Predicting Food Security with Machine Learning

Yujun Zhou, Kathy Baylis, Erin Lentz, Hope Michelson
Agricultural and Consumer Economics
University of Illinois

IPAD Seminar
10/17/2019

Motivation - The Problem

- 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 use the Integrated Food Security Phase Classification System (IPC).The IPC relies on a convergence of evidence approach rather than a formal model (IPC, 2012)
- The IPC has been accused of being too complex, requiring extensive information, and vulnerable to political influence (The Economist, 2017; De Waal, 2018)
- Need to have a data-driven, transparent framework to provide accurate, frequent, spatial granular predictions of food security crises

Motivation - The Efforts

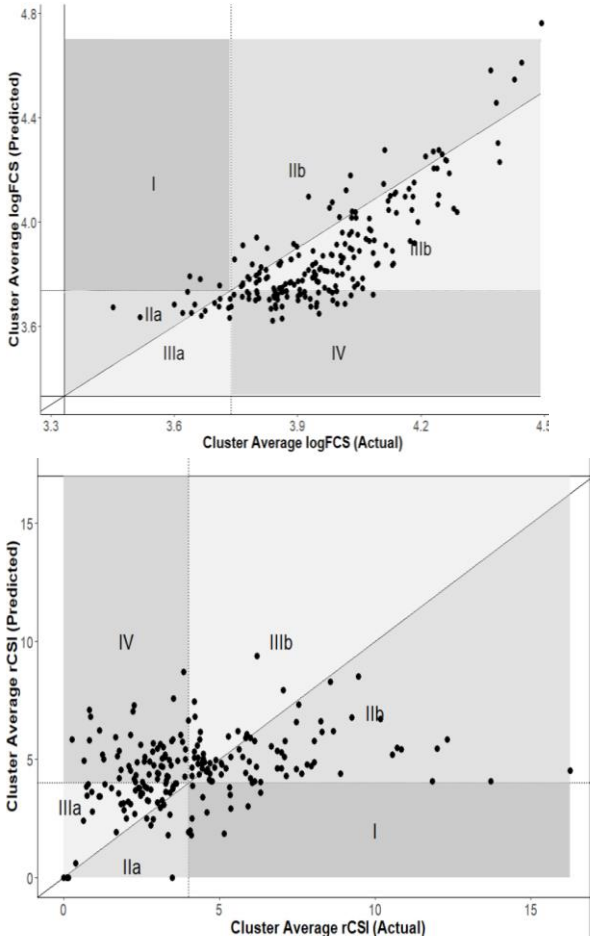
- Night-lights data to predict village-level poverty (asset index) (Chen and Nordhaus 2011; Henderson et al. 2012), noisy and lack of variation in certain areas
- Mobile phone data (Blumenstock et al., 2015; Steele et al., 2017) accurate but hard to scale
- Satellite imagery (Engstrom et al., 2017; Donaldson and Storeygard, 2016) - lack of labeled data to extract structured information
- Combining night-lights and satellite imagery in a CNN model (up to 70% accuracy) (Jean et al., 2016; Babenko et al. 2017)
- Reliance on satellite imagery and Nightlights data (Head et al. 2017)
    - works in some areas better than others (SSA; Nepal and Haiti are problematic)
    - does not capture changes in poverty over time
    - does not do so well with other development metrics: nutrition, health

Motivation - The Opportunity

- Recent increase in available data related to food security, geography, weather, and market price for food.
- These data are often evaluated in isolation and require a systematic framework of combining data from different sources, frequency and spatial scale.

Motivation - The Challenge

- Lentz et al. (2019) incorporate these data into a single predictive model of food security early warning and achieved R squared up to 0.65 and categorical accuracy up to 90%

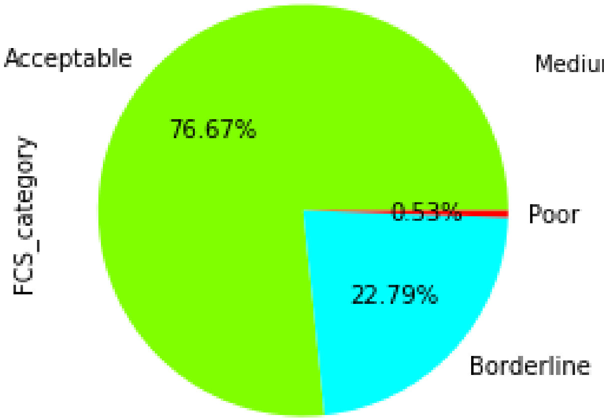


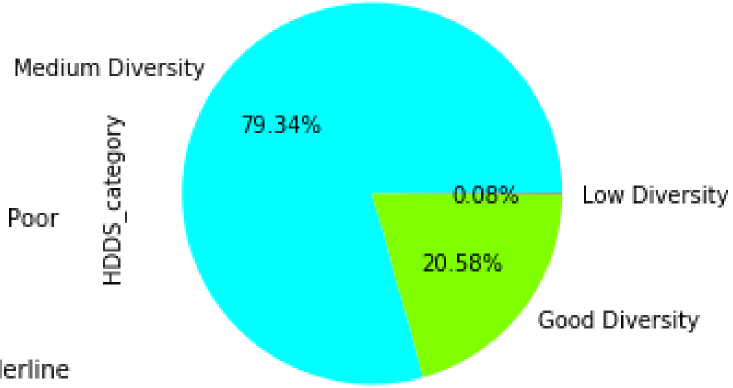
Zhou, Baylis, Lentz, and Michelson

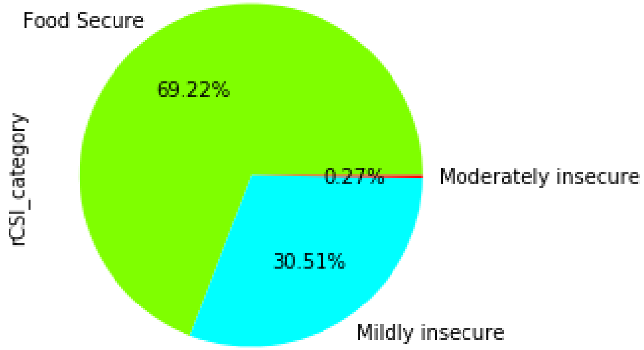
Motivation - The Challenge

- Capturing villages that face a potential threat of hunger, i.e. the most food insecure groups
- As with any rare event, the low rate of severe food insecurity in the baseline data tend to be ignored by the models
- Use machine learning and data techniques to address this imbalance to capture a higher fraction of these rare events.

Detecting Rare but Relevant households







FCS
Food Consumption Score

HDDS
Household Dietary Diversity Scale

rCSI
Reduced Coping Strategy Index

Research Question

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

- Successfully detect the rare events food insecurity, informing the when and where food shortages tend to happen

- A framework that is automatically updated, generalizable, scalable and cost-effective

Summary

- 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 (sampling, cross-validation, data segmentation) to improve prediction performance
- Correctly categorize 63-84 % of food insecurity categories and up to 20-57% of the most food insecure categories

Data

LSMS survey

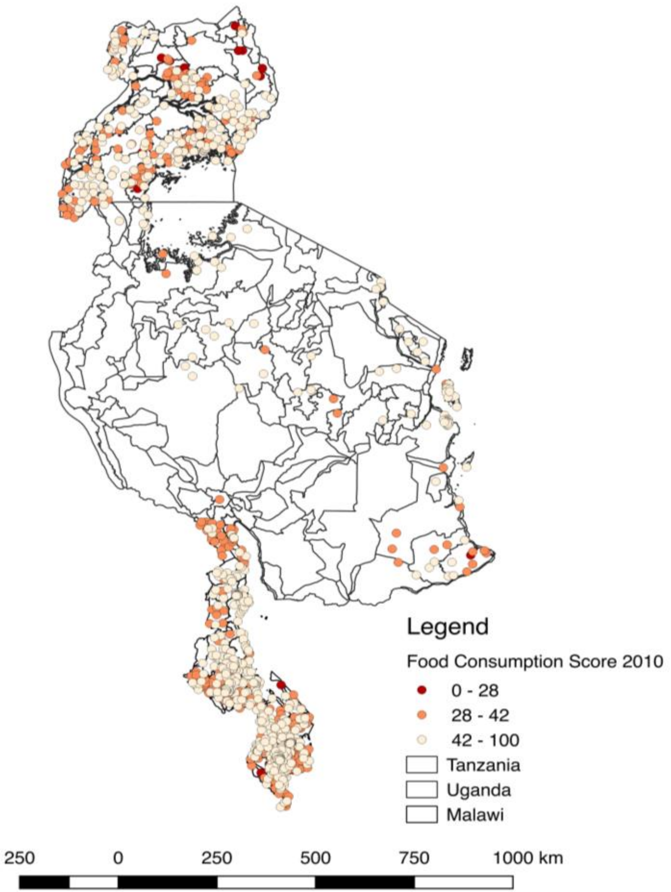
- Ground truth data in Uganda/Tanzania/Malawi
- Three different rounds each with broad spatial coverage
- Household assets, demographics, geolocation of clusters
- Distance to roads and markets reflects access to market and information

Prices

- Major grain prices (maize, rice, nuts, beans etc.) in main agricultural markets prices, collectedly monthly
- Lagged one month before the survey time

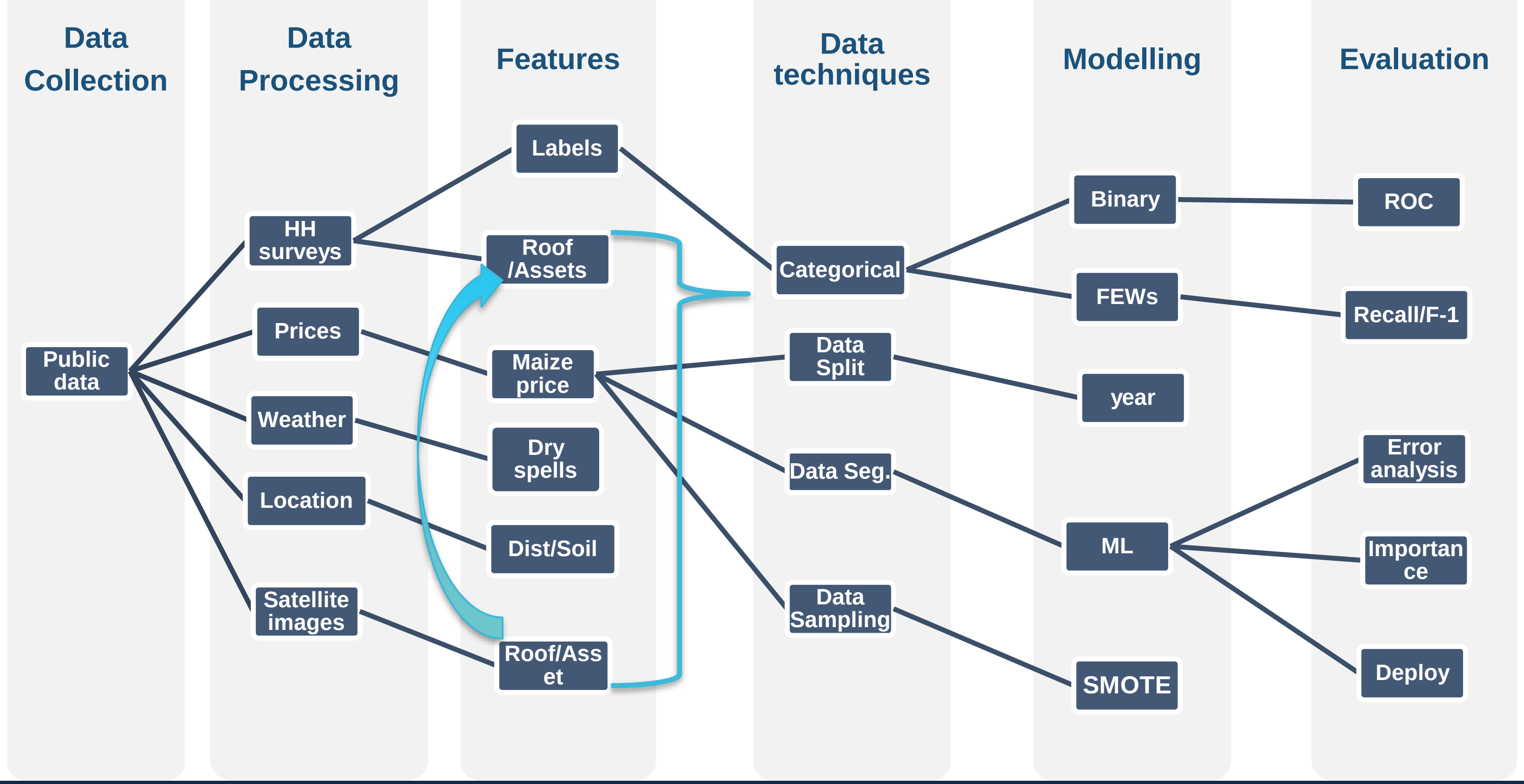
Weather

- Precipitation/temperature from remote sensing are relevant to agricultural production
- From the previous growing season



Zhou, Baylis, Lentz, and Michelson

Framework



Data Collection
Data Processing
Features
Data techniques
Modelling
Evaluation
Public data
HH surveys
Prices
Weather
Location
Satellite images
Labels
Roof /Assets
Maize price
Dry spells
Dist/Soil
Roof/Ass et
Categorical
Data Split
Data Seg.
Data Sampling
Binary
FEWs
year
ML
SMOTE
ROC
Recall/F-1
Error analysis
Importan ce
Deploy

Decisions, decisions

1. Categorical versus continuous prediction

2. If categorical, how do we address rare events?

3. How do we split the data?

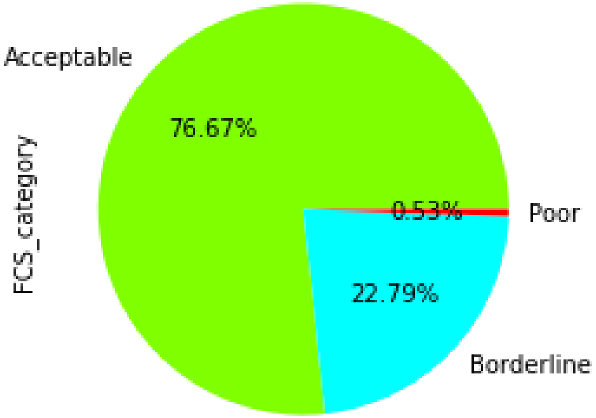
4. What algorithm do we use?

Categorical vs Continuous?

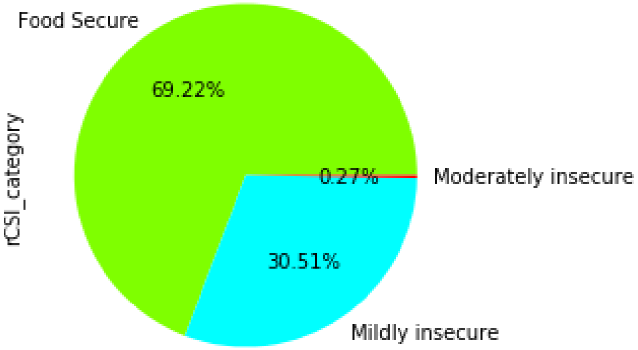
Focus on the categorical prediction for the given cutoffs of each food security measures.

- Recall rate of the insecure villages is more important than the over all 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.
- Apply the down sampling, over sampling, and synthetic data techniques to force the model to learn about the tail of the distribution

Detecting Rare but Relevant households



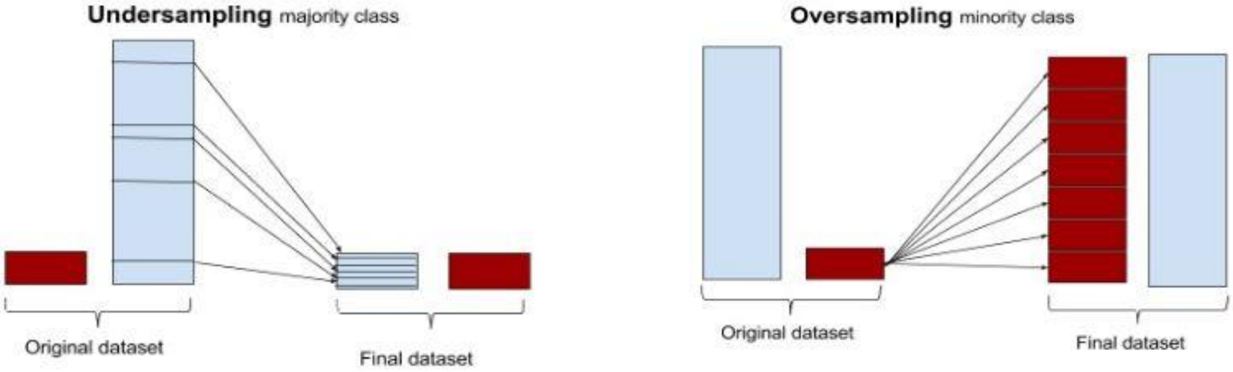
FCS
Food Consumption Score

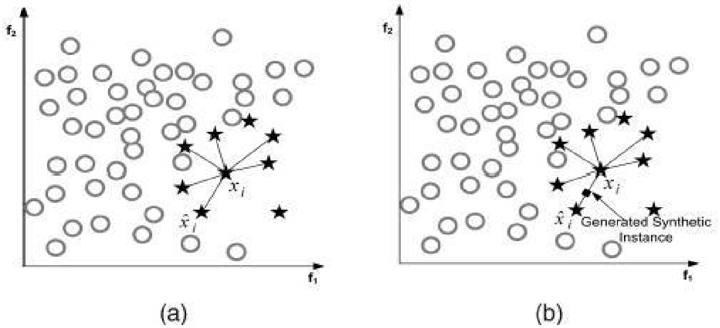


rCSI
Reduced Coping Strategy Index

Zhou, Baylis, Lentz, and Michelson

Methodology: Under sampling, Over sampling and Synthetic data



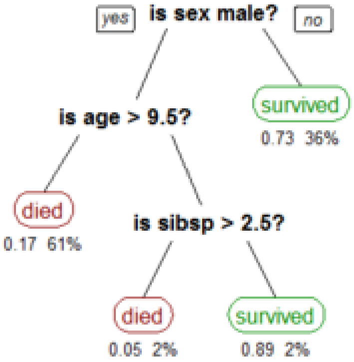


Zhou, Baylis, Lentz, and Michelson

Methodology: Classification Algorithms

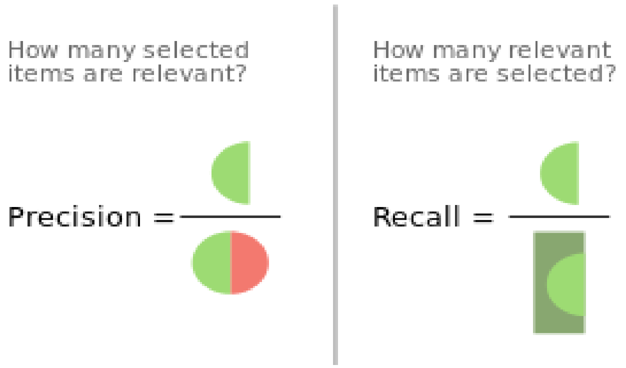
For structured data like ours (unlike text or pure image), tree-based methods are popular (Weiss, 2014; Tischio and Weiss 2019)

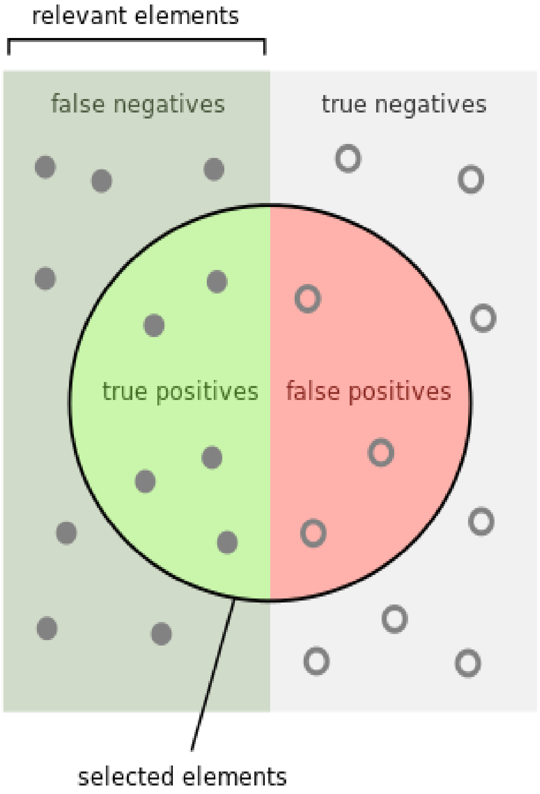
1. Classification Tree (base learner)
2. Random Forest (parallel)
3. Gradient Boosting (sequential)



Methodology: Result Measurements

1. Recall (are we getting all the insecure households ?)
2. Precision (are we mistakenly categorizing secure households as insecure?)
3. F-1 score (balance recall and precision)
4. Overall categorical accuracy





Zhou, Baylis, Lentz, and Michelson

Compare to baseline Model

- Logistic Regression
- Data split: year split (cross-validated)
- Data segmentation: by country
- Down/over sampling: None

Variable groups:

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

Methodology: Cross-validation

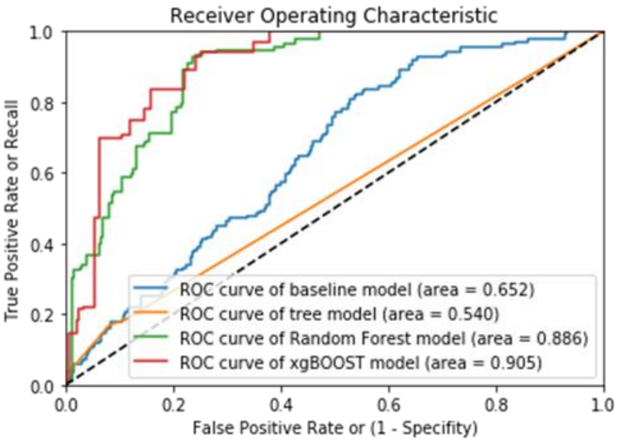
- Rare events of food insecurity tend to vary a lot year by year, i.e. 1 or 2 cases in a good year vs over 50 cases in a bad year

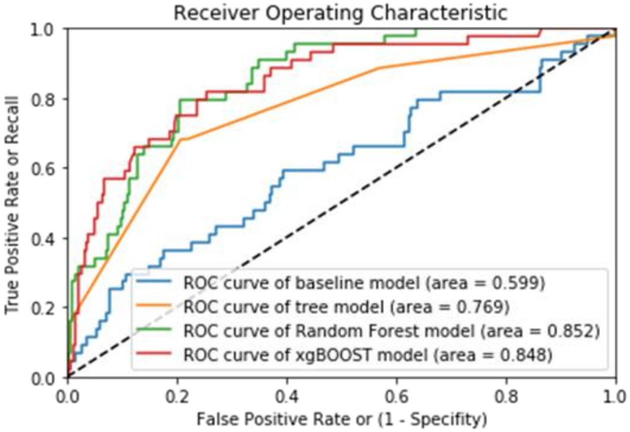
- Use any two years as training data to predict the third year

- Average out the performance after cross-validation to get more stable and trustworthy result

Results for binary cut-off

FCS

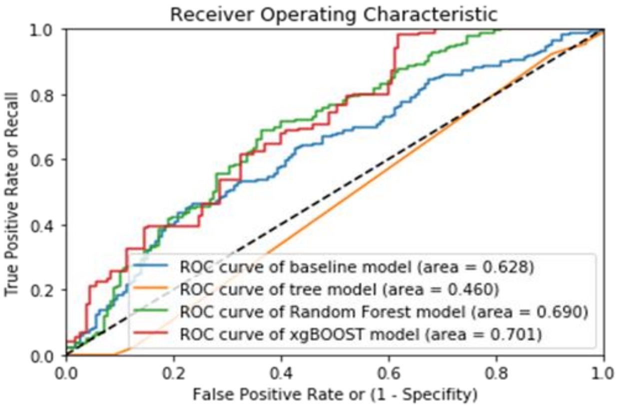


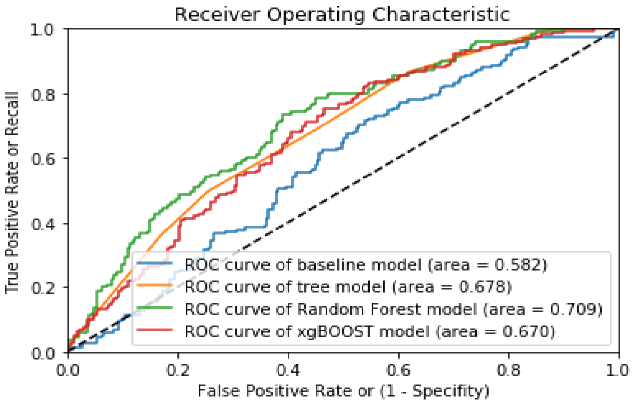


Malawi

Tanzania

rCSI





In table format… Binary Baseline vs ML algorithms, no oversampling (year split)

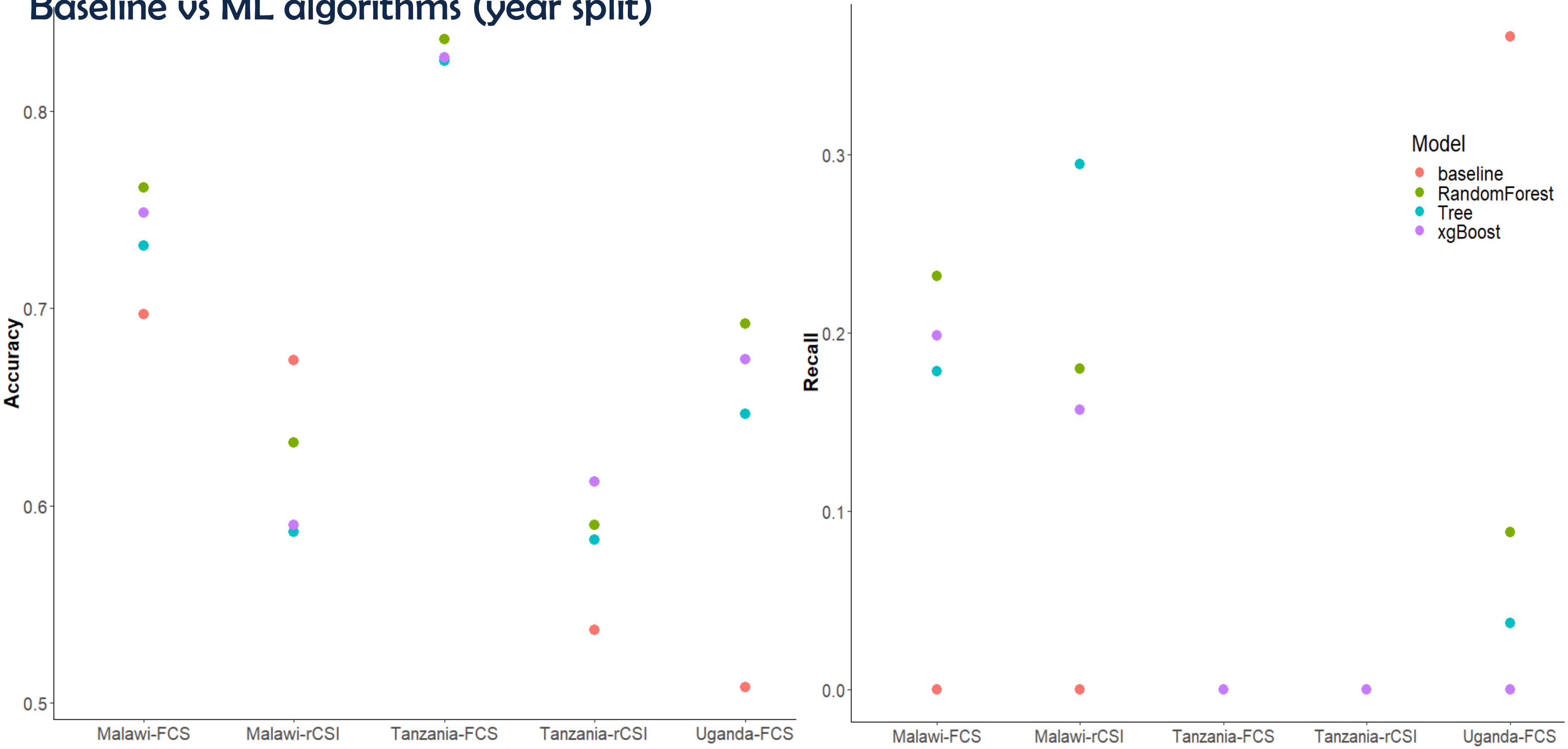
Similar accuracy, higher recall

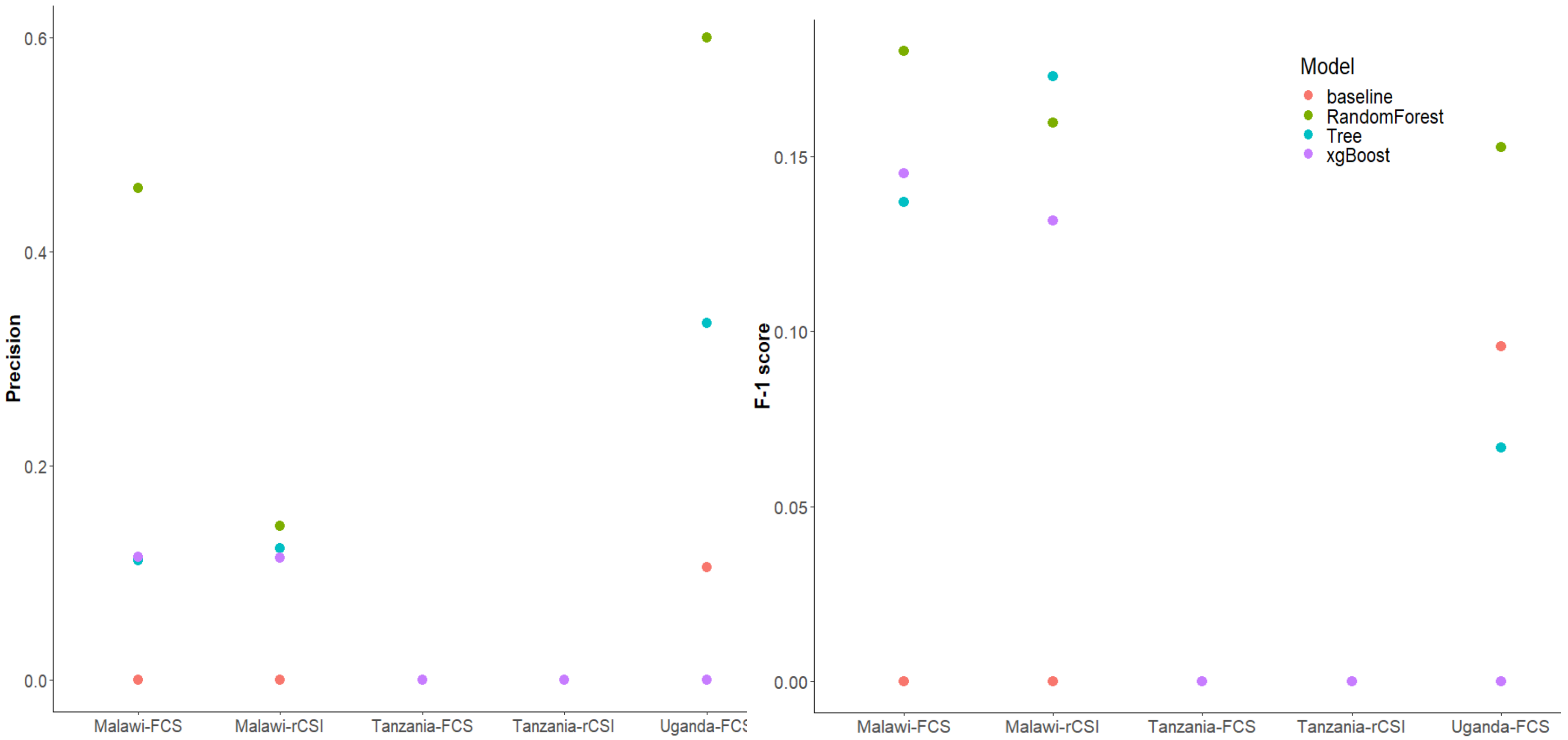
Country	Food Security Measure	Overall Accuracy (baseline)	Overall Accuracy (ML)	Recall Rate Insecure category (baseline)	Recall Rate Insecure category (ML)
Malawi 2010/11, 2013 to predict 2015/16	FCS	0.71	0.75-0.76	0.26	0.18-0.38
	rCSI	0.69	0.60-0.63	0.36	0.54-0.72
Tanzania 2010/11, 2012/13 to predict 2014/15	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 2010,2011 to predict 2012	FCS	0.67	0.59-0.71	0.36	0.33-0.36

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-TOMEK	Recall Rate Insecure category ML + ADASYN
Malawi	FCS	0.27 / 0.00	0.30 / 0.27	0.33 / 0.06	0.53 /0.15
	rCSI	0.36 / 0.00	0.33 / 0.00	0.42 / 0.20	0.46 / 0.20
Tanzania	FCS	0.01 / 0.00	0.08 / 0.00	0.22 / 0.01	0.23 / 0.43
	rCSI	0.32 / 0.00	0.41 / 0.00	0.44 / 0.20	0.47 / 0.06
Uganda	FCS	0.18 / 0.37	0.21 / 0.00	0.26 / 0.57	0.20 / 0.50

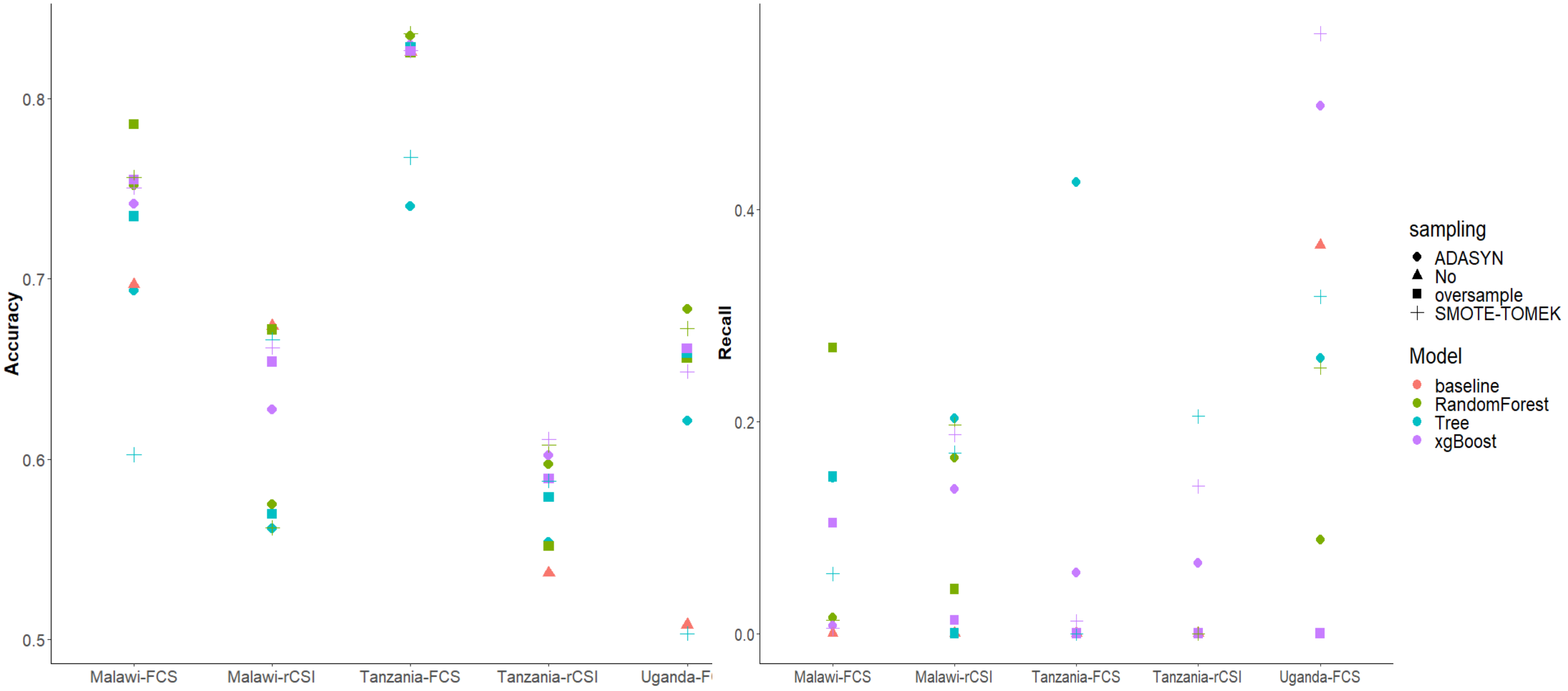
Baseline vs ML algorithms (year split)

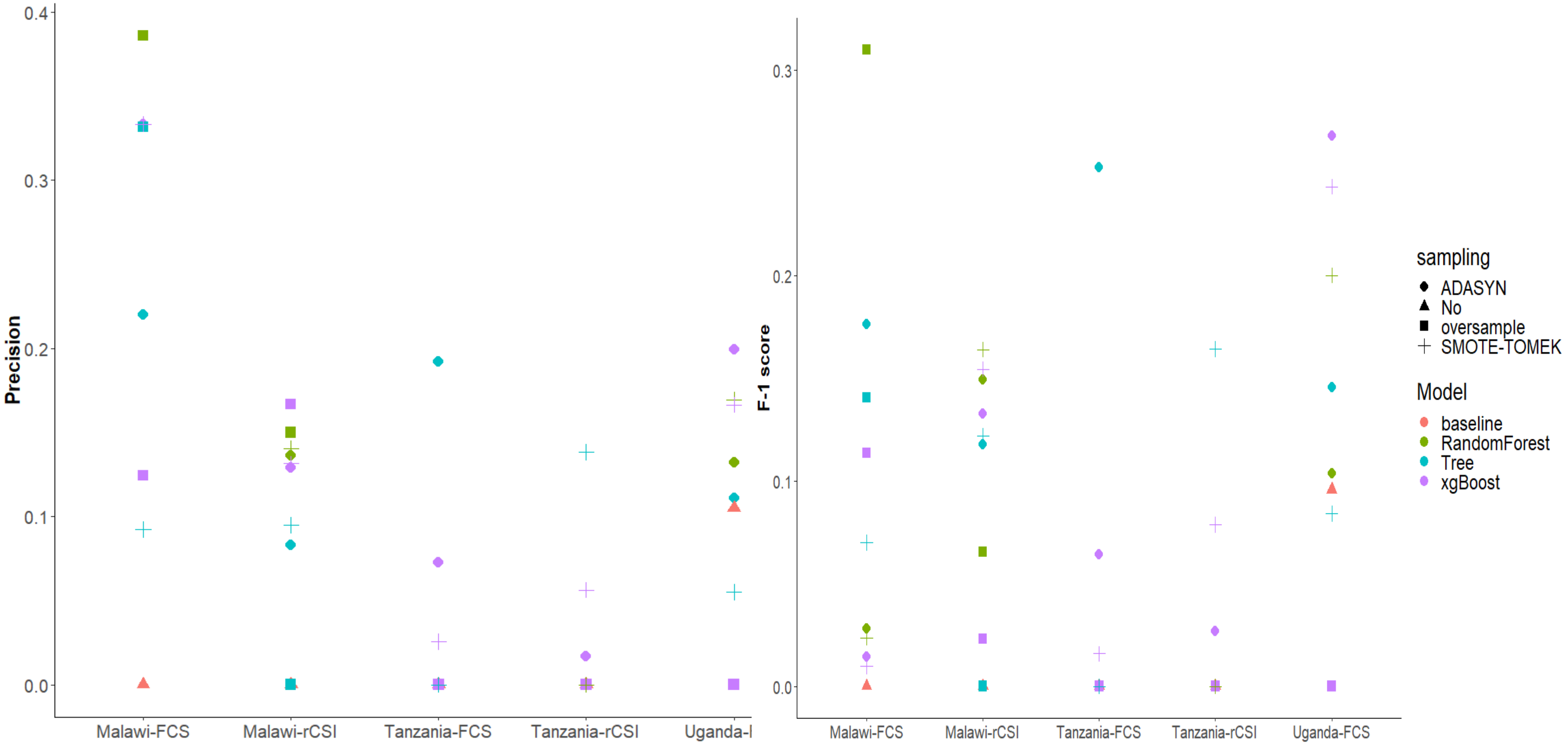




Precision
0.6
0.4
0.2
0.0
F-1 score
0.15
0.10
0.05
0.00
Model
baseline
RandomForest
Tree
xgBoost
Malawi-FCS
Malawi-rCSI
Tanzania-FCS
Tanzania-rCSI
Uganda-FCS

Baseline vs ML algorithms with down/over sample technique



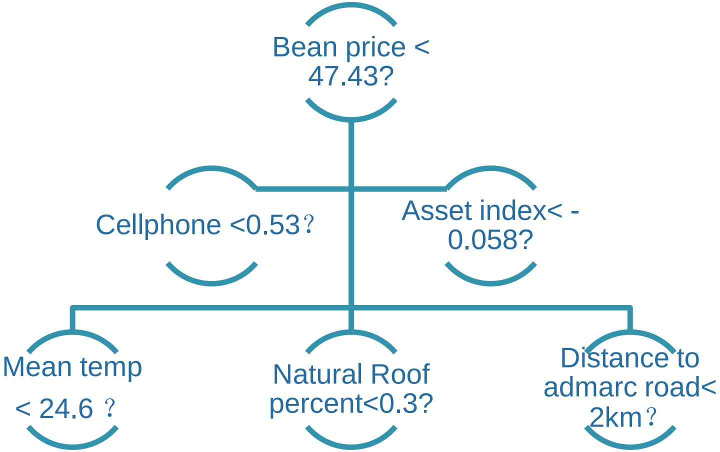


Precision
0.4
0.3
0.2
0.1
0.0
Malawi-FCS
Malawi-rCSI
Tanzania-FCS
Tanzania-rCSI
Uganda-I
F-1 score
0.3
0.2
0.1
0.0
Malawi-FCS
Malawi-rCSI
Tanzania-FCS
Tanzania-rCSI
Uganda-FCS
sampling
ADASYN
No
oversample
SMOTE-TOMEK
Model
baseline
RandomForest
Tree
xgBoost

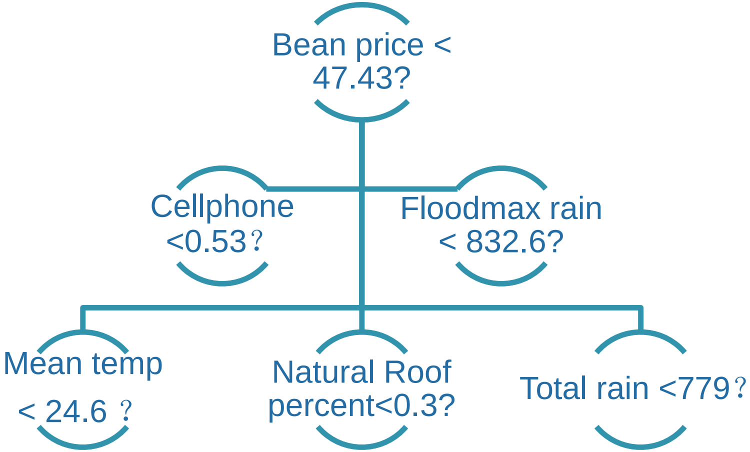
For most severe food security category with oversampling

Country	Food Security Measure	Overall Accuracy (baseline)	Overall Accuracy (ML)	Recall Rate Insecure category (baseline)	Recall Rate Insecure category (ML)
Malawi	FCS	0.70	**0.69-0.75**	0.00	0.01-0.27
	rCSI	0.67	**0.58-0.63**	0.00	0.00-0.20
Tanzania	FCS	0.83	**0.74-0.84**	0.00	0.00-0.40
	rCSI	0.54	**0.55-0.60**	0.00	0.00-0.52
Uganda	FCS	0.51	**0.62-0.68**	0.37	0.00-0.57

Feature importance



Original



Oversample

Random Forest Feature Importance

Variable	Importance	Std
number_celphones	0.12	0.11
cell_phone	0.09	0.10
roof_natural	0.05	0.06
asset_index	0.04	0.03
FS_month	0.04	0.02
floor_dirt_sand_dung	0.03	0.06
dist_popcenter	0.03	0.02
dist_road	0.03	0.02
percent_ag	0.03	0.02
maxdaysnorain	0.03	0.02
lhz_beans_price	0.03	0.02
dist_admarc	0.03	0.02
floodmax	0.02	0.02
clust_maize_mktthin	0.02	0.02

Original

Variable	Importance	Std
roof_natural	0.11	0.07
cell_phone	0.09	0.10
floor_dirt_sand_dung	0.08	0.04
number_celphones	0.05	0.10
roof_iron	0.04	0.06
day1rain	0.04	0.01
clust_beans_price	0.04	0.01
lhz_maxdaysnorain	0.03	0.02
lhz_nuts_mktthin	0.03	0.01
asset_index	0.03	0.02
household_head_age	0.03	0.02
clust_maize_price	0.03	0.02
dist_road	0.03	0.02

Oversample

XGBOOST Feature Importance

Variable	Importance
roof_natural	0.11
cell_phone	0.09
floor_dirt_sand_dung	0.08
number_celphones	0.05
roof_iron	0.04
day1rain	0.04
clust_beans_price	0.04
lhz_maxdaysnorain	0.03
lhz_nuts_mktthin	0.03
asset_index	0.03
Household_head_age	0.03
clust_maize_price	0.03
dist_road	0.03
clust_maize_mktthin	0.02

Malawi

Variable	Importance
Cellphone	0.11
num_cell	0.09
floor_dirt_sand_dung	0.07
roof_iron	0.06
asset_index	0.06
dist_popcenter	0.06
lhz_day1rain	0.06
lhz_maize_price	0.06
dist_road	0.06
Household_head_age	0.05
region19	0.05
percent_ag	0.05
region9	0.05

Tanzania

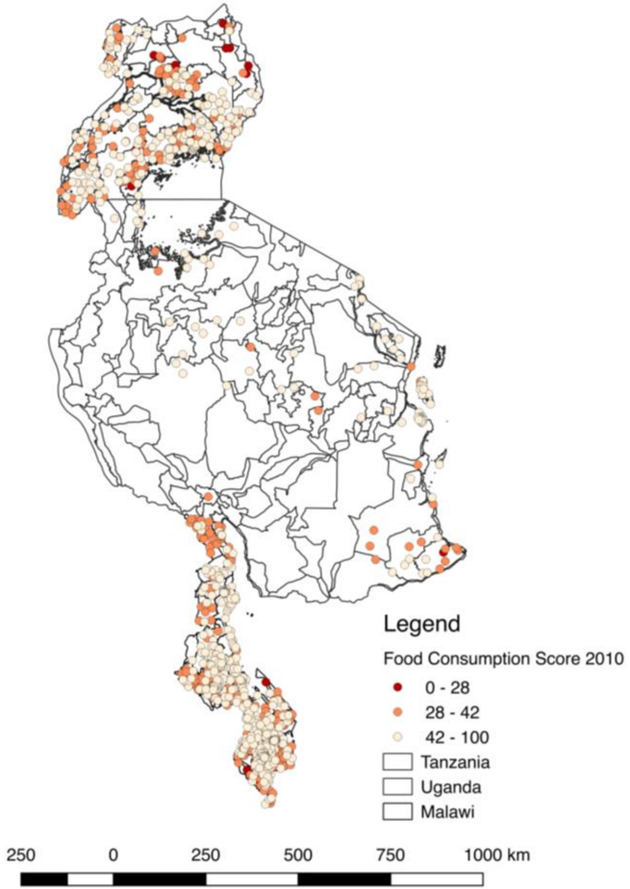
Variable	Importance
num_cellphones	0.15
roof_iron	0.15
region3	0.15
dist_road	0.12
floor_dirt_sand_dung	0.12
roof_natural	0.12
Cellphone	0.12
floodmax	0.08

Uganda

Methodology: Data Split

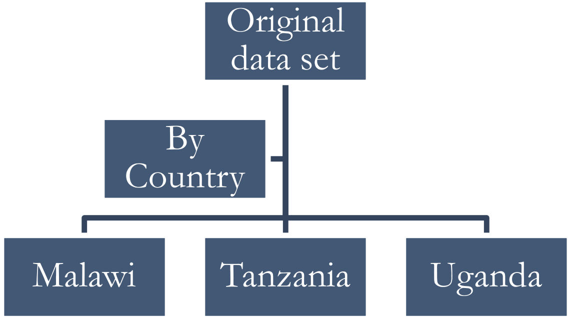
Split by year, by region or random

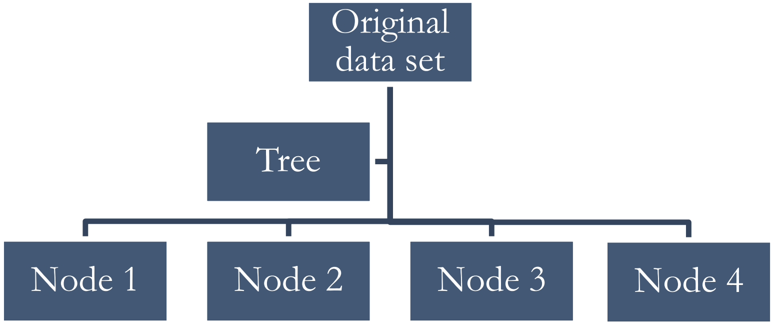
For different application purposes and different data structures



Methodology: Data Segmentation

1. By country
2. Entire dataset of three countries
3. By urban and rural
4. Auto-segmentation by training a shallow tree in each country based on observables





Ongoing work

- Price variable selection: which crop, where, and when
- Asset/roof variable prediction from satellite imagery
- Spatial-temporal correlation on data split
- Continuous approach: count of insecure HH for each cluster

Future Steps:

- Model generalization: What happens when we directly apply models trained on one country/region to predict another
- Model deploy and update: Compare the results of using one year, with a dynamic process of constantly updating model with new survey data
- Prediction at "grid" level where satellite imagery available

Conclusions

1. 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.

2. 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.

Thank you !