Food Security Analysis Report

Cheng Zhong

Sep 24, 2024

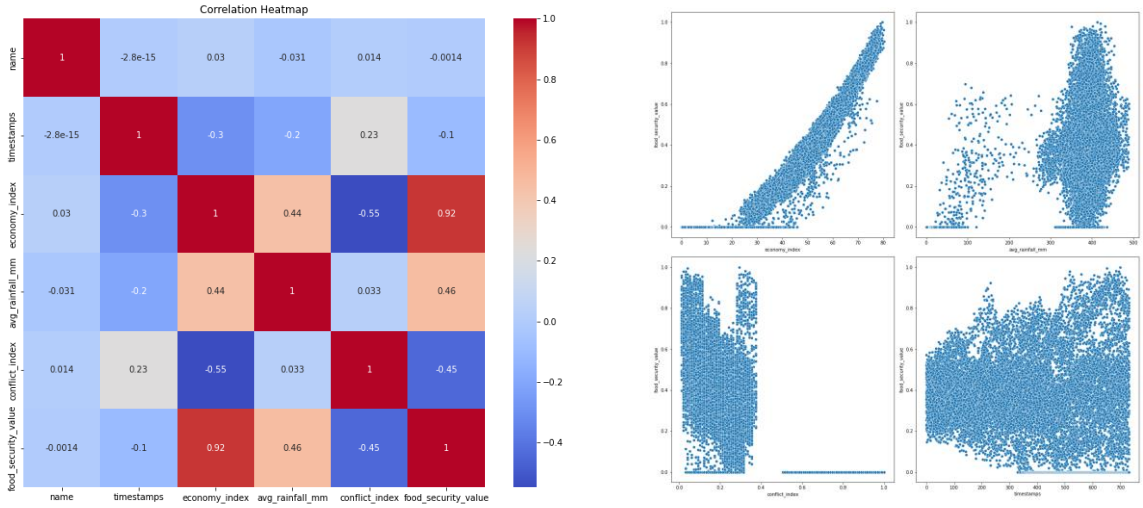
Data Preprocessing and Merging

Five datasets (*geometries*, *economy index*, *rainfall*, *conflict*, *food security*) were provided as raw input for the analysis, and data transformation techniques were used to unify the format, clean data, merging different sources, while keeping the authentic characteristics of data. For instance, the datasets are time series data with country specific information. Some countries have missing food security values which will be predicted later using inference models. Economy index, average rainfall, and conflict index are potential predictors.

These datasets require not only country key, but geometry information to merge. I converted geometry-based data to a common geospatial format under the same coordinate reference system. When merging *rainfall* and *geometries*, the coordinates and polygons do not have exact matches, so I adopted nearest spatial join of these two datasets based on the distance between their geometries, which will result in multiple rainfall entries for a single country. Then I calculated average rainfall to ensure each region has one record for one timestamp. Eventually, all datasets were merged based on spatial (country key or geometry) and temporal (timestamps) information.

Preliminary Analysis

After merging all datasets into one “unique source of truth”, I conducted exploratory data analysis and leveraged



visualization tools to derive some preliminary findings. The correlation heatmap shows there’s a strong positive correlation between economic index and food security, suggesting the countries with strong economies tend to have better food security. Average rainfall shows only a moderate positive correlation with food security, but the scatter plot shows extreme rainfall events (excess and deficit) could lead to decreased food security. The conflict index exhibits negative correlation with food security, confirming that countries experiencing high levels of conflict tend to have worse food security situations.

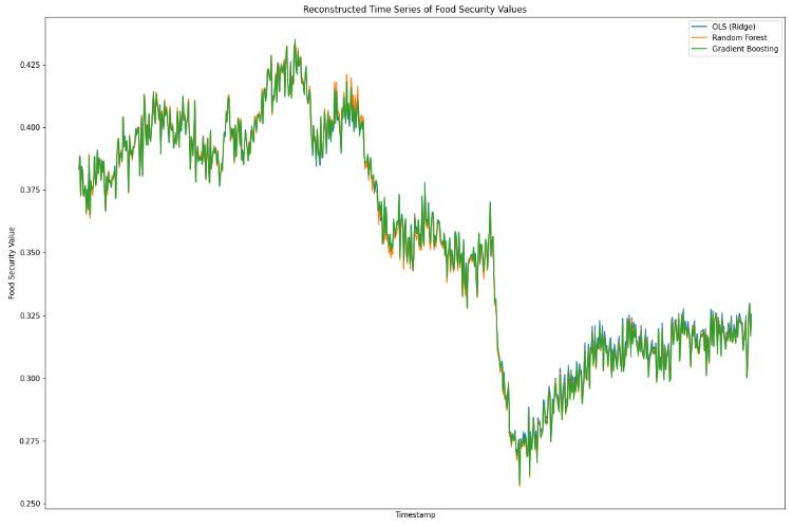
Inference Model and Reconstruction

Three models were adopted to infer food security scores for countries with missing data, including OLS (Ridge) Regression, Random Forest, and Gradient Boosting. These models were trained on countries with food security values

and predicted values for countries without. The OLS (ridge) model serves as a benchmark model, regularization was applied to handle potential multicollinearity and polynomial features could capture more complex relationships in the data. The Random Forest model performed well with high R2 score that the model explains about 92% of the variance in the target variable. The Gradient Boosting model performed marginally better than random forest, and both outperformed significantly then the ridge OLS. Both RF and GB have consistent performance across five folds split with time series cross-validation, and the average MSE and R2 are listed below.

- OLS (Ridge) - Average - MSE: 0.0072, R2: 0.7699
- Random Forest - Average - MSE: 0.0028, R2: 0.9204
- Gradient Boosting - Average - MSE: 0.0027, R2: 0.9237

Then feature importances were evaluated for all models. For Random Forest and Gradient Boosting, feature importances were derived from the model's feature importance attribute. For OLS, coefficient magnitudes were used to assess feature importance, with both linear and polynomial terms. The conclusion is economy index is the most important feature, while conflict index has very low importance. Counterintuitively, average rainfall has low importance, which may be due to the ability of importing and storing food when a country has strong economical power. Other interaction terms suggest nonlinear relationships exist behind the data. The best performing model is Gradient Boosting. Later, time series data were reconstructed using the prediction values for countries that have missing food security scores.



Summary

The tree-based models performed well in capturing general trends and country specific patterns. However, these nonlinear models may over relay on the economy index because the pseudo dataset contains very limited numbers of predictors and may have overfitting issues.

For future endeavors in complex real-world cases, more analysis could be implemented, trends and cyclical patterns of food security data may be explored; improve accuracy when merging with geographic information; geometry information could be one of the predictors; countries could be clustered with their geography, income, climate, etc.; tuning hyper parameters and avoid overfitting for the machine learning models.

With more context and information, this analysis could lead to more substantial discussions among economic factors, environmental conditions, social conflicts, and policy indications to improve food security with data-driven approaches.