INVESTIGATION OF VARIOUS FACTORS INFLUENCING THE CROP YIELD GLOBALLY.

Ali Sehpar Shikoh | March 4, 2022 | BrainStation Data Science Bootcamp

PROBLEM STATEMENT:

This project aims to identify key parameters that play a significant role in the determination of crop yield, i.e., the mass of harvest crop product in a specific area [1]. One of the essential aspects that are reported to have a noteworthy impact on crop yield is climate change [2]. Apart from climate change, there are additional important variables (i.e., population, agricultural intake, etc.) that might be of interest. It is required to identify what kind of influence do the key variables have and to what extent. Identifying key features will provide world organizations with a road map to boost crop yield by the implementation of intelligent agricultural practices, thereby improving food security worldwide.

BACKGROUND:

Food security is profoundly important for human begins all over the world. Further, it has been demonstrated that increased crop yields tend to reduce poverty significantly [1]. With the ever-increasing population, food availability is becoming a bigger challenge. However, food security is always studied under various climate change scenarios. However, many additional factors, i.e., human population, water availability, agricultural inputs, etc., might also play a key role in determining the crop yield [3]. Therefore, it is required to integrate variables like the ones mentioned above for the evaluation of food security in a complete and systematic fashion.

DATA SOURCE:

The bulk data set was for this project was acquired from the Food and Agriculture Organization (FAO) database owing to its comprehensive nature [4]. The entire dataset (having a size of around 1 GB) is updated periodically with the latest version containing data till the year 2019. When unzipped, the dataset contained 65 individual files incorporating numerous variables related to crop production, food balances, trade climate change, forestry, etc. These variables were reported on a yearly basis (and in rare cases on a quarterly basis) for various countries and country groups/regions (e.g., Asia, ASIAN countries, etc.) of the world. Variables that were considered to have an effect on crop yield (e.g., human population, livestock units, fertilizer used, manure applied, etc.) were handpicked and integrated into the dataset. Some of the variables (e.g., CO_2 emissions, pesticides, and insecticide used) were later on discarded as they were reported from 1991 to 2019 instead of 1961 to 2019, thereby significantly reducing the number of rows.

DATA PROCESSING:

Exhaustive preprocessing of various raw datasets was carried out. While doing so, it was realized that some of the columns contained multiple variables (e.g., area harvested, crop production, crop yield) that could function as individual columns. Such single columns were divided into multiple columns. Further, when processing the primary raw dataset that was related to crops, it was realized that a multitude of crops were grown around the world; it was decided to select six main crop groups (i.e., sugar crops, vegetables, etc.) to form the basis of the processed dataset. Statistics related to individual crops, i.e., production and area harvested, were used to calculate the yearly total yield for various countries, which later acted as the target variable. Apart from that, datasets related to other variables were also cleaned with the null and infinite value slots discarded or filled appropriately. Further redundant columns (like 'Year Code', 'Domain', 'Unit' etc.) were also discarded with the useful information (e.g., Unit = tones) mentioned included in the appropriate selected column names.

EXPLORATORY DATA ANALYSIS (EDA):

In general, an increasing trend was observed for variables included in the dataset on a yearly basis. Ample focus was paid to exploring temperature data as it is a direct indicator of climate change. On average, the world's temperature has increased by 1.50 ˚C, with European countries like Serbia, Montenegro, and Luxembourg being the most affected. In contrast, only one country, i.e., Nauru, saw a decrease in temperature over the years. When considering global crop yield, an overall increasing trend was observed on a yearly basis. In the case of individual crops, sugar-related crops turned out to have the highest yield, following vegetables, roots, and tubers. Country-wise highest total yield was observed for Peru, Figure 1.

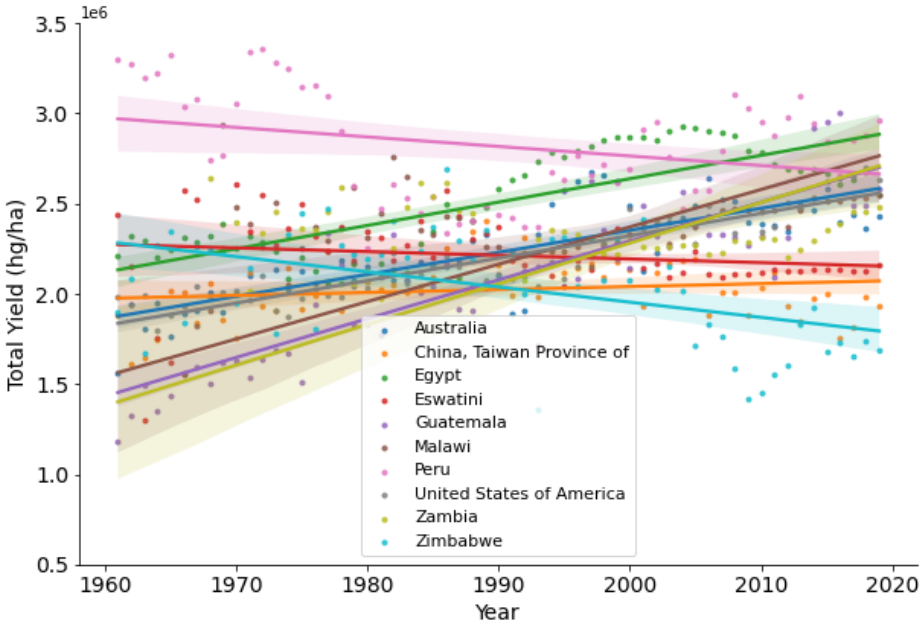

Figure 1: Annual yield of top ten countries plotted from 1961 to 2019.

Interestingly, many African countries (e.g., Egypt, Malawi, and Zimbabwe) were also noticed to have one of the highest total yields. However, when plotted on a yearly basis, the yield values were observed to decrease in two countries, i.e., Peru and Zimbabwe. Consistency in the annual yield for Taiwan was observed. When considering other independent variable columns, co-linearity was observed in among many of them.

MODELLING:

To start off with modelling, one hot encoding was utilized initially to convert the 'Area' column to represent each country mentioned as an individual column in the dataset in order to improve the R^2 value. Later, the encoded dataset was later combined with the original dataset. Later numerous regression models (four in total) were fitted to know more about the relationship between the independent and dependent variables. Initially, the models were fitted without any regularization and hyperparameter optimization. Later on, model-specific hyperparameters (at least two) were optimized (alongside regularization in the case of linear regression) and measured for the R^2 value. Further, all four models were evaluated based on various parameters, including mean squared error, mean absolute percentage error, median absolute error, and the residuals obtained after predicting the target variable using the test set.

Table 1: R^2 and mean absolute percentage error values for various models after hyperparameter optimization.

No.	Model	R^2 value for Test set (After Optimization)	Mean Absolute Percentage Error
1	XGBoost	0.957	0.120
2	Random Forrest	0.953	0.123
3	Decision Tree	0.893	0.172
4	Linear Regression	0.830	0.365

As seen in Table 1, the XGBoost regression model turned out to have the highest R^2 value alongside having the lowest means absolute percentage error after hyperparameter tuning.

FINDINGS:

The simplest model (i.e., a linear regression model with lasso regularization) and model with the highest R^2 value (i.e., XGBoost) were selected for further interpretation.

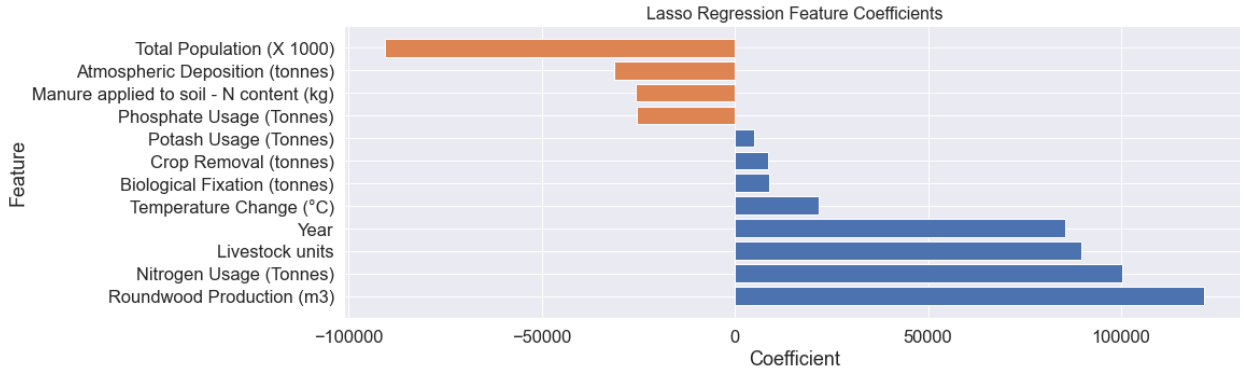

Figure 2: Lasso regression feature coefficients

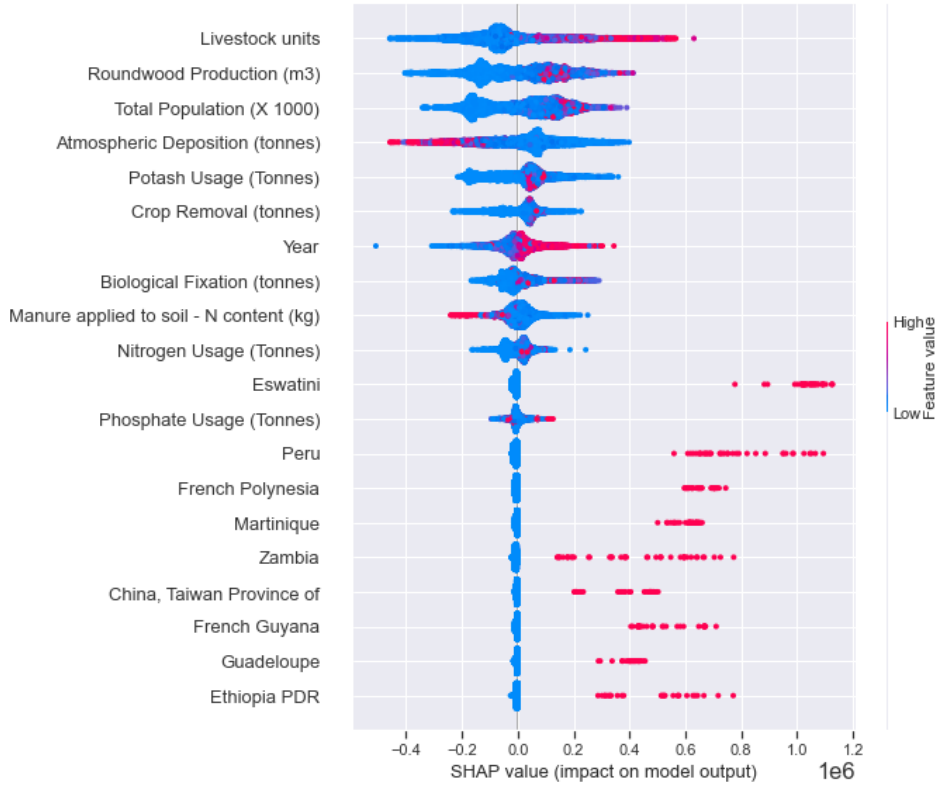

Figure 3: Graph related to the impact of various features and countries on model output obtained after fitting the dataset using XGBoost and interpreting using SHAP package.

To summarize, the key findings of the study are as follows:

- When interpreting Lasso regression, features like Roundwood production, usage of nitrogen-based fertilizers, and livestock units turned out to have the highest impact on increasing the total crop yield.
- Interestingly temperature tuned out to have a positive coefficient; however, it is believed not to be that impactful when compared with the above-stated features/variables. This was entirely unexpected.

- On the other hand, total populating and atmospheric deposition turned out to have the highest negative impact on the crop yield.
- In terms of countries, Peru, the People's Democratic Republic of Ethiopia, and Eswatini turned out to have the highest yield coefficients. In contrast, countries like USSR, Brunei Darussalam, and Djibouti turned out to have the highest negative coefficients.
- When interpreting the XGBoost model using SHAP, the total population turned out to have the highest negative impact irrespective of whether the population is high or low. On the other hand, livestock units turned out to have a positive effect on yield when their population was relatively high.

CONCLUSION:

An effort was made to know about the crucial features that tend to influence crop yield around the world the most. To do so, a dataset was gathered from the FAO website and processed to shape it in an acceptable format. Subsequently, exploratory data analysis was carried out on the combined and refined datasets. To find the significance of various features, four different models were fitted, optimized, and evaluated with the highest R^2 values (i.e., 0.957) achieved in the case of XGBoost. Later, two of the models were used for interpretation, i.e., lasso regression and XGBoost. Interestingly, the human population turned out to have the highest negative impact. Surprisingly, temperature change turned out to have a positive effect on yield.

NEXT STEPS:

The next step will be to incorporate additional independent variables that were collected from 1991 (instead of 1961 onwards) and tend to reduce the number of rows. These independent variables include agricultural emissions (i.e., carbon dioxide, methane, and nitrogen dioxide emissions) and agricultural inputs (i.e., insecticides and pesticides). Regression models were run on data gathered for the whole world and thus showed generic trends. It would be interesting to probe into country or region specific trends, especially in the case of countries showing a decline in positive dependent/independent variables (like crop yield) and an increase in negative variables (like temperature increase, emissions). This might also increase the number of variables available for data analysis as a comprehensive dataset with additional variables for all countries in the world is quite difficult.

REFERENCES:

[1] Liliane TN, Charles MS. Factors affecting yield of crops. Agronomy-Climate Change & Food Security; IntechOpen: London, UK. 2020 Jul 15:9.

[2] Shrestha S, Chapagain R, Babel MS. Quantifying the impact of climate change on crop yield and water footprint of rice in the Nam Oon Irrigation Project, Thailand. Science of the Total Environment. 2017 Dec 1;599:689-99.

[3] Kang Y, Khan S, Ma X. Climate change impacts on crop yield, crop water productivity and food security–A review. Progress in natural Science. 2009 Dec 10;19(12):1665-74.

[4] Faostat. [cited 2022Apr3]. Available from: https://www.fao.org/faostat/en/#home.