the actual prevalence of undernourishment.

Linear regression

Linear regression is a basic and commonly used type of predictive analysis. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

- Submission result: Root Mean Square Error (RMSE) = 9.9467

Ridge and Lasso regression

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.

Ridge regression:

In ridge regression, the cost function is altered by adding a penalty equivalent (L2) to square of the magnitude of the coefficients. Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.

- Submission result: Root Mean Square Error (RMSE) = 9.7778

Lasso regression:

The only difference from Ridge regression is instead of taking the square of the coefficients, magnitudes are taken into account. This type of regularization (L1) can lead to zero coefficients. Some of the features are completely neglected for the evaluation of output. So Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.

- Submission result: Root Mean Square Error (RMSE) = 9.9470

XGBoost

XGBoost provides a gradient boosting. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way.

- Submission result: Root Mean Square Error (RMSE) = 11.9036

Random Forest

A random forest classifier is an ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables. This classifier has become popular within the remote sensing community due to the accuracy of its classifications. Here trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

- Submission result: Root Mean Square Error (RMSE) = 9.1373

Conclusion

This analysis has shown that the prevalence of undernourishment of a country can be confidently predicted from its characteristics. However, in the original dataset, it contained many noises that need to be dealt with. Moreover, the dataset also contain feature that are very insignificant to predict the prevalence of undernourishment. After some data cleaning process and implementation of feature selection, we can discover that in particular, fertility rate, open defecation increase, obesity prevalence, avg supply of protein of animal origin, access to improved water sources increases have a significant effect on the prevalence of undernourishment of a country.

According to this analysis, it indicates that the prevalence of undernourishment can be properly predicted by selected features. In addition to this, Random forest algorithm has the best performance among all models.

Further improvement

1. Dealing with missing values:
    i. Current use: KNN
    ii. Alternatives: mean, median, by models
2. Adjust outlier values:
    i. Current use: 3 times standard deviation
    ii. Alternatives: Interquartile Ranges
3. Feature selection:
    i. Current use: Lasso cross-validation
    ii. Alternatives: Ridge, change alphas, select by linear correlation…etc.
4. Fitting model:
    i. Current use: Random forest
    ii. Alternatives: Neural network, Adaboost…etc.

Competition Result:

Here is the best score of our predict model

BEST SCORE	CURRENT RANK	# COMPETITORS	SUBS. TODAY
9.0324	25	368	3 / 3