Robust Statistical Learning for Food Security Forecasting

Gewei Cao

July 6, 2023

Abstract

Your abstract.

1 Introduction

"*Zart wäre einzig das Gröbste: daß keiner mehr hungern soll*" —-*Theodor W. Adorno*

In 2015, United Nations established Sustainable Development Goal 2 (SDG2), aiming to create a world free of hunger by 2030. However, with the breakout of Covid-19, WFP, UNICEF et al. (2022) estimated that there will be 78 million more undernourished people in 2030 than in a scenario in which the pandemic had not occurred. Meanwhile, the war in Ukraine hindered the recovery from the pandemic and disrupted the international food and energy market, pushing inflation higher (IMF (2022)). Due to this conflict, low-income food-deficit countries that rely on the import of food and fertilizers will suffer more from the high price of these kinds of commodities (WFP, UNICEF et al. (2022)).

According to the definition of the Food and Agricultural Organization of the United Nations (FAO), a person is food insecure when they lack regular access to enough safe and nutritious food for normal growth and development and an active and healthy life. In the year 2021, there were approximately 768 million people in this world facing hunger (undernourishment) (WFP, UNICEF et al. (2022)). Therefore, food insecurity is a crucial humanitarian and development challenge for the world.

From the perspective of the economy, food insecurity negatively affects economic development in several ways. Firstly, it can raise the risk of poor health outcomes such as malnutrition, obesity, and stunting, which can hinder the physical and cognitive development of children (e.g. WFP, UNICEF et al. (2022) Kang et al. (2018), Guerrant et al. (2008)). Food insecurity also increases the risk of chronic diseases (e.g. Seligman, Laraia and Kushel (2010),Weaver, Fasel et al. (2018),Laraia (2013)). Above both will reduce the productivity of the affected population, and increase the cost of health expenditure of the society.

In addition, food insecurity can contribute to social and political instability. For example, food insecurity can perpetuate violent conflicts, and contribute to a vicious cycle (Brinkman and Hendrix (2011), Hendrix and Brinkman (2013)). Food insecurity is also an important determinant of migration intentions and preparations (Smith and Floro (2020)). Furthermore, during Covid-19, the increased food insecurity may increase the migration pressure (Smith and Floro (2020)).

Food insecurity is apparently a critical humanitarian and development issue, particularly in Africa. According to WFP, UNICEF et al. (2022), in the year 2021, more than one-third of people in hunger are in Africa, the prevalence of undernourishment in Africa is 20.2%, and in Eastern Africa, it is 29.8%, in other words, there are around 136.4 million of people suffered undernourishment. Meanwhile, according to FAOSTAT, in 2021, the prevalence of severe food insecurity in Eastern Africa is 28.7%, and the prevalence of severe or moderate food insecurity is 66.9%. Africa, especially Eastern Africa, is facing a pressing food security challenge. The United Nations World Food Programme (WFP), Action Against Hunger, and many other humanitarian projects are concerning food insecurity in Africa. These projects need to identify or forecast food insecurity effectively, in order to make decisions and take action.

Moreover, predicting the occurrence of a food crisis or food insecurity helps policymakers and

international organizations to mitigate or prevent such negative events. Instead of causal inference, prediction is more important in the food insecurity context (Kleinberg et al. (2015)). With the development of machine learning (ML) and big data, starting with the study of Okori and Obua (2011), such data-driven technique implementations are emerging in this field (e.g. Lentz et al. (2019), Andree et al. (2020), Browne et al. (2021)). However, under the shocks of Covid-19 and the war in Ukraine, the robustness of the ML model for food insecurity prediction need to be considered. This study will focus on using ML to predict food insecurity in the Republic of Uganda, given the data and context of the breakout of Covid-19, trying to find whether exists a relatively robust ML algorithm when facing such shock and trying to find a high-performance ML model for the predicting task in Uganda.

This section is the introduction of the background and the motivation of this study. The next section will introduce the basic information and the food security situation in Uganda. Section 3 will ....

2 The Republic of Uganda

The Republic of Uganda, situated in Eastern Africa, is a landlocked nation bordered by Kenya to the east, South Sudan to the north, the Democratic Republic of the Congo to the west, Rwanda to the southwest, and Tanzania to the south. A significant part of Lake Victoria, which is shared with Kenya and Tanzania, is located in the southern region of the country. With an average elevation of 900 meters above sea level, Uganda's eastern and western borders are marked by mountains. According to the Uganda National Household Survey 2019/20 (UNHS), there are 40.9 million population in Uganda, and 49% of them are male. Meanwhile, 54% of the population in Uganda is below 18 years old. Within the working population, 68.1% of them work in the sector of agriculture, forestry and fishing.

Furthermore, Uganda's poverty rate is 20.3% in UNHS 2019/20, which was 18.7% in 2019, but after the breakout of the pandemic, it raised to 21.9%. In the round of the October/November 2022 Uganda-High-Frequency Phone Survey (HFPS), 56 % of the population was moderately food insecure and 15 % was severely food insecure. Thus, food insecurity in Uganda is severe, less than half population is able to access or buy enough food. Besides the shock of covid-19, additionally, the ongoing conflict in Ukraine, which has disrupted global trade, poses further challenges to food security in Uganda. For instance, according to the FAOSTAT data, in 2020, Uganda's agricultural fertilizer use of the nutrient nitrogen, phosphate P2O5, and potash K2O is fully dependent on import, but the Ukraine war is disrupting the supply chain of fertilizer (WFP, UNICEF et al. (2022)).

In addition, in UNHS 2019/20, the average Dietary Energy Consumption (DEC) in Uganda is 2393 kcal/person/day, more specifically, before the pandemic, DEC was 2437 kcal/person/day, while during the pandemic, DEC decreased to 2359, this number varies among different regions as well. Notably, 47% households in urban areas are classified as food poor by the Uganda government, and it is 22% in rural areas. Specifically, on average, Staples (cereal, roots and tubers) are consumed on a daily basis while meat, fruit and milk products are the least consumed in a week, we can find such patterns in the data description section as well.

In conclusion, as an Eastern African country, Uganda has issues of poverty and the threat of significant food insecurity. Accurate predictions of food insecurity can assist the people of Uganda, as well as public sectors and international humanitarian organizations, in implementing more cost-effective strategies to achieve development goals, such as the Uganda Nutrition Action Plan or SDG2.

3 Statistical Learning and Machine Learning

3.1 Introduction

Statistical learning refers to a set of tools for making sense of complex datasets (James et al. (2013)), and it is closely related to the concept and practice of Machine Learning (ML). Machine learning manages to fit complex data without simply overfitting and is able to give good out-of-sample predictions (Mullainathan and Spiess (2017)), and statistical learning theory provides the theoretical basis for many machine learning algorithms (Von Luxburg and Schölkopf (2011)). The

key idea of learning in this context is to find general patterns for a given sample data and utilize such patterns for different purposes. If we want to make predictions on output data y for input data x, it is called supervised learning (e.g. OLS), and if we only want to learn the structures and relationships for input data x but without output y, it is called unsupervised learning (e.g. Principal Component Analysis). In this paper, we only focus on supervised learning.

Generally, ML solves problems of regression and classification. Regression here means the output variable y is a continuous real-valued variable, and classification means y is a categorical variable, usually y is binary. Most supervised ML algorithms need a measure of how well they are performing when fitting a model, the measure is called loss function, $L(\hat{f}(x), y)$, parameters or functions which can minimize the expectation of loss function $E_{(x,y)}(L(\hat{f}(x), y))$ are the best parameters or functions we want. Note that we assume there are p features, i.e. $x_i \in R^p$ and X is a $N \times p$ matrix, where N is the number of observations.

In this paper, food security prediction is a classification problem, i.e. the response variable y is binary. Classification has advantages because one can construct multiple metrics to evaluate the performance of ML models and make policy implications. We call a food insecure observation a positive case and a food secure observation a negative case in the following table.

		True classes		
		Positive	Negative	Total
Predicted classes	Positive	TP	FP	P^*
	Negative	FN	TN	N^*
	Total	P	N	$P + N = P^* + N^*$

There are two types of important errors in the classification, Type I error (false positive) of inclusion (i.e., targeting those who are not food insecure) and Type II error (false negative) of exclusion (i.e. not targeting those who are insecure). Usually, type II error is more important in the food insecurity context. Some widely used metrics in classification are:

- Accuracy: $Accuracy = \frac{TP+TN}{P+N}$ totally correct prediction
- Precision: $Precision = \frac{TP}{P^*}$ correct prediction over predicted positive
- Recall (or Sensitivity): $Recall = \frac{TP}{P}$ correct prediction over true positive
- False positive rate: $FPR = \frac{FP}{N}$ false prediction over true negative
- F1 score: $F1 = \frac{2TP}{2TP+FP+FN}$ harmonic mean of the precision and recall

In addition, the receiver operating characteristic (ROC) curve is used to show the performance of a classification model at all classification thresholds. The ROC curve shows True Positive Rate (TPR) (or called recall) and False Positive Rate for different thresholds. The closer the curve to the top left corner, the better the model. Because this means there exists a threshold with low FPR and TPR. To measure such a pattern, the area under curve (AUC) is a good measure. The higher AUC, the more effective the model.

We will compare different algorithms in this food security prediction issue by comparing the AUC for each model. The most basic algorithm (benchmark) is Logistic Regression, and other tree-based algorithms will be used. Finally, we want to find the most robust and precise algorithm.

3.2 Learning Working Flow

3.2.1 Data cleaning

After collecting UNHS data and other open-source data, we

3.2.2 Training & Testing data split

3.2.3 Problem of imbalanced data

Our data is imbalanced, the proportion of food severely insecure ("Poor" class) is less than 5%. When data is imbalanced, the ML algorithms tend to treat the minority class as noise and ignore their influence. For example, if we have 100 households, and only 1 of them is food insecure, then

if we make a naive guess that any given household is food secure, we will still obtain a 99% accuracy.

To avoid the above problem, we have two potential solutions. Firstly, we can adjust the model, and make it more sensitive to the minority class, such as changing the metrics in the hyper-parameter tuning step or changing the probability threshold of being food insecure.

Secondly, we can use resampling techniques to generate synthetic samples for the minority class. We will implement two techniques, and compare their performances. The first is the Synthetic Minority Oversampling Technique (SMOTE), which create new synthetic data by interpolation between several positive instances that lie together (Fernández et al. (2018)). The second is the Adaptive Synthetic Sampling Approach (ADASYN), which is based on the idea of adaptively generating minority examples according to their distributions (Fernández et al. (2018)).

Finally, other techniques such as class weighting or more complicated resampling techniques will not be included in this study. The techniques above are expected to help us overcome the issue caused by imbalanced data, and we will compare their performance in further sections.

3.2.4 Hyper-parameter tuning

3.2.5 Performance evaluating

3.3 Logistic Regression

Logistic regression is quite famous in econometrics, usually, it is solved via maximum likelihood estimation, and the loss function L_{log} is

$$L_{log} = -\sum_{i=1}^{N}[ln(1-p_i) + y_i ln(\frac{p_i}{1-p_i})]$$

where

$$p_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

If we simply minimize the L_{log} and find the β vector, in such high dimensional and complex food security prediction problem, there is a risk of over-fitting. Thus, we use regularization to mitigate such risk. Regularization controls the scale of parameters, and in our logistic regression case, we add L_1 penalty to L_{log}, i.e.

$$min_\beta L_{log} + \lambda \sum_{j=1}^{P} |\beta_j|,$$

regression with L_1 penalty is called Lasso regression. In the Lasso regression case, regularization not only controls the scale of parameters, but also selects a set of most significant features i.e. some β_j will be given 0 value. The tuning parameter (hyperparameter) λ controls the influence of the penalty term on the minimization problem, and it is given exogenously, usually chosen via cross-validation.

3.4 Tree-based Methods

A typical decision tree is intuitive, it partitions the feature space into a set of regions and then fits a simple model (like a constant) in each one (Hastie et al. (2009)). Figure 1 is an example of a decision tree. By doing this, a decision tree can reflect the non-linearity and complex interactions of our data. Those regions are called terminal nodes, and each terminal node can make a regional prediction, for instance, the majority class or the mean of the response variable for that terminal node. The split for each pair of nodes is made according to some form of loss functions or some criterions, such as entropy or information gain in a classification context. There could be some stopping limits, otherwise, a tree can have many terminal nodes as observations. Meanwhile, to avoid over-fitting, a technique called tree pruning is introduced. Tree pruning first needs a fully grown tree T_0, then prune it to control the number of terminal nodes.

However, the above decision tree algorithm is still non-robust and has high variance. Thus, we introduce techniques of bagging and boosting. Bagging uses totally B bootstrapped training samples and fits a tree for each sample, then average the predictions of these trees to make a prediction. Different from bagging, boosting trees do not need bootstrapped samples. It uses a weak learner,

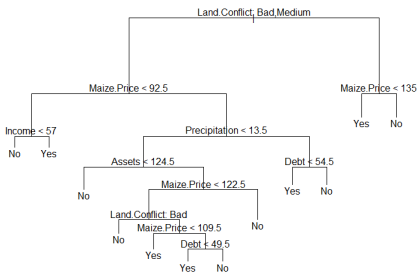


Figure 1: Example of a single decision tree

for instance, a tree (stump) only has one split and two terminal nodes. With such weak learners, for each fit, we only update our $\hat{f}(x)$ at a little learning rate λ, for example, $\lambda = 0.01$. Then we fit B times, and obtain the boosted model as $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$. In other words, for each round of fitting, we modify our data, in other words, we fit the new "residual" each round through boosting. Nevertheless, in the context of classification, instead of "residual" in regression context, the $log(odds) = log(\frac{p}{1-p})$ is usually used in each round of fitting.

3.4.1 Random Forest

The random forests algorithm is a bagging tree-based method. Although a single decision tree is intuitive and can include high dimensional interactions and non-linearities of features, it could have a high variance for testing data. In random forests, with the bagging of trees, we can reduce the variance of function, and grow decorrelated trees.

To be specific, we draw a boostrap sample of size N and use m out of p features (predictors) to grow a tree T_b, repeat B times, and we have a set of trees $\{T_b\}_1^B$. Then we average the prediction result of $\{T_b\}_1^B$ to make the final prediction, we can take the mean of single predictions in the regression case and make a majority vote in the classification case.

3.4.2 XGBoost

eXtreme Gradient Boosting (XGBoost) is proposed by Chen and Guestrin (2016), it is a regularized gradient boosting method. XGBoost substitutes loss function $L(y_i, \hat{y}_i)$ by its second-order Taylor expansion $L(y_i, \hat{y}_i^b) = L(y_i, \hat{y}_i^{b-1} + f^b(x_i)) \approx L(y_i, \hat{y}_i^{b-1}) + L'(y_i, \hat{y}_i^{b-1}) f^b(x_i) + \frac{1}{2} L''(y_i, \hat{y}_i^{b-1}) f^b(x_i)^2$, where $f^b(x_i)$ is the predicted value of x_i for bth round fitting, and $\hat{y}_i^{b-1}$ is the fitted value for the former total $(b-1)$ rounds. Thus, $\hat{y}_i^{b-1}$ is constant in the bth round. Rewrite $L'(y_i, \hat{y}_i^{b-1})$ as g_i and $L''(y_i, \hat{y}_i^{b-1})$ as h_i. Meanwhile, for XGBoost decision trees, we need a regularizer for pruning and controlling the scale of predictions by $\Omega(f_b) = \gamma|T| + \frac{1}{2}\lambda \sum_{j=1}^{|T|} f_j^b(x)^2$. Then, we can rewrite the optimization problem in round b as

$$\min_{(f_1^b(x), f_2^b(x), \ldots, f_T^b(x))} \sum_{j=1}^{|T|} (f_j^b(x) G_j + \frac{1}{2} f_j^b(x)^2 (\lambda + H_j)) + \gamma|T|,$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. I_j is the set of all observations that belong to the jth terminal node, and there are T terminal nodes. In addition, remember that $f_j^b(x)$ is the predicted value of terminal node j which is constant for all $x \in I_j$.

For a given structure of a tree, we compute the predicted value for terminal node j:

$$f_j^b(x) = -\frac{G_j}{H_j + \lambda}$$

In addition, the splits for a single tree are made according to the relationships of $\frac{G_j^2}{H_j + \lambda}$ and γ. How the splits are made will determine the structure of a tree.

4 Data of Uganda

Uganda National Household Survey (UNHS) is undertaken by the Uganda Bureau of Statistics. Our food security indicator (response variable) and some of predictors are provided via UNHS. The UNHS of 2019/20 and 2016/17 data are used. In the 2016/17 data, there are 15,912 households were interviewed, and in the 2019/20 data, there were 15,659 households. The 2016/17 data covers from July 2016 to June 2017, consecutively 12 months. Besides, the 2019/20 data covers from September 2019 to February 2020 and July 2020 to November 2020, a total of 11 months. After the data-cleaning process, some households will be dropped. Meanwhile, other kinds of open-sourced data will also be used. In addition, this paper focuses on machine learning and prediction for specific data, the population weight in the survey data will not be used, because in a descriptive study (e.g. predict the national scale of food insecure rate in Uganda), imputation and adjustments could be made.

4.1 Food Security Indicator

Considering the data availability, I choose the Food Consumption Score (FCS) as the food security indicator in Uganda. The FCS is a composite score based on dietary diversity, food frequency, and relative nutritional importance of different food groups. FCS was first used in Southern Africa in 1996 and is frequently used in food insecurity prediction papers (e.g. Zhou et al. (2022),Lentz et al. (2019) and Martini et al. (2022)). FCS is calculated as $FCS = \Sigma_{i=1}^{9} x_i w_i$, where the x_i is the number of days in the past 7 days that a household consumed food item i, and w_i is the nutrition weight for i, i and w_i are shown in table 1. Thus, $FCS \in [0, 112]$.

Food groups(i)	Weights (w_i)	Food groups(i)	Weights (w_i)
Main Staples	2	Meat and fish	4
Pulses	3	Dairy	4
Vegetables	1	Oil and Fat	0.5
Fruits	1	Sugar	0.5
Condiments	0		

Table 1: FCS Components (WFP (2008))

FCS data in this paper is calculated with the UNHS data, rather than a standard FCS questionnaire. After calculating FCS, households are categorized into 3 classes, for those $FCS \leq 21$ are in the "Poor" diet group, for those $FCS \in (21, 35]$ are in the "Borderline" group, and for those $FCS > 35$ are in the "Acceptable" group. The threshold of 21 and 35 are usually used to indicate food consumption and security status. The value 21 comes from an expected daily consumption of staples and vegetables for the last 7 days, and 35 comes from an expected daily consumption of staples and vegetables complemented by a frequent (4-day/week) consumption of oil and pulses. Meanwhile, WFP (2008) suggested that the use of FCS and the threshold value need to be validated.

Figure 1 shows the cumulative mean food consumption frequencies for Uganda 2019/20 UNSH, the green dashed line is the threshold of 21, the blue dashed line is 28, and the red dashed line is 35. 28 is an alternative threshold for "Poor" diet in WFP (2008), because in some cases, for low FCS households, the sugar and oil consumption may be daily, and combined with daily staple food consumption, FCS already reaches 21, but this 21 does not reflect the borderline diet. Therefore, 7 is added to 21 and 35. From Figure 1, we can find that neither the consumption of sugar nor fat/oil is not reaching 7 for low FCS groups, thus in this paper, we use 21 and 35 as the thresholds. The case for 2016/17 UNHS can be found in the appendix.

In addition, we need to evaluate if the created FCS is a good proxy for food security (WFP (2008)). The validation of created FCS score is shown in Table 2, and we can conclude that our created FCS can be used as a good indicator of food security. Note that for Subjective Income Stability, 1 means very unstable, 2 means somewhat stable, and 3 means stable.

In summary, the 2019/20 FCS of 13302 households are calculated, the FCS distribution of UNHS is shown in Figure 2, and the distribution of 2016/17 FCS data can be found in the appendix. Figure 2 shows that most households are in the "Acceptable" group. Notably, at the bar of the "Poor" threshold of 21 (right-hand side of the green dashed line), which is $FCS \in [21, 21.9]$, this

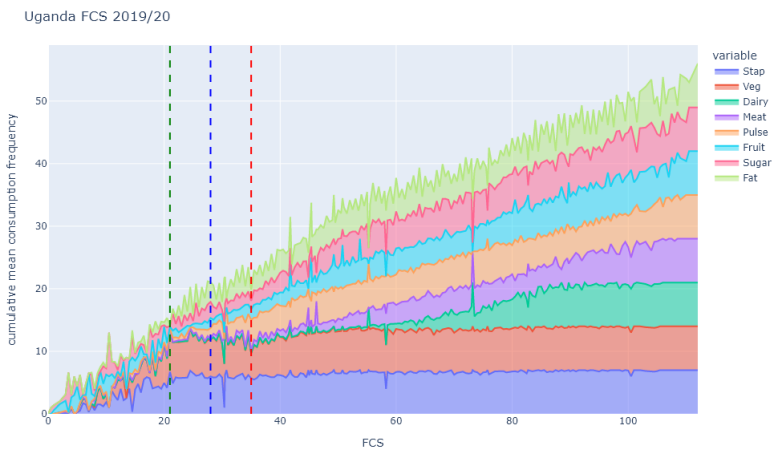


Figure 2: cumulative mean consumption frequency 2019/20

Variable	Corr. Coef.	p-value
Subjective Income Stability	0.231 (Pearson)	0.000
Subjective Income Stability	0.235 (Spearman)	0.000
Total Expenditure on Food	0.605 (Pearson)	0.000
Proportion of Food (value) Received in Kind/Free	-0.155 (Pearson)	0.000

Table 2: Correlation Coefficients with FCS

high bar contains 230 observations, and 214 of them are equal to 21 (in the "Poor" group). 4.16% of the households are in the "Poor" food security situation, and 17.72% of households are in the "Borderline" group. If we exclude those who have 0 FCS, then 3.99% households are in the "Poor" group, more data cleaning details are going to be explained later.

4.2 Predictors

4.2.1 Open Source Data

Nightlight data could be a proxy of economic variables (e.g. Weidmann and Schutte (2017), Yeh et al. (2020)), especially in those developing countries, whose detailed economic data are not available. Because nightlight data is open-sourced, researchers can utilize it at a low cost. Meanwhile, it provides more information on local development in developing countries. This paper uses the nightlight data provided by the Earth Observation Group[1] (Elvidge et al. (2013)), and in order to be consistent with UNHS, the data of Uganda's subnational administrative boundaries (map of Uganda) provided by UN The Humanitarian Data Exchange[2] project is used to extract the mean

[1] https://eogdata.mines.edu/products/vnl/
[2] https://data.humdata.org/

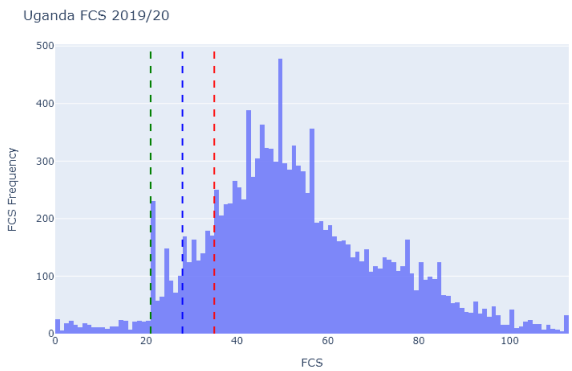


Figure 3: Distribution of FCS in Uganda 2019/20

nighttime illumination for a given region.

The nightlight data is extracted on the scale of District and County. Although the map data is provided by the Uganda Bureau of Statistics, a few names of administrative counties in both datasets do not match, and for those not matched, an imputation by the data of the corresponding District happens.

The climate variables for this paper are monthly sum precipitation (mm) and monthly 2 meters range above ground temperature (celsius degree). *These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program.*[3] They are extracted with the map data as the nightlight section on the district scale because there are 169 cells of climate data, but the number of counties is 194 in UNHS, thus, keep the temperature and precipitation variation in the district scale should be enough. In addition, another source of rainfall data is also used, it is provided by VAM Food Security Analysis[4], and for those districts without NDVI data, an imputation of Uganda national scale data happens.

The conflict data are obtained at *Armed Conflict Location & Event Data Project (ACLED),*[5] more could be found at Raleigh et al. (2010). There are totally six kinds of conflicts recorded in Uganda from 2016 to 2020: Strategic developments, Battles, Riots, Explosions/Remote violence, Violence against civilians, and Protests. Each type of conflict event is counted for a given district and date. Meanwhile, different from Martini et al. (2022) which used the time difference fatalities, I generated the total fatalities for each month for all kinds of conflicts, to reflect the fierceness of the conflict. For all seven generated variables, the missing values are replaced by the mode of that variable values for the covered two years.

NDVI (Normalized Difference Vegetation Index) is used as a predictor for food security in the ML context as well (e.g. Martini et al. (2022) and van der Heijden et al. (2018)). The NDVI used in this paper is downloaded from VAM Food Security Analysis[6]. For those districts without NDVI data, an imputation of Uganda national scale data happens. Lagged data are added.

4.2.2 UNHS data

In UNHS data, wealth indicators, income, and living standard indicators are used as predictors. A detailed variable list could be found in the appendix. Some categorical variables are needed to be transformed into different forms, such as one-hot encoding (dummy variable encoding) or frequency encoding (replacing categories by their frequency value).

5 ML Results

5.1 Introduction and Design

5.2 Basic Model Design (FCS = 35 as Cut-Off)

There are 11 months in UNHS 2019/20. What we want is a nowcasting ML model, i.e. given the previous data, we want to predict the food insecurity situation at present. Because the survey of FCS is costly and time-consuming, this kind of nowcasting model can provide a low cost and real time prediction. To be specific, we use the first 10 months of data as a training dataset, which covers the pre- and during-Covid data. Meanwhile, the last month's data is used as a testing dataset. Facing the shock of Covid, we want to find an ML model which is more robust.

In our food insecurity context, we want to keep balance among the *recall* (ratio of true food insecure households are detected), *precision* (ratio of true food insecure households in the predicted food insecure class), and the *accuracy* (overall correct ratio), to better trade-off the cost and effectiveness of food insecurity detecting and aiding. Therefore, I choose the ROC (AUC) as a performance metric. For a higher AUC value, it implies there exists a threshold probability, which satisfies a high True Positive Rate (recall) and a low False Positive Rate. We can find that when using all features, tree-based methods outperform others.

[3] https://power.larc.nasa.gov/data-access-viewer/
[4] https://dataviz.vam.wfp.org/seasonal_explorer/rainfall_vegetation/visualizations
[5] www.acleddata.com
[6] https://dataviz.vam.wfp.org/seasonal_explorer/rainfall_vegetation/visualizations

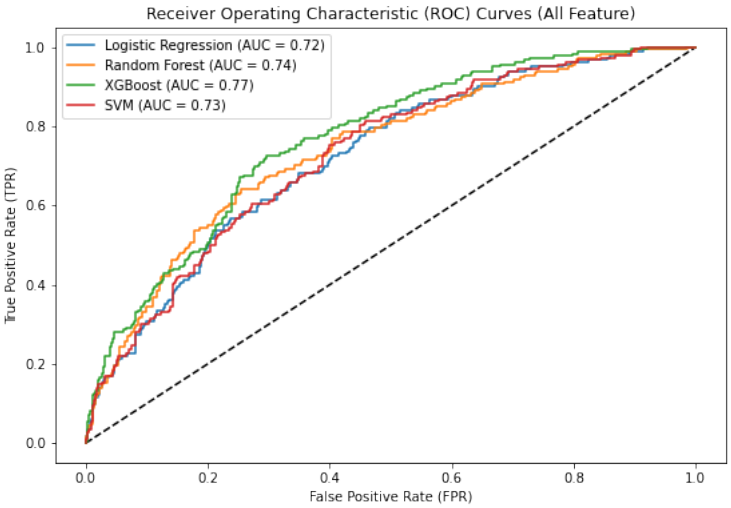


Figure 4: ROC when using all features

ML Models	accuracy	precision	recall	f1	AUC
Logistic Regression	0.774	0.625	0.165	0.261	0.724
Random Forest	0.726	0.451	0.604	0.516	0.740
XGBoost	0.788	0.638	0.280	0.389	0.767
SVM	0.676	0.391	0.610	0.476	0.728
ML Models (70% recall)	accuracy	precision	recall	f1	AUC
Logistic Regression	0.627	0.360	0.700	0.475	0.724
Random Forest	0.668	0.395	0.700	0.506	0.740
XGBoost	0.714	0.443	0.700	0.542	0.767
SVM	0.633	0.365	0.700	0.479	0.728
ML Models (90% recall)	accuracy	precision	recall	f1	AUC
Logistic Regression	0.478	0.304	0.900	0.455	0.724
Random Forest	0.485	0.306	0.900	0.457	0.740
XGBoost	0.537	0.331	0.900	0.484	0.767
SVM	0.497	0.312	0.900	0.464	0.728

Table 3: Performances all features case(Proba. = 0.5 as cut-off except SVM)

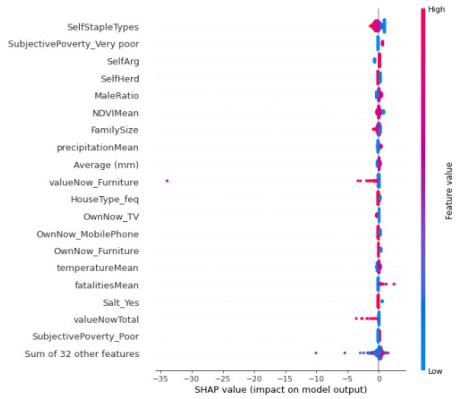
In addition, table ( ) shows how AUC influences the trade-off among accuracy, precision and recall. When we set a high recall (70% and 90%), the XGBoost has the highest accuracy, precision and f1 score.

Using the SHAP value ranking top 25 in the logistic regression and random forest, and the feature importance ranking top 25 of the random forest, we find there are 15 variables that appeared in all ranking lists. This implies that they are influential in the context of food insecurity detection. The figures of SHAP and Feature Importance show us, that how many types of staple food a household grows (*SelfStapleTypes*) is a critical predictor, as well as subjective poverty. Meanwhile, wealth and income variables, demographic variables and some open-source macro variables (e.g. NDVI) also contribute to the prediction.

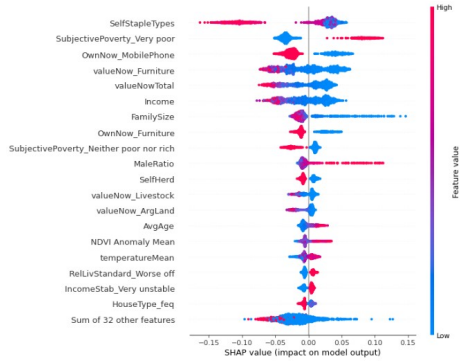
We will use these selected features in the next section.

5.3 Model with Feature selection (FCS = 35 as Cut-Off)

After the feature selection, we use the 15 selected features to build our models. In this section, we have reduced the dimensions of data from more than 40 to 15, then we can use the resampling techniques to overcome the data imbalance. For a high-dimensional dataset, the oversampling techniques cannot capture the distribution of our data very well, because of the sparsity. The following table shows the performance comparison of different scenarios.



(a) SHAP of Logistic Regression



(b) SHAP of Random Forest

Figure 5: SHAP value

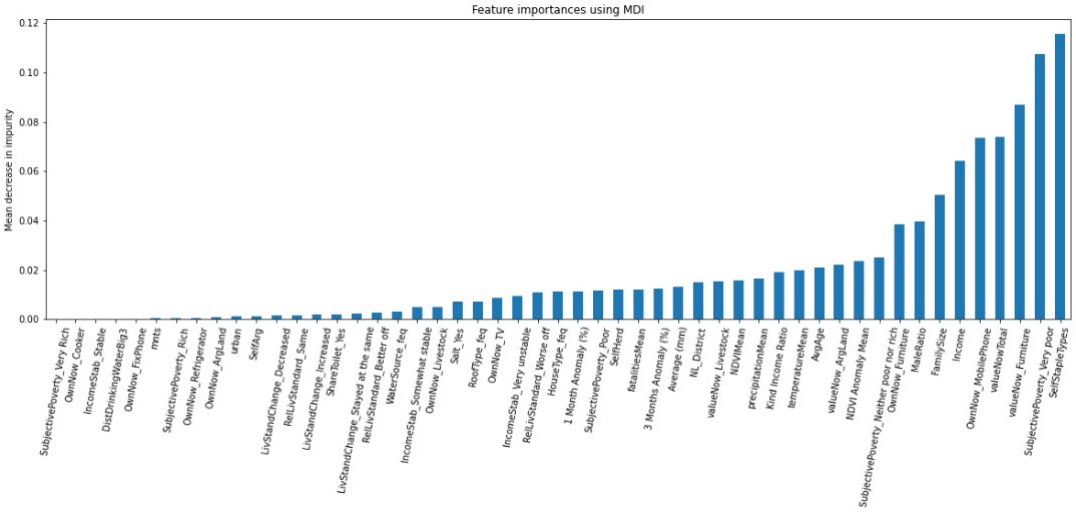


Figure 6: Feature Importance for Random Forest

Models	accuracy	precision	recall	f1	AUC
Logistic	0.773	0.628	0.148	0.24	0.709
Logistic SMOTE	0.668	0.381	0.599	0.466	0.71
Logistic ADASYN	0.651	0.378	0.692	0.489	0.705
Random Forest	0.718	0.435	0.555	0.488	0.721
Random Forest SMOTE	0.746	0.469	0.379	0.419	0.712
Random Forest ADASYN	0.749	0.476	0.385	0.426	0.71
XGBoost	0.765	0.547	0.159	0.247	0.722
XGBoost SMOTE	0.761	0.508	0.352	0.416	0.733
XGBoost ADASYN	0.754	0.489	0.374	0.424	0.73
SVM	0.672	0.386	0.604	0.471	0.716
SVM SMOTE	0.758	0.5	0.022	0.042	0.612

Table 4: Performance With 15 Features

5.4 Basic Model Design (FCS = 21 as Cut-Off)

Models	accuracy	precision	recall	f1	AUC
Logistic	0.95	0.667	0.051	0.095	0.786
Random Forest	0.855	0.188	0.538	0.278	0.813
XGBoost	0.956	0.875	0.179	0.298	0.833
SVM	0.746	0.138	0.744	0.233	0.816
Models (70% recall)	accuracy	precision	recall	f1	AUC
Logistic	0.7	0.112	0.7	0.193	0.786
Random Forest	0.724	0.121	0.7	0.206	0.813
XGBoost	0.753	0.134	0.7	0.225	0.833
SVM	0.795	0.16	0.7	0.26	0.816
Models (90% recall)	accuracy	precision	recall	f1	AUC
Logistic	0.448	0.078	0.9	0.144	0.786
Random Forest	0.575	0.1	0.9	0.179	0.813
XGBoost	0.599	0.107	0.9	0.193	0.833
SVM	0.637	0.115	0.9	0.204	0.816

Table 5: Model Performance

Appendix

Figures

Figure 7: cumulative mean consumption frequency 2016/17

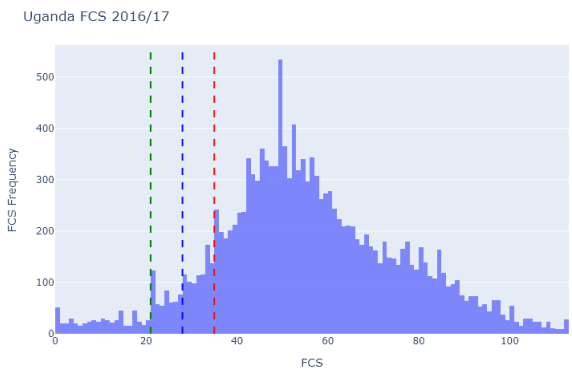


Figure 8: Distribution of FCS in Uganda 2016/17 (15144 obs.)

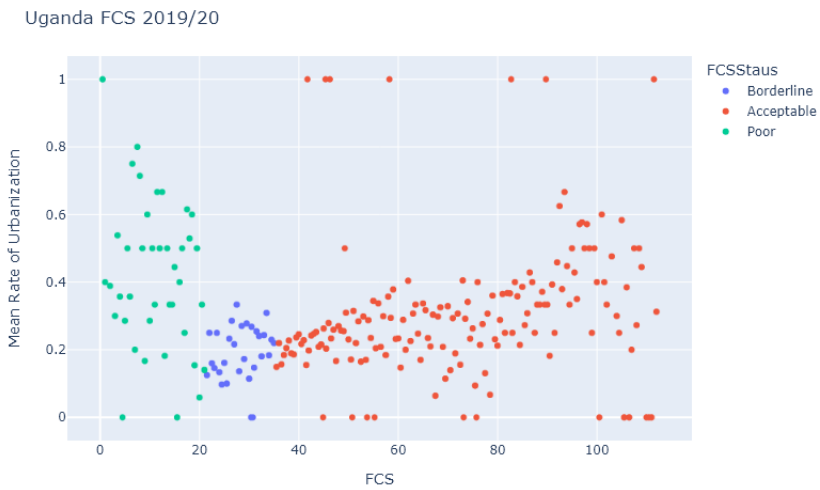


Figure 9: FCS vs. Urbanization in Uganda 2019/20

Tables

Variable Name	Type
Type of dwelling	category
Source of drinking water	category
Distance of drinking water	category
Sharing toilet	binary
Total income	continuous
Share of given in total income	continuous
Having salt	binary
Relative living standard	category
Income stability	category
Change of living standard	category
Subjective poverty	category

Table 6: UNHS Predictors

References

Andree, Bo Pieter Johannes, Andres Chamorro, Aart Kraay, Phoebe Spencer and Dieter Wang. 2020. "Predicting food crises.".

Brinkman, Henk-Jan and Cullen S Hendrix. 2011. "Food Insecurity and Violent Conflict: Causes." *Consequences, and Addressing the Challenges, World Food Prgramme* .

Browne, Chris, David S Matteson, Linden McBride, Leiqiu Hu, Yanyan Liu, Ying Sun, Jiaming Wen and Christopher B Barrett. 2021. "Multivariate random forest prediction of poverty and malnutrition prevalence." *PloS one* 16(9):e0255519.

Chen, Tianqi and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. pp. 785–794.

Elvidge, Christopher D, Kimberly E Baugh, Mikhail Zhizhin and Feng-Chi Hsu. 2013. "Why VIIRS data are superior to DMSP for mapping nighttime lights." *Proceedings of the Asia-Pacific Advanced Network* 35(0):62.

Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk and Francisco Herrera. 2018. *Learning from imbalanced data sets*. Vol. 10 Springer.

Guerrant, Richard L, Reinaldo B Oriá, Sean R Moore, Mônica OB Oriá and Aldo AM Lima. 2008. "Malnutrition as an enteric infectious disease with long-term effects on child development." *Nutrition reviews* 66(9):487–505.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2 Springer.

Hendrix, Cullen and Henk-Jan Brinkman. 2013. "Food insecurity and conflict dynamics: Causal linkages and complex feedbacks." *Stability: International Journal of Security and Development* 2(2).

IMF. 2022. "World Economic Outlook: War Sets Back the Global Recovery." *IMF* .

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112 Springer.

Kang, Yunhee, Víctor M Aguayo, Rebecca K Campbell and Keith P West Jr. 2018. "Association between stunting and early childhood development among children aged 36–59 months in South Asia." *Maternal & Child Nutrition* 14:e12684.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan and Ziad Obermeyer. 2015. "Prediction policy problems." *American Economic Review* 105(5):491–495.

Laraia, Barbara A. 2013. "Food insecurity and chronic disease." *Advances in Nutrition* 4(2):203–212.

Lentz, Erin C, Hope Michelson, Katherine Baylis and Yang Zhou. 2019. "A data-driven approach improves food insecurity crisis prediction." *World Development* 122:399–409.

Martini, Giulia, Alberto Bracci, Lorenzo Riches, Sejal Jaiswal, Matteo Corea, Jonathan Rivers, Arif Husain and Elisa Omodei. 2022. "Machine learning can guide food security efforts when primary data are not available." *Nature Food* 3(9):716–728.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2):87–106.

Okori, Washington and Joseph Obua. 2011. Machine learning classification technique for famine prediction. In *Proceedings of the world congress on engineering*. Vol. 2 pp. 4–9.

Raleigh, Clionadh, rew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47(5):651–660. **URL:** *https://doi.org/10.1177/0022343310378914*

Seligman, Hilary K, Barbara A Laraia and Margot B Kushel. 2010. "Food insecurity is associated with chronic disease among low-income NHANES participants." *The Journal of nutrition* 140(2):304–310.

Smith, Michael D and Maria S Floro. 2020. "Food insecurity, gender, and international migration in low-and middle-income countries." *Food Policy* 91:101837.

van der Heijden, Wesley, Marc van den Homberg, Martijn Marijnis, Marijke de Graaff and Hennie Daniels. 2018. Combining open data and machine learning to predict food security in Ethiopia. In *2018 International Tech4Dev Conference: UNESCO Chair in Technologies for Development: Voices of the Global South*.

Von Luxburg, Ulrike and Bernhard Schölkopf. 2011. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*. Vol. 10 Elsevier pp. 651–706.

Weaver, Lesley Jo, Connor B Fasel et al. 2018. "A systematic review of the literature on the relationships between chronic diseases and food insecurity." *Food and Nutrition Sciences* 9(05):519.

Weidmann, Nils B and Sebastian Schutte. 2017. "Using night light emissions for the prediction of local wealth." *Journal of Peace Research* 54(2):125–140.

WFP, VAM. 2008. "Food consumption analysis: calculation and use of the food consumption score in food security analysis." *WFP: Rome, Italy* .

WFP, WHO, UNICEF et al. 2022. "The state of food security and nutrition in the world 2022.".

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon and Marshall Burke. 2020. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." *Nature communications* 11(1):2583.

Zhou, Yujun, Erin Lentz, Hope Michelson, Chungmann Kim and Kathy Baylis. 2022. "Machine learning for food security: Principles for transparency and usability." *Applied Economic Perspectives and Policy* 44(2):893–910.