**Predict Food Security with Machine Learning**

Yujun Zhou, Kathy Baylis, Erin Lentz, Hope Michelson

July 25, 2019

**Abstract**

Food insecure and famine have reemerged in the recent decade. Identifying food insecurity crises rapidly and accurately can speed up humanitarian responses to mitigate casualties from hunger and save lives. Our model is based on readily available data on prices, geography, and demographics that are spatially and temporally granular. By incorporating machine learning techniques, we are able to further improve the accuracy of capturing the clusters of households facing potential threat of hunger. The same procedure is applied to three sub-Saharan African countries: Malawi, Tanzania and Uganda to predict out-of-sample. We are able to capture categorize 69-88 percent of the insecure clusters, which is ZZ percent higher compared to the baseline model. Our result show that a data-driven model with the help of machine learning methods can greatly improve its performance on capturing the food insecure households despite the imbalance in the data. We aim to further develop the data flow into a system that is automatically updated, generalizable, scalable, and cost-effective in predicting areas of potential food shortage to assist responses to food insecurity crisis.

**Keywords:** food insecurity, machine learning, early warning, Sub-Saharan Africa, famine

**Predict Food Security with Machine Learning**

1. **Introduction**

We lack the ability to identify food insecure populations in time to intervene. Humanitarian response tends to trail the onset of food security crises. Currently, governments and NGOs in the sub-Saharan Africa region use the Integrated Food Security Phase Classification System (IPC) as the early warning system. The IPC is updated quarterly for each livelihood zones, which makes it difficult for the system to detect specific villages that are in danger of hunger in the near term. Recent increase in available data related to food security, geography, rainfall, and market price for food provides us with the opportunity of making the food security predictions to a finer geographic level and updates at monthly or even weekly frequencies.

Following the framework in Lentz et al. (2019), this paper incorporates these data into a single predictive model. We expand the study area to Malawi, Uganda, and Tanzania to test the framework with more heterogeneity in geography, environment, and socioeconomic status. We use the same framework of incorporating data at different scales and the same procedure of transforming the raw data to usable forms for these three countries. The machine learning algorithms used for prediction are the same for the countries but the hyperparameters are tuned on the training dataset of each country separately. This makes the model generalizable for other data scarce countries and areas with some previous household survey data (LSMS or DHS), but to provide insights on food security that are both spatially and temporally granular.

With an emphasis on capturing the majority of food insecure households, this paper uses various machine learning techniques to increase the performance in detecting the insecure food clusters. The primary goal of food security prediction is thus identifying the minority class of food insecure among the imbalanced data where the majority of the data in most years are secure (even though malnutrition is still a problem). Similar to a fraud detection problem, severe food insecurity crises are rare but too important that we cannot afford to miss a single one of them. It usually costs more to miss a village under food shortages than to falsely sending food assistance to areas that are not in desperate need of them. Therefore, we concern more with recall (percent of truly food insecure clusters that were classified as such) than precision (percent of positive classifications that are truly positive) in

the trade-off between recall and precision. This is important because choosing the right metric to optimize matters for model selection and parameter tuning and ultimately the ability to provide desirable performance on correctly detecting the minority class that are food insecure. This paper uses a Cost-sensitive learning as well as a weighted f-1 score to maximize the recall without too much sacrifice on precision.

However, the ballows us to create a balanced data-set that, in theory, should not lead to classifiers biased toward one class or the other. However, in practice, these simple sampling approaches have flaws. Oversampling the minority can lead to model overfitting, since it will introduce duplicate instances, drawing from a pool of instances that is already small. Similarly, undersampling the majority can end up leaving out important instances that provide important differences between the two classes. SMOTE, which actually creates new instances of the minority class by forming convex combinations of neighboring instances. This allows us to balance our data-set without as much overfitting, as we create new synthetic examples rather than using duplicates. This however does not prevent all overfitting, as these are still created from existing data points.

We aim to further develop the data flow into a system that is automatically updated, generalizable, scalable, and cost-effective in predicting areas of potential food shortage.

The procedure is ready for constant updates with new data coming to make the most up-to-date predictions.

To build an early warning system of food security in areas where data are scarce and data collection is costly (Hutchinson,1991)

That captures the majority of food insecure households through data techniques

That is automatically updated, generalizable, scalable and cost-effective

Build ML models to predict cluster-level food security status for targeting, aid purposes in times of food shortage

Use LSMS data for Malawi, Tanzania and Uganda as ground truth

Use market price of food staples, weather shocks in growing seasons, and geospatial features around clusters to predict potential food security challenges

Use data techniques (oversampling, data segmentation) to improve prediction performance

Correctly categorize 69-88 % of food insecurity categories



Faster response during food crises saves lives and resources. Crises are increasing in frequency and severity in many parts of the world. Identifying the scale and scope of these crises in a timely and accurate fashion is essential for food aid and humanitarian responses. However, policymakers often lack the information required to identify the right populations to target programming and resources (Barrett and Headey 2014). By 2012, only 27 of Africa’s 48 countries had conducted at least two comparable household level surveys (Beegle et al. 2016) because it is costly to do so. The data gap hinders the efforts to effectively targeting the population in need and calls for the use of data and method that are cost-effective and accurate.

Novel data and data methods can be used to fill this data gap. Nightlights data (Chen and Nordhaus 2011; Henderson et al. 2012.) can serve as a proxy for economic activity, especially when comparing across countries. However, in remote rural or better off urban areas, the nightlight intensity varies little over time, hiding substantial changes in economic outcomes. Mobile phone data (Blumenstock et al.,2015; Steele et al.,2017) is more frequent and less expensive compared to census surveys. However, in the short term, it is not feasible to roll out cellphone surveys in entire sub-Saharan Africa, and the biases associated with using relying on cell phone-sourced information to infer population statistics are as of yet, not well understood. Very high-resolution satellite imagery is becoming cheaper but suffers the lack of structure (Engstrom et al., 2017; Donaldson and Storeygard, 2016). Recent studies have combined Convolutional Neural Network (CNN) models and transfer learning (Jean et al., 2016; Babenko et al. 2017) to make an inference based on the information in the satellite imageries. These models can explain up to 60% - 75% of the variation at the village level wealth and asset measures in several sub-Saharan Africa countries. However, the reliance on the information in the satellite imagery (specifically, building size, roof type, road conditions) limits its performance on development indicators other than wealth or assets. Head et al. (2017) apply the Jean et al. (2016) approach to a set of various development indicators and across several countries. Their research finds that the prediction performance degrades quickly on health and nutrition outcomes (no better than random guessing in some cases). The reliance on nightlight data on this approach also limits the prediction accuracy when applied in countries with different socioeconomic conditions. The external validity and interpretability of this deep learning-based approach call for a method tailored for food security predictions.

This paper uses data-driven framework to predict the onset of food crises. Combining remote sensing data with household surveys and price data, the model is able produces the most spatially and temporally granular predictions of food security. With an emphasis on the structure of the prediction error, this paper uses various machine learning techniques to increase the accuracy and reducing the type II error in predicting food security status. The empirical application of the method is in Malawi, Uganda and Tanzania, using the Living Standard Monitoring Survey (LSMS) as the reference data.

1. **Data:**

***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.

***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.

We plan to use readily available data to model the food security status of village clusters in Uganda, Malawi, and Tanzania. 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).

The variables used to predict food security are high-frequency data including precipitation, temperature, market prices, soil quality and geographic variables, which 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 important predictors. Household-level data including demographics and assets from LSMS are also included.

1. **Method**

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. Categorical vs. continuous: focus on the categorical prediction for the given cutoffs of each food security measures

Focus on the categorical prediction for the given cutoffs of each food security measures.

Recall rate of the insecure villages is more important than the overall accuracy.

Typically, classifiers are more sensitive to detecting the majority class and less sensitive to the minority class.

Close to the actual policy scenarios where policy makers need to locate places with most insecure households.

1. Result metrics: recall, precision, AUROC, instead of accuracy

Methodology: Result Measurements

Recall (are we getting all the insecure households ?)

Precision (are we mistakenly categorizing secure households as insecure?)

f-1 score (balance recall and precision)

Overall categorical accuracy

Methodology: Detecting Rare but Relevant households

Cost-sensitive learning: penalize misclassifications of the minority class more heavily by having a cost function, which is equal to the inverse of the class proportions

extra reward for identifying the minority class over the majority class, Cost-sensitive learning changes this, and uses a function C(p, t) (usually represented as a matrix) that specifies the cost of misclassifying an instance of class t as class p. penalize misclassifications of the minority class more heavily than we do with misclassifications of the majority class, in hopes that this increases the true positive rate. A common scheme for this is to have the cost equal to the inverse of the proportion of the data-set that the class makes up. This increases the penalization as the class size decreases.

***Baseline Model***

* Classification Tree
* Hyperparameters (e.g. max\_depth, max\_features) automatically selected by training data
* Data split: year split
* Data segmentation : by country
* Down/over sampling: None

Variable groups:

1. Mkt : food price, market thinness
2. Asset: cellphone ownership, floor/roof material, asset index
3. Weather: dry spells, average temperature and rain
4. Location: elevation, distance to road, urban/rural

***Sampling: down-sampling, over-sampling, and SMOTE***

Apply the down sampling, over sampling, and synthetic data techniques to force the model to learn about the tail of the distribution

Down sampling, Over sampling and Synthetic data

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.

***Classification Algorithms***

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.

For structured data like ours (unlike text or pure image), tree-based methods are popular.

Tree-based: Classification Tree (baseline and base learner), Random Forest, Extreme Gradient boosting

1. Classification Tree (baseline and base learner)
2. Random Forest ( parallel )
3. Extreme Gradient boosting (sequential)

Xgboost performs better in capturing the minority classes than random forest, especially with oversampling techniques

b. Anomaly detection: clustering methods, One-class SVMs, and Isolation Forests

clustering methods, One-class SVMs, and Isolation Forests

Boosting (Freund and Schapire 1997) is very similar to bagging but trees are grown sequen- tially. Each tree that is grown uses information from the previously grown trees. In boosting we don’t use bootstrapping, but fit the tree on a modified version of the original data set. In this ensemble technique, errors are corrected by sequentially adding new models to the existing models until no more improvements can be made. Boosting algorithms include some parameters which slow down the learning process. This slow learning process gives us better predictions (James et al. 2013). Another method related to boosting is the gradient boosting which was developed in several papers such as Breiman (1996); Friedman, Hastie, and Tibshirani (2000); Friedman (2001). In this method, new models are created to pre- dict the errors of the previous models. The gradient descent algorithm is used to optimize an arbitrary differentiable loss function while adding new models, it can be thought as a combination of gradient descent and boosting

Extreme Gradient Boosting (xgboost), developed by Chen and Guestrin (2016), is an efficient, flexible and portable variant of the gradient boosting model of Friedman, Hastie, and Tibshirani (2000) and Friedman (2001). Xgboost has been a winning tool for several Machine learning competitions (Adam-Bourdarios et al. 2015; Chen and Guestrin 2016). Xgboost relies on the same principles with gradient boosting but compared to gradient boosting, it is more efficient and faster since it uses sparsity aware algorithms and better processor utilization.

***Data Split***

How year split prevails the other, quote Robert et al 2017

***Data Segmentation***

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.

By country

Entire dataset of three countries

By urban and rural

4. Auto-segmentation by training a shallow tree in each country based on observables

***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.

***Error analysis***

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

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

***Model deploy and update***

compare the results of using one year, with a dynamic process of continually updating the model with new survey data

The models are trained using a training set in one year. The accuracy of the models is evaluated using out-of-sample data in another year.

Our model tries to explain these variations in food security by the spatial-temporal variation in food availability and food access. Specifically, we align weather data with the crop growing season to describe the temporary shocks in food availability. We also align households with their most relevant market price, as shocks to income and household consumption budget.

To deal with the high dimensionality and nonlinearity problem, we apply machine learning methods including regularization methods (LASSO, Elastic net) and Ensemble learning models (Random Forest and Gradient boosting) to improve prediction accuracy. Another issue with the prediction accuracy is that the model works well on the quality measures of food security (FCS and HDDS), but a lot worse on the quantitative measure (rCSI). Partially due to the survey questions to the construction of the variable, the distribution of the rCSI appears to be highly skewed, a long-tail with a mass of points around zero. We employed oversampling and down sampling techniques to improve prediction performance.

The choice of the best model is the key in this paper. The criterion for choosing the best model is a fixture of prediction accuracy, recall rate (reducing type II error) and model interpretability.

**4. Results**

Currently, the prediction accuracy in around an r squared of 0.6– 0.7 at the cluster level and around 0.4 at the household level. Preliminary results suggest assets related variables are explaining most of the variance in food security: cellphone ownership, asset index, and roof types. Market access variables, for example, distance to road and distance to markets are also important. Total rainfall, temperature and the start of the rainy season play a big role as well.

Table 1: Baseline vs ML algorithms (year split) : Higher accuracy, similar recall

**5. Discussion**

**6. Conclusion**

Combined with data techniques, machine learning methods not only improve prediction accuracy in general, but particularly on households that are vulnerable to food price shocks.

An automated, updated and scalable food security system based on publicly available data, advanced data techniques can assist the work of food aid and humanitarian responses in a timely, transparent, and efficient fashion

**Table 1: Baseline vs ML algorithms: year split**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Food Security Measure | Overall  Accuracy  (baseline) | Overall  Accuracy  (ML) | Recall Rate Insecure category  (baseline) | Recall Rate Insecure category  (ML) |
| Malawi  2010/11, 2013 to predict 2015/16 | FCS | 0.70 | 0.69 | 0.04 / na | 0.00 / na |
| HDDS\* | 0.49 | 0.73 | 0.49 / na | 0.66 / na |
| rCSI | 0.39 | 0.69 | 0.23 / 0.00 | 0.79 / 0.00 |
| Tanzania  2010/11, 2012/13 to predict 2014/15 | FCS | 0.89 | 0.88 | 0.02 / 0.00 | 0.21 / 0.00 |
| HDDS\* | 0.79 | 0.80 | 0.99 / na | 0.95 / na |
| rCSI | 0.63 | 0.63 | 0.37 / 0.00 | 0.38/0.00 |
| Uganda  2010/11 to predict 2012 | FCS | 0.74 | 0.77 | 0.37/0.11 | 0.16 / 0.00 |
| HDDS\* | 0.88 | 0.88 | 1 / 0.00 | 1 / 0.00 |

**Table 2: Baseline vs ML algorithms with down/over sample technique**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Food Security Measure | Recall Rate Insecure category  (Baseline) | Recall Rate Insecure category  ML + Oversample | Recall Rate Insecure category  ML + SMOTE | Recall Rate Insecure category  ML +   ADASYN |
| Malawi  2010/11, 2013 to predict 2015/16 | FCS | 0.04 / na | 0.93 / na | 0.10 /na | 0.16 /na |
| HDDS\* | 0.49 / na | 0.68 / na | 0.83 / na | 0.71 / na |
| rCSI | 0.23 / 0.00 | 0.80/ 0.00 | 0.16 / 0.00 | 0.13 / na |
| Tanzania  2010/11, 2012/13 to predict 2014/15 | FCS | 0.02 / 0.00 | 0.63/0.00 | 0.49/0.00 | 0.58/0.00 |
| HDDS\* | 0.99 / na | 0.66/na | 0.92/na | 0.85/na |
| rCSI | 0.37 / 0.00 | 0.44/ 0.00 | 0.42 / 0.00 | 0.46 / 0.00 |
| Uganda  2010/11 to predict 2012 | FCS | 0.37/0.11 | 0.24/ 0.00 | 0.44/0.33 | 0.43/0.33 |
| HDDS\* | 1 / 0.00 | 0.89/0.00 | 1 / 0.00 | 1 / 0.00 |

**Table 3: Data Split Comparisons**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | Food Security Measure | Year split | | Regional split | | Random split | |
|  |  | Accuracy | Recall | Accuracy | Recall | Accuracy | Recall |
| Malawi | FCS | 0.80 | 0.93 / na | 0.79 | 0.59 / 0.00 | 0.94 | 0.93 / 0.00 |
| HDDS\* | 0.68 | 0.68 / na | 0.90 | 0.89 / na | 0.94 | 0.93/ 0.00 |
| rCSI | 0.71 | 0.80/ 0.00 | 0.80 | 0.64 / na | 0.93 | 0.83 / 0.00 |
| Tanzania | FCS | 0.80 | 0.93 / na | 0.79 | 0.59 / 0.00 | 0.94 | 0.93 / 0.00 |
| HDDS\* | 0.68 | 0.68 / na | 0.90 | 0.89 / na | 0.94 | 0.93/ 0.00 |
| rCSI | 0.71 | 0.80/ 0.00 | 0.80 | 0.64 / na | 0.93 | 0.83 / 0.00 |
| Uganda | FCS | 0.76 | 0.24 / 0.00 | 0.77 | 0.27 / 0.00 | 0.79 | 0.56 / 0.60 |
| HDDS\* | 0.69 | 0.89 / 0.00 | 0.51 | 0.49 / 0.00 | 0.75 | 0.77/ 0.00 |

**Table 4: Data Segmentation Comparisons (keep the same testing set)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Overall | Food Security Measure | By country  With oversample and ML | One dataset  With oversample and ML | Rural Clusters only  With oversample and ML |
| Malawi  predict 2015/16 | FCS | 0.93 / na | 0.00 / na | 0.01 / na |
| HDDS | 0.68 / na | 0.62 / na | 0.46 / na |
| rCSI | 0.80/ 0.00 | 0.08 / 0.00 | 0.53 / na |
| Tanzania  predict 2014/15 | FCS | 0.63/0.00 | 0.00 / 0.00 | 0.19 / 0.00 |
| HDDS | 0.66/na | 0.96 / na | 0.97 / na |
| rCSI | 0.44/ 0.00 | 0.08 / 0.00 | 0.86 / 0.00 |
| Uganda  predict 2012 | FCS | 0.24/ 0.00 | 0.00 / 0.00 | 0.16 / 0.00 |
| HDDS | 0.89/0.00 | 0.89/ 0.00 | 0.91 / 0.00 |

****

**Fig. 1 Map of FCS**

****

**Fig. 2 Top tree split**

****

**References**

Babenko, Boris, et al. "Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico." *arXiv preprint arXiv:1711.06323* (2017).

Barrett, Christopher B., and Derek Headey. "A proposal for measuring resilience in a risky world." (2014).

Beegle, Kathleen, et al. *Poverty in a rising Africa*. The World Bank, 2016.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata." *Science* 350.6264 (2015): 1073-1076.

Castelluccio, Marco, et al. "Land use classification in remote sensing images by convolutional neural networks." *arXiv preprint arXiv:1508.00092* (2015).

Chen, Derek. "Temporal Poverty Prediction using Satellite Imagery." (2017)

Chen, Xi, and William D. Nordhaus. "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences* 108.21 (2011): 8589-8594.

Dang, Hai-Anh, Dean Jolliffe, and Calogero Carletto. "Data gaps, data incomparability, and data imputation: a review of poverty measurement methods for data-scarce environments." (2017).

Donaldson, Dave, and Adam Storeygard. "The view from above: Applications of satellite data in economics." *Journal of Economic Perspectives* 30.4 (2016): 171-98.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. "Poverty from space: using high-resolution satellite imagery for estimating economic well-being." (2017).

Head, Andrew, et al. "Can Human Development be Measured with Satellite Imagery?" *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. ACM, 2017.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. "Measuring economic growth from outer space." *American economic review* 102.2 (2012): 994-1028.

Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

Kussul, Nataliia, et al. "Deep learning classification of land cover and crop types using remote sensing data." *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017): 778-782.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*(46), E9783-E9792.

Steele, Jessica E., et al. "Mapping poverty using mobile phone and satellite data." *Journal of The Royal Society Interface*14.127 (2017): 20160690.