Methodology

This section explains the main approach and techniques used in this paper. In summary, we use readily available data to model the food security status of village clusters in Uganda, Malawi, and Tanzania, with various machine learning related techniques.

1. Food Security measurement

We predict three measures of food security used by international humanitarian organizations including USAID and the World Food Programme (WFP): the reduced coping strategies index (rCSI), the household dietary diversity score (HDDS) and the food consumption score (FCS). Different from the results measurement in previous works (Lentz et al. (2019) and Jean et al., (2016), among others), this paper focuses on the categorical prediction for the given cutoffs. First of all, this is close to the actual policy scenarios where policymakers are trying to capture all the insecure households in a potential famine year. Predicting when and where the food security crisis will happen is more important than having an accurate assessment of the food security status of the general population. In technical terms, this study focuses on the recall rate of insecure households, rather than the overall prediction accuracy. This measurement makes sense as the crisis prediction bears a resemblance to the problem of anomaly detection: in most years and most areas, households are not in direct threat of hunger. The food insecure households are usually a small portion among the general population, even though the malnutrition problem is prevalent and persistent. Models aiming to maximize the overall accuracy tend to capture characteristics that are rich in the majority of the population and fails to understand the insecure households enough. To force the models to gain as much information as possible regarding the imminent threat of food shortage, we applied the downsampling and oversampling technique to force the model to learn about the tail of the distribution. These methods are broadly used to imbalanced datasets like the one we have about food security status.

2. Explanatory data

The variables used to predict food security are high-frequency data, including precipitation, temperature, market prices, soil quality, and geographic variables. These data are generally collected remotely and are widely available. Household roof type is used as a crude proxy of poverty that can be accurately captured from satellite imagery. Cellular phones are access to financial resources, market information, and remittance flow (Eagle et al. 2010, Blumenstock et al. 2016) also serve as significant predictors. Household-level data, including demographics and assets from LSMS, are also included.

3. Data segmentation and data split

For data segmentation, we compare the results of models trained on the entire dataset of three countries, with models trained by each country separately and then weight the recall rate by the number of households in each country. Similarly, we split the data even further by separate models trained on each region of the nations. Lastly, we train a shallow tree in each country based on observables to automatically split the data into several subsets.

We use three different training, and testing data split methods to feature various applications. Yearly split is using one year as a training set to forecast places where the famine may happen in the future. The regional division is used to predict rural or more remote areas given current information in more accessible regions. Random split is used as a demonstration of general prediction or the “best guess” on unknown households based on all the information that we have.

4. Classification Algorithms and result

In this paper, we tried baseline classification algorithms such as classification tree and SVM.

To reduce the overfitting problem involved with the large of variables we have in the model, we also tried ensemble learning such as random forest and XGboost. The ensemble learning methods improve model performances by averaging and sequentially improving the base trees.

Parameter tuning and feature importance analysis

For result measurement, we use the recall rate of the insecure category as the primary measure to evaluate the success of the models. Among models with similar recall-rate, we also compare other metrics, including AUC and precision, to find the best model.

5. Error analysis of the primary model by region, by group, by month: presented in table and map.

6. Model generalization issues: what happens when we directly apply models trained in one country to predict another.

7. Model deploy and update: compare the results of using one year, with a dynamic process of continually updating the model with new survey data