Project Luther - Summary

Title: What other factors can be used to predict cereal yield?

Wen Fung Leong

Intuitively, we are aware of the motivation to increase the world agricultural productivity is to ensure there is enough food to feed the growing world population that is estimated to double every 61 years [1]. Other than the total population, there must be other determinants (factors) that influence the productivity. A direct way to solve this is to find a linear model that can reveal the relationship between the factors of interests and total cereal yield in the world and in each country. The major steps to find the linear model and its coefficients are summarized below.

Data Sets (Python code: web_scraping.py)

Several websites were explored to search for possible factors (features). Among them, the websites that required payment for downloading the data were excluded. The remaining ones are government or nonprofit organizations, which are The Food Security Portal[2], The World Bank Data[3], Climate Change Knowledge Portal[4], Wikipedia[5], and Food and Agriculture Organization of United Nations[6]. The data for nineteen features' and target (i.e., cereal yield) downloaded or scraped from these website. Functions were written for these tasks (Please see to `web_scraping.py`). Majority of these features, stored in csv files, contain data starting from early 1960s to 2015 (or later) for each country, continent, and special region. Only the monthly and temperature and precipitation data are limited for time frame of 1991-2015. Due to differ in formatting and major lack of data in those downloaded files, only 14 features were considered for the model. Note that both Beautiful Soup and Selenium were used for gathering the data.

Data Cleaning (Python code: data_cleaning.py)

First, the average, minimum, and maximum of the temperature and precipitation data were extracted from the 400+ excel files that were downloaded from the Climate Change Knowledge Portal[4]. Second, each feature was converted to individual DataFrame format and was later combined to a big DataFrame for analysis. Third, while converting the features to their DataDrame, those with missing data were imputed via regression or substitution imputation. Third, the country codes that are shared by different data sets (including the Country and Continent Codes List [7]) were extracted, and there are 206 countries. The list drops to 188 countries after excluding the 88 countries that have no data for specific feature or target. Lastly, heatmap and plots were used to visualize the features' distributions, and the correlation relationship among them and the target. At this point, several features were dropped for the next step. The remaining number of features was 13. Functions were written for the handling the data (Please see `data_cleaning.py`).

Feature Engineering (Python code: feature_engineering.py)

Functions to calculate the RMSE, as well as to fit the linear regression and regularization models (Lasso and Ridge regressions) were written. First the countries were converted to categorical variables, resulting in an extra of 118 features. To determine the best set of features, the original 13 features were divided into several subsets to evaluate the goodness-of-fit and the residual plots of the three models. When heteroscedasticity was shown on the residual plots, log-transformed was applied on the target to mitigate this issue. Note that all these plots were generated from the standardized validation set, while the standardized training set were used to trained the models. After several rounds of analysis, all 13 plus 118 features were retained because the R-square was 0.921 and RMSE was around 0.2 for linear regression model, an indication that the fit is very good.

At one point, I tried to reduce the number of features by selecting longitude and latitude (from Koneill's country centroid [8]) as the features to represent the countries. Unfortunately, the less than 0.6 R-square value and lack of time stopped me from continuing the investigation.

Model Selection (Python code: linear_models_selection_CV.py)

LassoCV and Ridge CV (with 10 folds) were used to estimate the alphas' value for the Lasso and Ridge regression model. With these alpha values, the standardized training and validation sets were used in the cross validation procedure to select the best model. Linear regression seemed to be the best model, with R-square of 0.913; while Lasso regression came in a close second place with R-square of 0.910. Finally, the linear regression model was used to fit standardized training and validation sets in order to estimate the features' coefficients. After that standardized testing data were tested to ensure the Linear regression yielded high R-square and RMSE values.

Analysis on the coefficients: The large positive coefficients are associated with the agricultural arable land and average temperature, while the large negative coefficients are associated to percent rural population and employment. These four coefficients have high predictive power. It also tells us that temperature is an important determinant for high cereal yield; while precipitation doesn't have the same influence because most agricultural land have efficient irrigation system. The decline in rural population and percent employment in agriculture don't have significant influence on the increasing cereal yield, indicating the advancement in technology, and bioscience, and agricultural fields have been the leading factors that contribute to the high crop productivity.

References:

[1] Rosenberg, Matt. "Population Growth Rates." ThoughtCo, Jun. 14, 2018, thoughtco.com/population-growth-rates-1435469.

[2] The Food Security Portal. "Data Sets." Retrieved from http://www.foodsecurityportal.org/api

[3] The World Data Bank. "World Bank Open Data." Retrieved from https://data.worldbank.org

[4] Climate Change Knowledge Portal. Retrieved from http://sdwebx.worldbank.org/climateportal/index.cfm?page=downscaled_data_download&menu=historical

[5] International wheat production statistics. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/International_wheat_production_statistics

[6] Food and Agriculture Organization of United Nations. http://www.fao.org

[7]JohnSnowLabs. "Country and Continent Codes List." Retrieved from https://datahub.io/JohnSnowLabs/country-and-continent-codes-list

[8]Koneill. "Country centroids." Retrieved from https://worldmap.harvard.edu/data/geonode:country_centroids_az8