**Predicting Food Security with Machine Learning**

Yujun Zhou, Kathy Baylis, Erin Lentz, Hope Michelson

November 13, 2019

**Abstract**

Hunger is on the rise throughout Africa, with famine threatening millions across several countries. Identifying food insecurity crises rapidly and accurately can enable humanitarian responses to mitigate casualties from hunger and save lives. We develop a predictive model based on readily available, spatially granular data on prices, geography, and demographics. Using machine learning techniques, we are able to improve the accuracy of predicting those villages that face a potential threat of hunger. As with any rare event, one challenge with predicting food insecurity is the low rate of severe food insecurity in the baseline data. We use several different approaches to address this imbalance to allow us to capture a higher fraction of these rare events. We apply our procedure to three sub-Saharan African countries: Malawi, Tanzania, and Uganda to predict food security in out-of-sample villages. We correctly identify up to 60 percent of the most food-insecure clusters, when the baseline model using a logistic regression did not detect any of them. We further explore which data splits perform best under the spatial-temporal correlations between observations in the panel dataset to reduce overfitting in actual implementation. Our result shows that a data-driven model with the help of machine learning methods can significantly improve its performance on capturing the food insecure households despite the imbalance in the data. Our paper demonstrates that this approach could be used in a scalable, automatically updated prediction model that could enhance the current famine early warning systems.

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

**Predicting Food Security with Machine Learning**

1. **Introduction**

Hunger 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 providing food aid and organizing humanitarian responses to mitigate the long-run effects of food insecurity. Without timely identification to target the vulnerable population, food aid often fails to arrive the areas in time where the assistance is needed the most (Barrett and Headey 2014). One of the many reasons that prevent building a successful early warning system is that data are scarce, and data collection is costly (Hutchinson 1991). The data gap hinders efforts to effectively target the population in need and calls for the use of data and methods that are cost-effective and accurate. 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 uses a Delphic system that requires detailed on-the-ground data and is updated quarterly for each livelihood zones, making it difficult to identify specific villages that might be at risk of hunger in the near term.

The recent increase in available data on geography, weather, and market price for food staples provides us with the opportunity to predict food security more frequently at a finer geographic level. The use of remotely sensed data to predict socio-economic outcomes is a growing endeavor. Nightlights data (Chen and Nordhaus 2011; Henderson et al. 2012) can serve as a proxy for economic activity, but variations in the nightlight intensity are too low in remote rural or better off urban areas to detect any substantial changes in economic outcomes. Mobile phone data (Blumenstock, Cadamuro, and On (2015); Steele et al., 2017) also hold potential for identifying economic outcomes and are more frequent and less expensive than census surveys. However, geocodes are often limited to cell towers, 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 the lack of labeled data in the imageries makes it challenging to extract structured information from the raw images (Engstrom et al., 2017; Donaldson and Storeygard, 2016). Recent studies using a Convolutional Neural Network (CNN) and transfer learning (Jean et al., 2016; Babenko et al. 2017) make promising progress utilizing the information in 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 for time-varying development indicators. Head et al. (2017) finds that the prediction performance of the Jean et al. method degrades quickly on health and nutrition outcomes to no better than random guessing in some cases. The reliance on nightlight data in 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.

To the best of our knowledge, Lentz et al. (2019) is one of the few papers that combines publicly available spatially and temporally granular data to predict village-level food security status that significantly improves the prediction accuracy without significant cost in data collection and model training. Building on the framework in Lentz et al. (2019), in this paper, we construct a prototype of an early-warning system that is automatically updated, generalizable, scalable, and cost-effective in predicting areas of potential food shortage. Like Lentz et al., we incorporate data from different sources, dimensions, and scales into a single predictive model of food security status. We use machine learning models to predict cluster-level food security status for targeting aid in times of food shortage. Variables in the model include the market price of food staples, weather shocks in growing seasons, and geospatial features around village clusters. We combine data techniques such as oversampling with the machine learning models to improve the prediction of food insecure categories. The models correctly capture 30-60 % of the most food insecurity categories among the three countries for different food security measures. The main contribution of this paper is to improve the prediction of clusters with potential food crisis in an imbalanced data setting. We achieve this by choosing an objective function that balances overall accuracy against the ability to capture the food insecure, along with techniques like data sampling.

Instead of predicting the overall accuracy of food security status in previous studies, this study focuses on correctly detecting the clusters that are food insecure. Similar to a fraud detection problem, severe food insecurity crises are rare but too valuable to miss. The failure of identifying villages with food shortage is more costly than falsely sending food assistance to areas that do not have a shortage of food. Therefore, we focus more on identifying all the insecure clusters to maximize the recall rate in classification, where the recall rate is defined as the percent of truly insecure households correctly captured by the model. We care less about minimizing the number of secure clusters misclassified as unsafe. In technical terms, we put a higher weight on the recall rate than the precision rate for classifying the food security categories. Choosing the right criterion to optimize matters for model selection and parameter tuning. Ultimately, we want the model to correctly detect the minority classes that are food insecure without too many false positive cases.

Along with choosing the optimization criterion, we also explore the effect of different up and down sampling approaches to identify the food-insecure groups. Various sampling techniques create a balanced dataset that forces the classifiers to learn about the characteristics of the minority class. One such method is oversampling the minority class at the risk of model overfitting. SMOTE creates synthetic new data of the minority class by forming convex combinations of neighboring points, as a way to reduce the overfitting in oversampling.

As a natural extension to Lentz et al. (2019), this study expands the study areas to Malawi, Uganda, and Tanzania with more years of data to test the framework with more heterogeneity in geography, environment, and socioeconomic status. For example, Uganda has two growing seasons, and the main food staples are matoke and cassava, while most areas in Malawi and Tanzania have only one growing season and rely on maize as the staple food. This means adjustments to the local climate and agricultural markets, such as having the weather variables during the local growing seasons, grabbing markets data on the staples that take up more significant shares in the household budget in that specific area. The machine learning algorithms and data techniques used for prediction are the same kinds for the three countries, but the hyperparameters are tuned on the training dataset of each country separately. Despite the heterogeneity in socioeconomic and geographic conditions across the three countries, we are able to produce similar prediction results using a unified framework, variables types, and model tuning procedures. This speaks to the validity of the method in expanding to other developing countries with market price data available. At the same time, the model remains flexible and adaptable enough to capture the differences between countries such as climate, crops, and different levels of infrastructure and offer insights on the variables of importance in each country. We compare different methods and protocols of handling the raw data, selecting the optimal model, and setting up parameter tuning procedures in order to come up with a standardized data flow that maximizes our chances of making the model generalizable for potentially other areas in the world.

This paper uses a data-driven framework with machine learning techniques to predict the onset of food crises. Combined with remote sensing data, household surveys, and price data, the models are able to produce 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 reduce the misclassification of food insecure clusters. The framework developed in this paper has important policy implications for accurate target and aid areas of potential food shortage in data-scarce environments.

1. **Data:**

***Food Security measurement***

We predict three measures of food security used by the international humanitarian organizations, including USAID and the World Food Programme (WFP): the reduced coping strategies index (rCSI) and the food consumption score (FCS). The FCS gives nutrient-related weighting to the count of different food categories that a household consumes in the past seven days, to come up with a weighted score of food quality. Higher values of the FCS indicate more diversity of nutrition intake and higher food security. The rCSI reflects the number of coping strategies a household uses to address possible food shortages with higher values of the rCSI indicating lower food security. The rCSI is believed to capture inadequate quantities of food consumed, which is consistent with acute food insecurity.

Governments and international agencies apply cut-offs to categorize food security status rather than use the continuous measure (Vaitla et al., 2017). Consequently, this paper focus on the categorical prediction for the given cutoffs instead of the continuous measures of socio-economic outcomes in previous works (Jean et al., (2016), Steele et al. (2017), and Lentz et al. (2019), among others). The food security category is close to the actual policy scenarios where policymakers are trying to capture all the insecure households in a potential famine year in the currently used IPC system. Our study does not use the household dietary diversity score (HDDS) measure like the other papers do since the variation in the categorical HDDS measure is close to 0, especially after averaging at the cluster level. In other words, a naïve guess that all clusters on average have medium HDDS would beat any models available.

***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. We handcrafted weather-related variables such as the first day of rain, length of dry spells, growing degree days, and heating degree days from the raw precipitation and temperature data during the previous growing seasons specific for each country. We gather the market prices for main food grains such as maize, rice, and groundnuts for the major markets in each country and align the villages to the prices in their nearest markets. To help forecast future food security status, we use prices with one to three months prior to the household survey time. The tree-based models help us choose from a variety of price variables of different products, lag length, and format, with more details in the discussion section. For missing data in the market prices, we construct market thinness measures defined as the number of weeks with price information missing in a given month. Variables regarding wealth status, cellphone ownership, and demographic characteristics are created using answers from the LSMS surveys. Cellular phones are access to financial resources, market information, and remittance flow (Eagle et al. 2010, Blumenstock et al. 2016). Although our results rely on these variables collected from the surveys, we use satellite imagery predicted roof types, consumption aggregates and asset index as proxies for the predictors based on information from the household surveys. Household roof type is used as a crude proxy of poverty that can be accurately captured from satellite imagery.

1. **Method**

This section explains the main approaches 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.

***Categorical vs. continuous measure***

We focus on the categorical prediction for the given cutoffs of each food security measure as it captures the policy scenario where the policymakers need to target communities that are likely to fall under some average food security threshold. We do care about the overall fit of the prediction on the actual food security measures, and we achieve a similar performance of model fit compared to previous studies at around 0.7 R squared. Since this paper focuses on successfully detecting the villages in need of food assistance, we use categorical measures of food security to transform the prediction task into a classification problem. In this way, we can utilize data techniques such as choosing the right result metrics, sampling techniques to improve the chance of detecting insecure villages.

According to the FEWSNET, an FCS from 0 to 28 is considered “Poor”; 28 to 42 is “Borderline”; and above 42 is “Acceptable. An rCSI above 42 is considered “Severely insecure”; 17 to 42 is “Moderately insecure”; 4 to 17 is “Mildly insecure”; and 0 to 4 is “Food Secure.” These cutoffs used in the fieldwork by NGOs and governments allow us to transform the continuous food security measures into categorical using these cutoffs. The goal of the prediction problem then becomes detecting the clusters in the most insecure category (“Poor” or “severely/moderately insecure”) if there is one. In some cases, these insecure clusters comprise less than 1% in the data which makes it harder to detect using conventional methods. In Figure 1, we present the location and distribution of FCS by category in year 2010 for all the three countries to give the readers an idea about the relative portions of food insecure villages and where they were.

An alternative to strictly following the cutoffs to have three or four categories as outcome variables is to have the outcome variable as binary by treating the safest category as 0, and the rest as “potentially or currently insecure.” The binary split has simple interpretations of the results, and we can use measures like the ROC curve to demonstrate the models’ ability to classify the unsafe category accurately. We use the results on the binary splits to be a coarser classification of clusters that may require assistance. One significant advantage in treating data as binary is that we have more data in the minority class to make the models work better.

***Result metrics***

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. Models aiming to maximize the overall accuracy tend to capture characteristics that are rich in the majority of the population and fail to understand the insecure households enough. We want to maximize the recall rate to try and get all the insecure households, without a too low precision rate so that we do not mistakenly categorize all the secure households as insecure. The F-1 score, essentially the weighted harmonic averages of precision and recall, serves as balanced measures of the two. To summarize, the prediction results will be evaluated based mainly on recall of the insecure category but also in consideration of the performance of other measures.

In the binary case, we use the widely used measure called receiver operating characteristic (ROC) curve. The curve plots the true positive rate (TPR) or recall against the false positive rate (FPR) at different values of threshold, where the threshold is defined as the probability cutoff of treating an instance as 1. The false-positive rate measures the cases we have a “false alarm.” The area under the ROC curve (AUC) is the universal statistic for model comparison in the machine learning practices.

***Sampling***

Food secure households and villages make up the majority of the data. When we feed the prediction algorithms with training data made up of these proportions, the models naturally better identify the characteristics of the secure households more than the insecure ones. As a result, the models tend to predict villages in the testing dataset to be secure. To force the models to gain as much information as possible about the food-insecure households, we apply downsampling and oversampling techniques to create a more balanced training dataset in terms of the outcome variable while the testing set remains intact. Specifically, we downsample the clusters in the food secure category and oversample the observations that are food insecure to artificially create a training set where the insecure households make up half of the observations. These methods are broadly used to deal with imbalanced datasets. The main disadvantage with oversampling is that by making exact copies of the minority class, the models tend to overfit. We also use methods like SMOTE (Synthetic Minority Oversampling Technique) to creates synthetic new data of the minority class by forming convex combinations of neighboring points as a way to reduce the overfitting in oversampling. Specifically, we use SMOTE-TOMEK, which combines the oversampling method of SMOTE and then undersamples the synthetic data with Tomek links to avoid overfitting. Another method we use is ADASYN (Adaptive Synthetic), which adds a random noise in the created synthetic new data.

As an alternative to the sampling technique, the cost-sensitive learning approach penalizes misclassifications of the minority class more heavily by having a cost function, which is equal to the inverse of the class proportions (Elkan 2001). This function produces an extra reward for identifying the minority class over the majority class. This approach penalizes the misclassifications of the minority class more heavily than the misclassifications of the majority class, in hopes that this increases the true positive rate. However, defining the “right” cost is not always easy, and they might vary from one case to another. The sampling approach that we use in this paper can be seen as a wrapper-based method that can make any learning algorithm cost-sensitive. The sampling effectively imposes non-uniform misclassification costs on different categories (Elkan, 2001). Weiss, McCarthy, and Zabar (2007) show that oversampling may be preferable for smaller data sets like ours, and cost-sensitive learning is more suitable for datasets with over 10,000 observations. Based on the reasons listed above, our study chooses the sampling approach over cost-sensitive learning.

***Data split***

In the case of spatial and temporal correlation, splitting the data set into training and testing sets is harder than one would think, as the assumption of independence between the training and the testing set is not easily satisfied. The classical assumption of a random split in generating the testing data set may not hold since the points randomly assigned to the testing may have a strong spatial and temporal correlation with some of the points in the training set. Because of these correlations, models with high prediction accuracy on the training set would appear to have higher than actual accuracy in the testing set as well (Meyer et al. 2018). Consequently, the training process would prefer models and parameters that are complex enough to understand every bit of detail in the training set. These models would tend to overfit in a genuinely independent testing set as is demonstrated in simulation results in Robert et al. (2017). Taking these into consideration, this paper chooses the year split as the data split strategy. In other words, we use one year as the training set to predict food security in another year. Survey data collected in another year or wave have weaker ties with the training data because the time difference is long and that not all villages are repeatedly visited in another year. Spatial correlation between points that are close in space with similarity but are separated into training and testing might create upward bias on the out-of-sample performance. Similar argument can be made for a purely random split method, as the independence assumption of the testing set would no longer be valid.

***Cross-validation***

Cross-validation (CV) is a resampling-based technique for the estimation of a model’s predictive performance and is often used in hyper-parameters tuning and model selection. In this paper we use a two-tier nested CV. First, we use any two years of data to predict the other year and average the results on the three test sets to get generalized performance metrics. The reason why we do this is because the number of rare events of food insecurity can vary a lot from one year to another and the averaging provides a relative stable performance to understand generalization error of the underlying model. Also, we apply this stratified CV method to avoid the spatiotemporal correlation issue mentioned above. In the model training phase, we use a K-fold CV grid search for hyper-parameters tuning and model selection. K-fold CV randomly divides the data into K subsets of approximately the same size and then having each subset used successively as the test set.

***Data segmentation***

Data segmentation is another choice to make in defining our training dataset. Due to the heterogeneity of different countries and regions, including more data in the training data, sometimes is adding more noise than information. The concern of fitting a generic model on all the data that we can get our hands on is that the model tend to work well on the country or region that take up the larger portion of the observations and perform poorly on places with less survey data. Especially when combined with the oversampling method, the model fit on the whole sample perform worse in some areas more than the others (with more discussions in the error analysis section). Also, the price variables across countries are not completely the same and as a result, we lose a few price variables on some of the grains when we train the model on all the countries. We compare the results of models trained on the entire dataset of three countries, with models trained by each country separately.

***Classification algorithms***

For structured data like ours (unlike unstructured data like text or pure images), tree-based methods are popular, as they work well with the nonlinearities and a large number of variables with different scales in the data. Decision trees also make no distributional assumptions on the data to reduce the misspecification error. We start with a simple machine learning algorithm known as Decision Trees. The decision tree splits the training data based on several input variables, such as does the village undergoes a long dry spell in the last year? If so, is the cluster rural or urban? The splits created by the questions divide the training data into different subset of data known as the “leaves.” In this classification problem, each leaf is associated with a probability of falling into the secure or insecure category and the probability is trained using the training dataset. The main hyperparameter of the classification trees is the depth of the trees. Deeper trees tend to capture more complexity of the training data but may suffer from overfitting when applied to the testing set. Shallow trees may suffer from losing essential splits in the data and underfit the training set. As a result, tree-based ensemble learning methods like the Random Forest, and Gradient Boosting help improve model performances by averaging and sequentially improving the base trees, which is particularly helpful when there is a large number of potential variables in our model. Random Forest is a kind of bagging algorithm, short for “Bootstrap aggregating.” The idea of the algorithm is to create many deep trees based on randomly selected subsamples of the training data with randomly selected variables determined at each split of the trees (the bootstrap aspect) and then using the weighted average of the results from the trees (the aggregating aspect) to reduce the total variance of the deep trees.

Boosting differs from the bagging method in that trees are grown sequentially instead of in parallel. Modelers usually start with a simple, shallow tree for prediction and then use the error in prediction from the previously grown tree to adjust the weights in the next iteration. Errors are corrected by sequentially adjusting the weights in the existing models until no more improvements can be made. Weiss (2014) suggested that boosting methods tend perform well at classifying minority examples because boosting places more emphasis on misclassified instances in the training set and the errors are more likely to come from the minority classes. Based on the performance across twenty-nine datasets, Tischio and Weiss (2019) find that models like Decision Trees, Adaboost, and Gradient Boosting perform better than other algorithms in the presence of imbalanced data. The gradient boosted tree improves the initial decision tree model by sequentially adjusting the model in the direction of the negative gradient of the loss function defined as the squared distance of predicted and actual values. The Xgboost (Extreme Gradient Boosting) algorithm relies on the same principles as gradient boosting but is more efficient and faster since it adds regularization to the loss function to prevent overfitting.

***Baseline model***

As a comparator, we use a standard logistic regression as the baseline model. We estimate each country separately using all but the last year of data, and then use the model results to predict the last year, without downsampling or oversampling methods. We use the same variables used in machine learning models such as food prices, market thinness, cellphone ownership, floor/roof material, asset index, length of dry spells, average temperature, and the total amount of rain.

1. **Results**

The results section shows many improvements in performance of the machine learning models compared to the baseline model. The results suggest that the machine learning algorithms, combined with data techniques, help capture the rare food insecure instances when the conventional methods cannot.

Figure 2 shows the comparison of the model performance using the machine learning model and the baseline model using logistic regression. The performance is measured by overall accuracy, recall rate, and F-1 score. The results of the baseline model is shown in red; Random Forest, Tree, and xgBoost models are shown in green, blue and purple respectively. On the left panel of Figure 2, across the different measures of food security and country combination, we see the overall classification accuracy to be quite high, demonstrating that our framework, in general, predicts food security well. Using a simple logistic regression, we successfully classify the FCS in Tanzania 90% of the time. In general, the prediction accuracy is around 70% to 90% for FCS measure and around 60% to 70% for rCSI. Machine learning algorithms overperform the baseline model in most cases. In Uganda-FCS, the accuracy is 20% higher than the baseline using the xgBoost model. This makes sense as the machine learning models can either reduced the variance from a complex model or reduce bias sequentially by updating the model from training errors.

However, most of these correctly predicted samples come from the food secure categories. As we move to the middle and right panel of Figure 2, the baseline model fails to detect any of the most food-insecure villages, showing as 0 recall rate and 0 f-1 scores across all the country-measure combinations. In some cases, there were only 1 to 9 villages that are in that category, representing 1% -3% of the sample in the test data. One exception being the Malawi-FCS combination, where we have more than 100 villages (around 20% of the test sample) in this case, as there are 0 villages with “poor” food consumption score, so we treat the “borderline’ category as the insecure category that we want to predict. Even so, the baseline model was not able to detect any of them. Even more surprising is how our machine learning models also fail to predict the insecure villages in the testing set. The random forest model outperforms the other ones but is also only able to capture 7% to 9% of the most vulnerable villages. The main reason for this is the lack of training data with labels in the minority class. In this case, the training data in the previous years also contain very few cases of food-insecure villages. The majority of the sample comprised of villages not suffering from imminent danger of hunger. Models, machine learning, or conventional, trained on these data tend to capture more the characteristics of the food secure villages and unable to identify signs in the data that signify food insecure.

As is discussed in the method section, we apply oversampling and downsampling on the training data to create a more balanced dataset. Figure 3 shows the results of different sampling techniques combined with different machine learning models, compared to the baseline model with no data sampling. Results using simple oversampling of the minority class are shown in squares; SMOTE-Tomek in a cross; ADASYN in circles; and baseline model in red triangles. Like in Figure 2, we have random forest model in green, tree model in blue, and xgBoost in purple. When combined with sampling, the performance difference in machine learning methods and our baseline model diverges, as we improve the prediction accuracy on the insecure households (higher recall rate). The overall accuracy does not improve much or is even lower, compared to the performance in figure. The balanced training data have forced models to take up more structures of the data in the insecure category, and hence more misclassification on the secure households appears.

On the other hand, we see an increase in the recall rate in almost all the country-measure combinations, except for Tanzania-rCSI where we did not identify any of the six insecure clusters. The case of Tanzania-FCS is also extreme, only one insecure cluster in the testing set. The only case that we are able to capture that one and achieve 100% recall rate is the combination of Tree and oversampling. However, it is at the cost of misclassifying several secure ones since the F-1 score is as low as 0.1 in this case. We observe better results in other two countries. In Malawi, the recall rate on FCS has increased to 0.5 and 0.6 with reasonable F-1 score; the combination of oversample and xgBoost reaches a recall of 0.3 on rCSI when the other model cannot detect the insecure clusters well. In Uganda, xgBoost outperforms other methods in two different sampling techniques to successfully detect half of the food insecure villages. These results are not ideal, but at least they are a significant improvement from the baseline recall rate of 0 to 60% without too many misclassified samples.

The combination of machine learning algorithms and sampling is interesting. Across different sampling methods, the simple tree model performs surprisingly well, especially combined with simple oversampling. The idea might be that it is an extreme combination that overfits the training sample, and in the balanced training set with lots of duplicates of insecure clusters, the tree model is more likely to classify an unknown cluster as insecure (but luckily not too much). Aggregated or ensemble models like Random forest and gradient boosting are not universally better than their single counterparts (i.e. classification tree). They are better when we value the overall stability of the results. However, they are not necessarily better when our focus is on successfully detecting a few observations that comprise 1% to 5% of the entire testing sample. Likewise, we can see how the xgBoost-oversampling combination performs better than random forest in several cases in detecting the rare events when the random forest does better on overall accuracy and recall without the sampling. This is because the random forest model brings down the variance by averaging the trees with randomly selected features. Even though the model is more stable in testing out-of-sample in terms of overall accuracy on all the categories, it performs worse on detecting the lower end of the distribution. We can find evidence on these model characterizations in the binary classification case presented in Figure 4 as well. This finding may benefit future modeling strategy on detecting rare but significant events.

Figure 4 presents the results of binary classification of “safe vs not” clusters where one is defined as not in the safest category. Thanks to the binary feature, we can use the universal classification model comparison tool known as the ROC curves. By the way we define the outcome variables, the dataset is balanced. Across the five different country-measure combinations, we can see how the machine learning methods perform much better than the baseline model. In particular, random forest and xgBoost models get close to 0.9 AUC on the FCS cases and 0.7 for the rCSI when the baseline model is only slightly better than random guessing (0.6 AUC). In this case, the simple classification pales to the ensembled counterparts as the result metrics is more on the overall classification accuracy.

On Segmentation: Table 4: Data Segmentation Comparisons. Comparison of one dataset/ by country / urban-rural

1. **Discussion**

***Parameter tuning***

* Discussion on tree depth, max features and how they impact the results
* Discussion on variables that explains the majority of the variations in data, e.g., cellphone ownership

***Feature importance analysis***

***Error analysis***

* Regional error analysis
* Urban vs. rural
* By food security measure

1. **Conclusion**

****

**Fig. 1 Map of FCS in 2010**

****

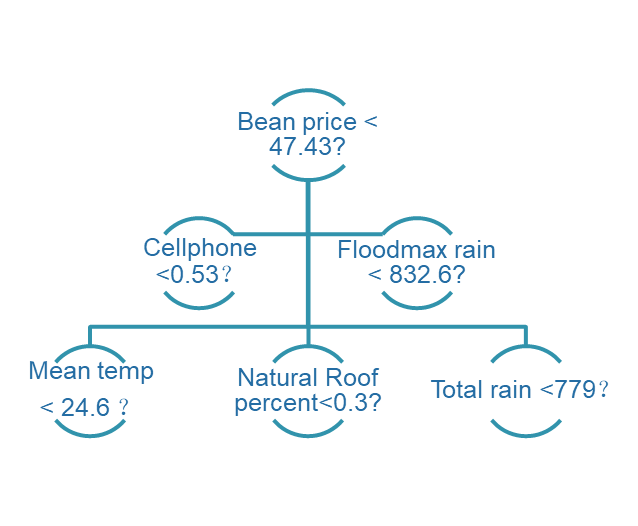
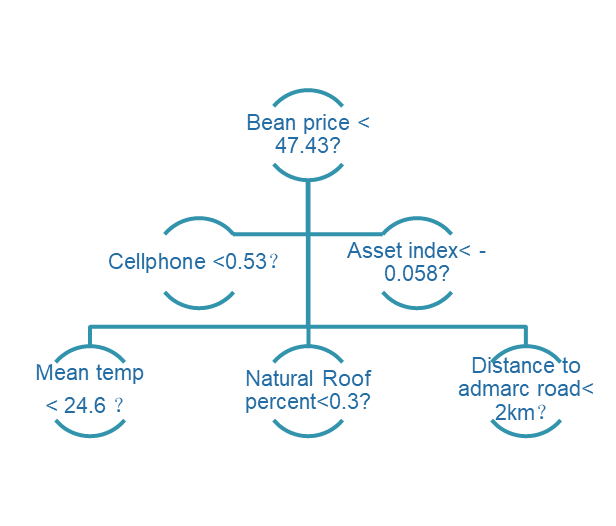
**Fig. 2: ROC curves of Baseline vs. ML algorithms using year split with binary cutoff**

**Table 1: Baseline vs. ML algorithms using year split with FEWSNET cutoff**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Food Security Measure | Overall  Accuracy  (Baseline) | Overall  Accuracy  (ML) | Recall Rate Insecure Category  (Baseline) | Recall Rate Insecure Category  (ML) |
| Malawi | FCS | 0.71 | 0.75-0.76 | 0.26 | 0.18-0.38 |
| rCSI | 0.69 | 0.60-0.63 | 0.36 | 0.54-0.72 |
| Tanzania | FCS | 0.81 | 0.82-0.84 | 0.06 | 0.08-0.29 |
| rCSI | 0.55 | 0.59-0.63 | 0.29 | 0.43-0.54 |
| Uganda | FCS | 0.67 | 0.59-0.71 | 0.36 | 0.33-0.36 |

**Table 2: Baseline vs. ML algorithms using year split with FEWSNET cutoff**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Food Security Measure | Recall Rate Insecure Category  (Baseline) | Recall Rate Insecure Category  Ml + Oversample | Recall Rate Insecure Category  Ml + Smote-Tomek | Recall Rate Insecure Category  Ml +   Adasyn |
| Malawi | FCS | 0.71 | 0.75-0.76 | 0.26 | 0.18-0.38 |
| rCSI | 0.69 | 0.60-0.63 | 0.36 | 0.54-0.72 |
| Tanzania | FCS | 0.81 | 0.82-0.84 | 0.06 | 0.08-0.29 |
| rCSI | 0.55 | 0.59-0.63 | 0.29 | 0.43-0.54 |
| Uganda | FCS | 0.67 | 0.59-0.71 | 0.36 | 0.33-0.36 |

**Appendix:**

**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.

Elkan, Charles. "The foundations of cost-sensitive learning." In *International joint conference on artificial intelligence*, vol. 17, no. 1, pp. 973-978. Lawrence Erlbaum Associates Ltd, 2001.

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.

Meyer, Hanna, Christoph Reudenbach, Tomislav Hengl, Marwan Katurji, and Thomas Nauss. "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation." *Environmental Modelling & Software* 101 (2018): 1-9.

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

Ray M. Tischio, and Gary M. Weiss. Identifying Classification Algorithms Most Suitable for Imbalanced Data, *Proceedings of the 15th International Conference on Data Science* 2019, Las Vegas, NV.

G.M. Weiss. Mining with Rarity: A Unifying Framework. *SIGKDD Explorations*, 6(1): 7-19, 2004.

Weiss, Gary M., Kate McCarthy, and Bibi Zabar. "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?." Dmin 7, no. 35-41 (2007): 24.