Predicting Food Security with Machine Learning

Yujun Zhou, Kathy Baylis, Erin Lentz, and Hope Michelson

Selected Paper prepared for presentation at the International Agricultural Trade Research Consortium's (IATRC's) 2019 Annual Meeting: Recent Advances in Applied General Equilibrium Modeling: Relevance and Application to Agricultural Trade Analysis, December 8-10, 2019, Washington, DC.

Predicting Food Security with Machine Learning

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

IATRC Annual Meeting
Washington DC,
December 8-10, 2019

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 has large data requirements and has been accused of political influence

Need to improve prediction of food security crises

The opportunity

- Recent increase in available data related to food security, rainfall, and prices.
- These data are often evaluated in isolation.

Incorporate these data into a single predictive model of food security early warning.

Objective

- To build an early warning system of food security in areas where data are scarce and data collection is costly
  - That captures the majority of food insecure households through data techniques
  - That can be automatically updated, generalizable, scalable and cost-effective

Follow on flurry of prediction using remotely sensed data

- Village-level poverty (asset index) using night lights
- Combining night-lights and satellite imagery in a CNN model (up to 70% accuracy)
- …works in some areas better than others (SSA; Nepal and Haiti are problematic)
- …and does not capture changes in poverty over time
- …and does not do so well with other development metrics

What we do

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

What we do

- 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

What we do

- 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

What we do

- Build ML models to predict cluster-level food security status for targeting, aid purposes in times of food shortage
- Use LSMS data for Malawi, Tanzania and Uganda as ground truth
- Use market price of food staples, weather shocks in growing seasons, and geospatial features around clusters to predict potential food security challenges
- Use data techniques (oversampling, data segmentation) to improve prediction performance

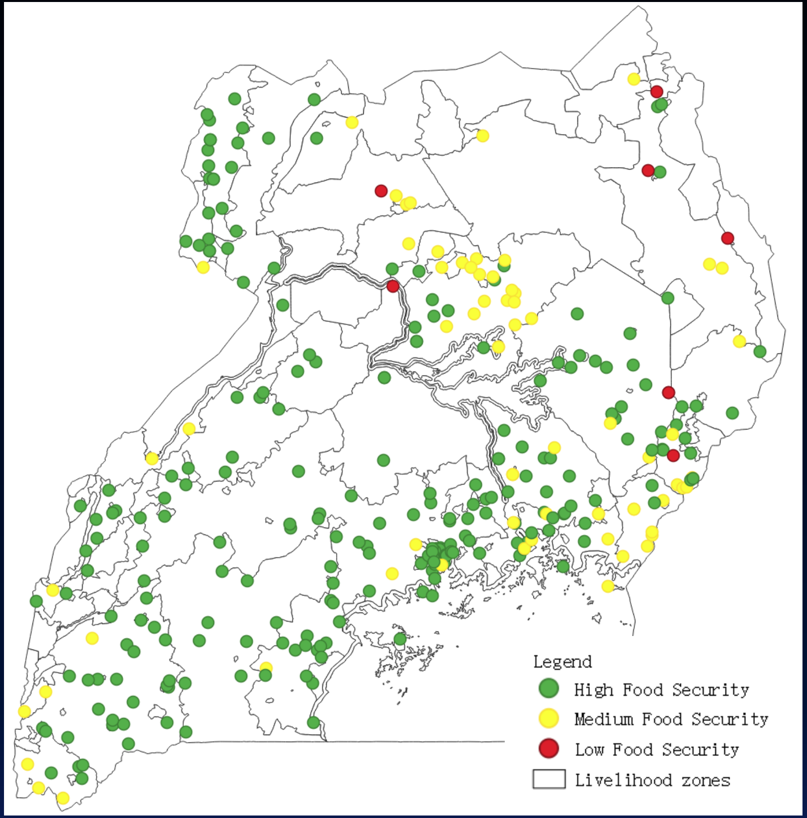
What we do

- Build ML models to predict cluster-level food security status for targeting, aid purposes in times of food shortage
- Use LSMS data for Malawi, Tanzania and Uganda as ground truth
- Use market price of food staples, weather shocks in growing seasons, and geospatial features around clusters to predict potential food security challenges
- Use data techniques (oversampling, data segmentation) to improve prediction performance
- Correctly categorize 63-84 % of food insecurity categories and up to 20-57% of most food insecure category.

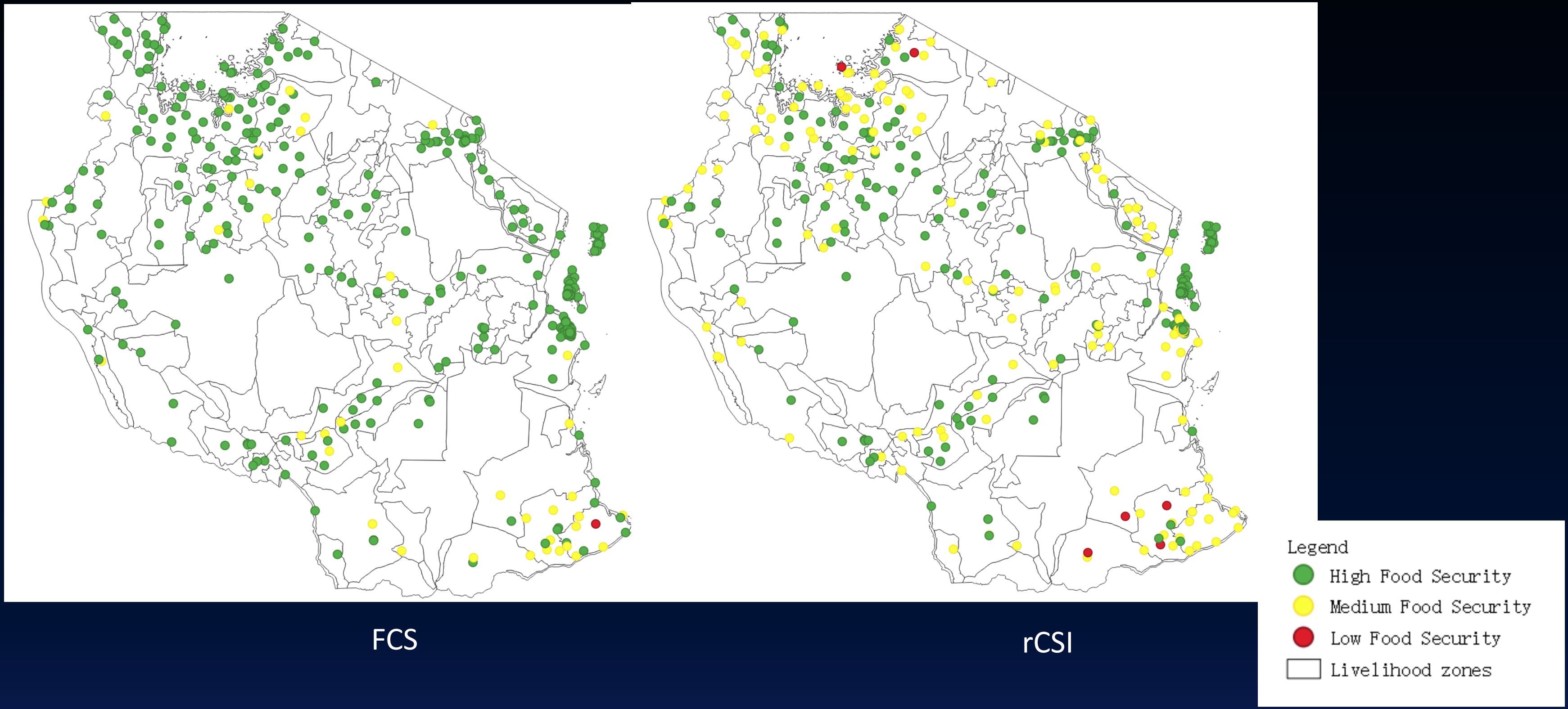
Data

- LSMS survey ground truth data
- Cluster averages
- Categorized using cutoffs
- Uganda/Tanzania/Malawi
- Three different rounds with broad spatial coverage

Uganda FCS



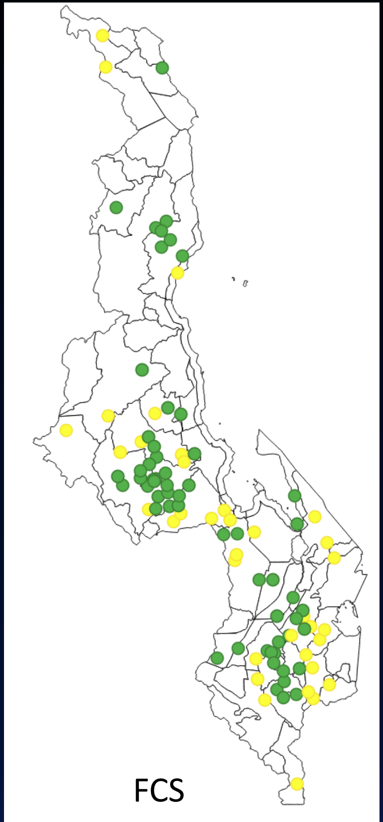
Tanzania

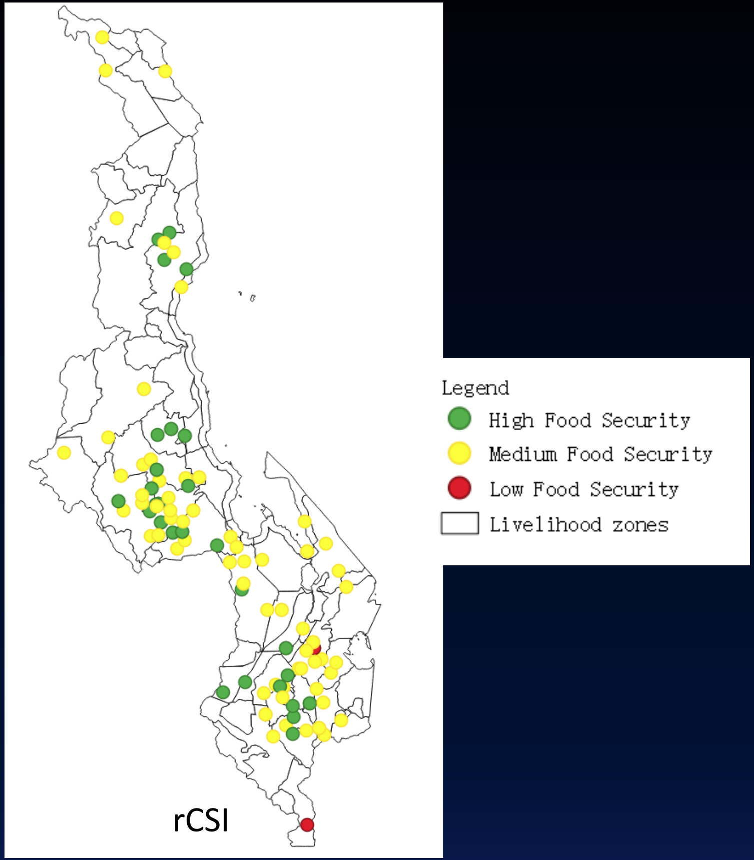


FCS

rCSI

Malawi


FCS


Legend
● High Food Security
● Medium Food Security
● Low Food Security
□ Livelihood zones
rCSI

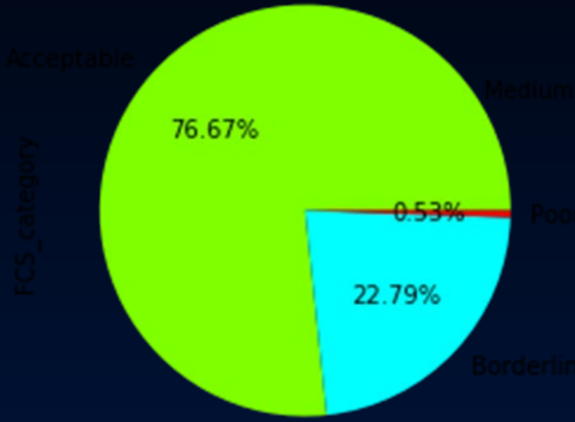
Decisions, decisions

1. Categorical versus continuous prediction
2. If categorical, how do we address rare events?
3. What algorithm do we use? And how do we assess it?
4. How do we split the data?

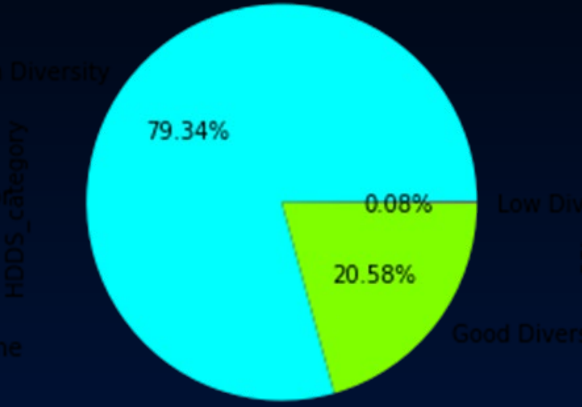
Categorical vs. Continuous

- Focus on categorical prediction for the food security cutoffs
    - Policy-relevant
    - Classifiers are more sensitive to the majority class
    - Recall rate of the insecure villages is more important than accuracy
    - Apply down sampling, over sampling, and synthetic data techniques to force the model to learn about the tail of the distribution

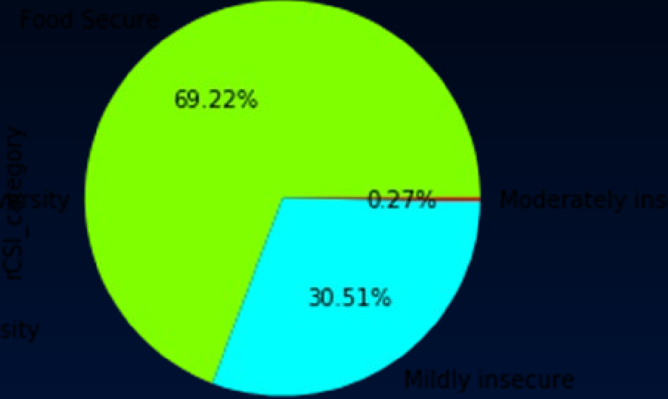
Challenge: detecting rare but relevant households



FCS

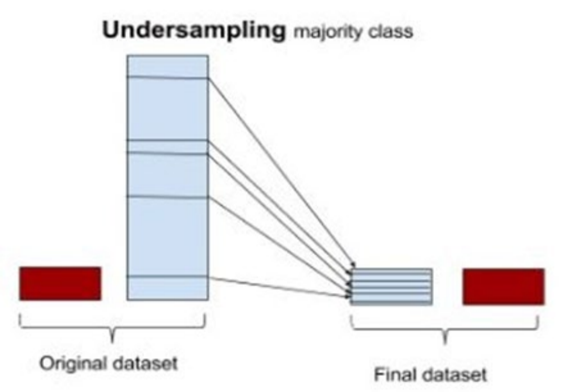


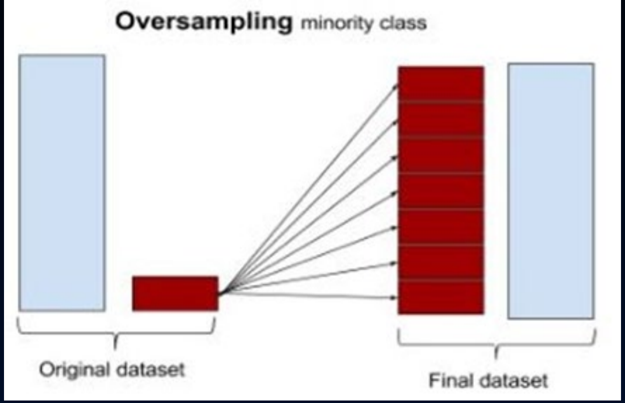
HDDS



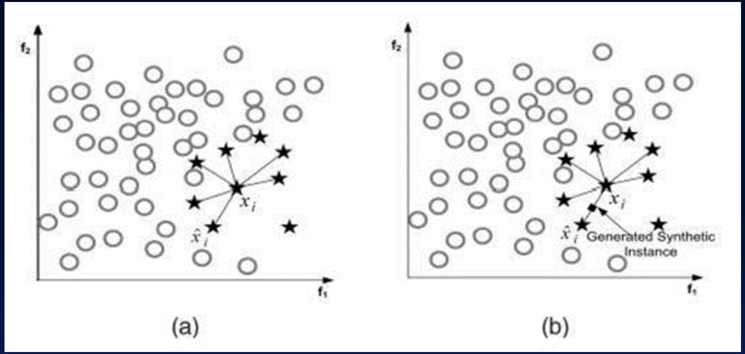
rCSI

Methods: Sampling design





SMOTE



Classification Algorithm

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

Compare to a Baseline

Logistic Regression

Data split: year split (cross-validated)

Data segmentation : by country

Down/over sampling: None

Variable groups:

Market: food price, market thinness

Asset: cellphone ownership, floor/roof material, asset index

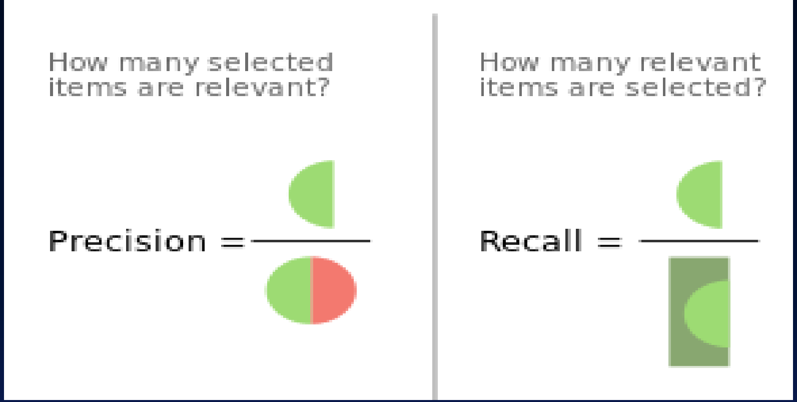
Weather: dry spells, average temperature and rain

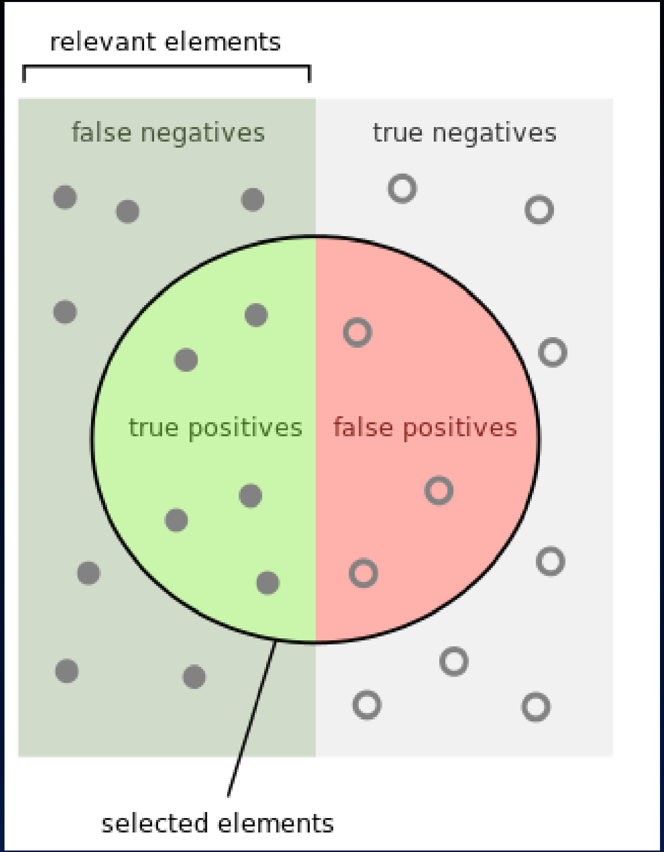
Location: elevation, distance to road, urban/rural

At village, district and regional level

Results Metrics

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

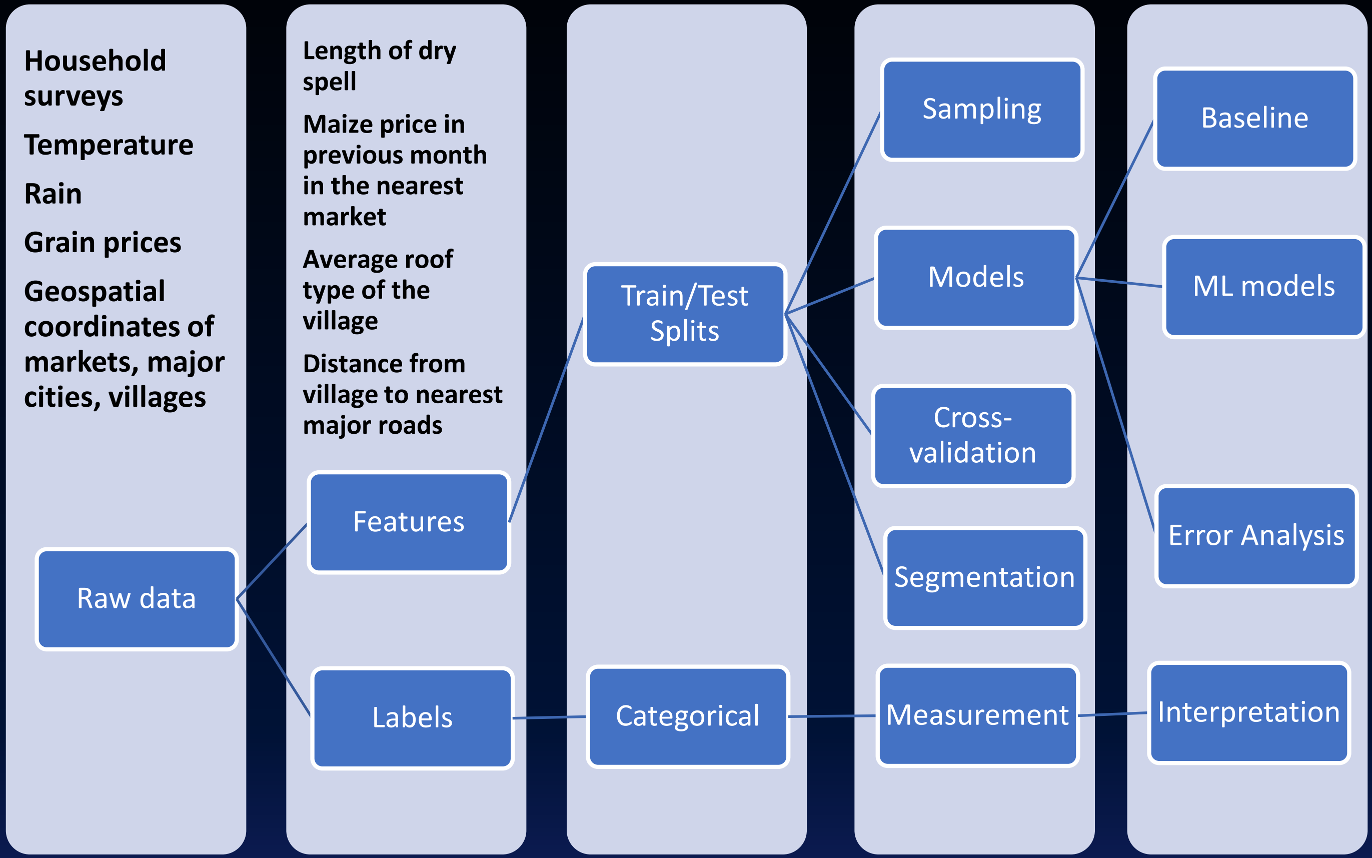

How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =



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 > 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

Putting things together…



Household surveys

Temperature

Rain

Grain prices

Geospatial coordinates of markets, major cities, villages

Raw data

Length of dry spell

Maize price in previous month in the nearest market

Average roof type of the village

Distance from village to nearest major roads

Features

Labels

Train/Test Splits

Categorical

Sampling

Models

Cross-validation

Segmentation

Measurement

Baseline

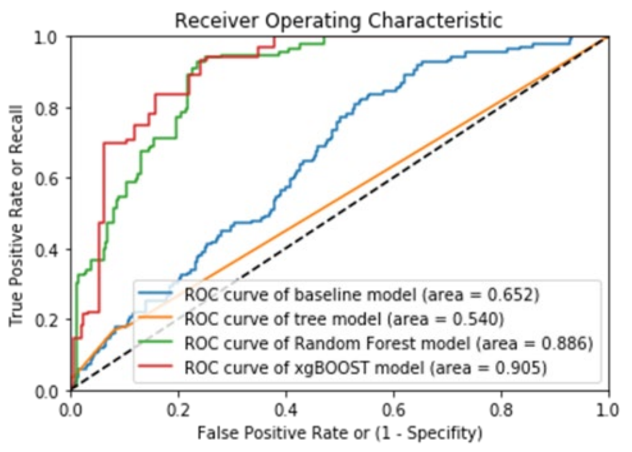
ML models

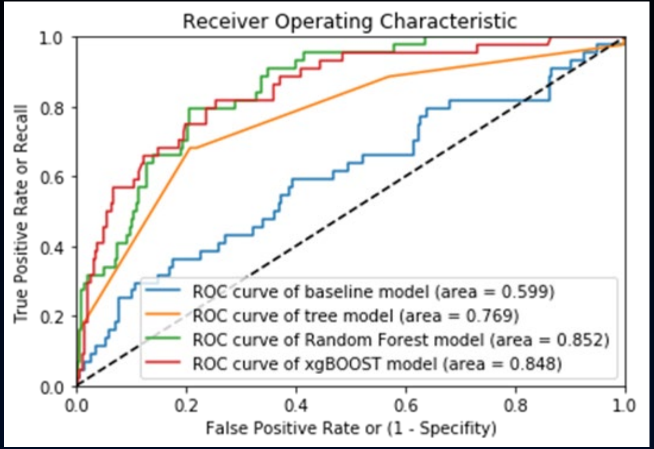
Error Analysis

Interpretation

Results for binary cutoff

FCS

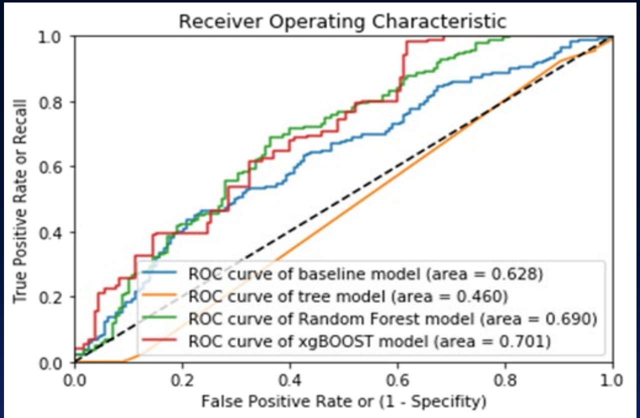


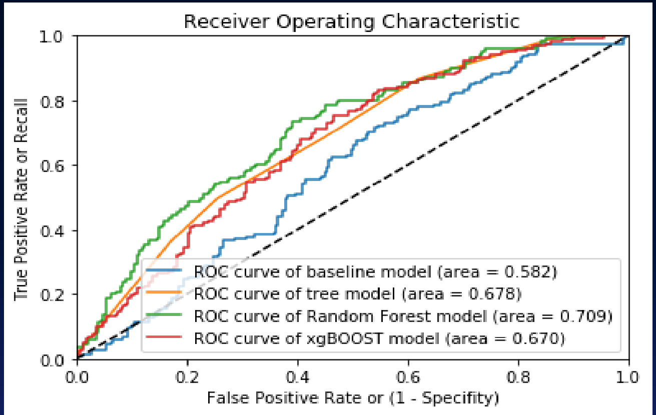


Malawi

Tanzania

rCSI





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

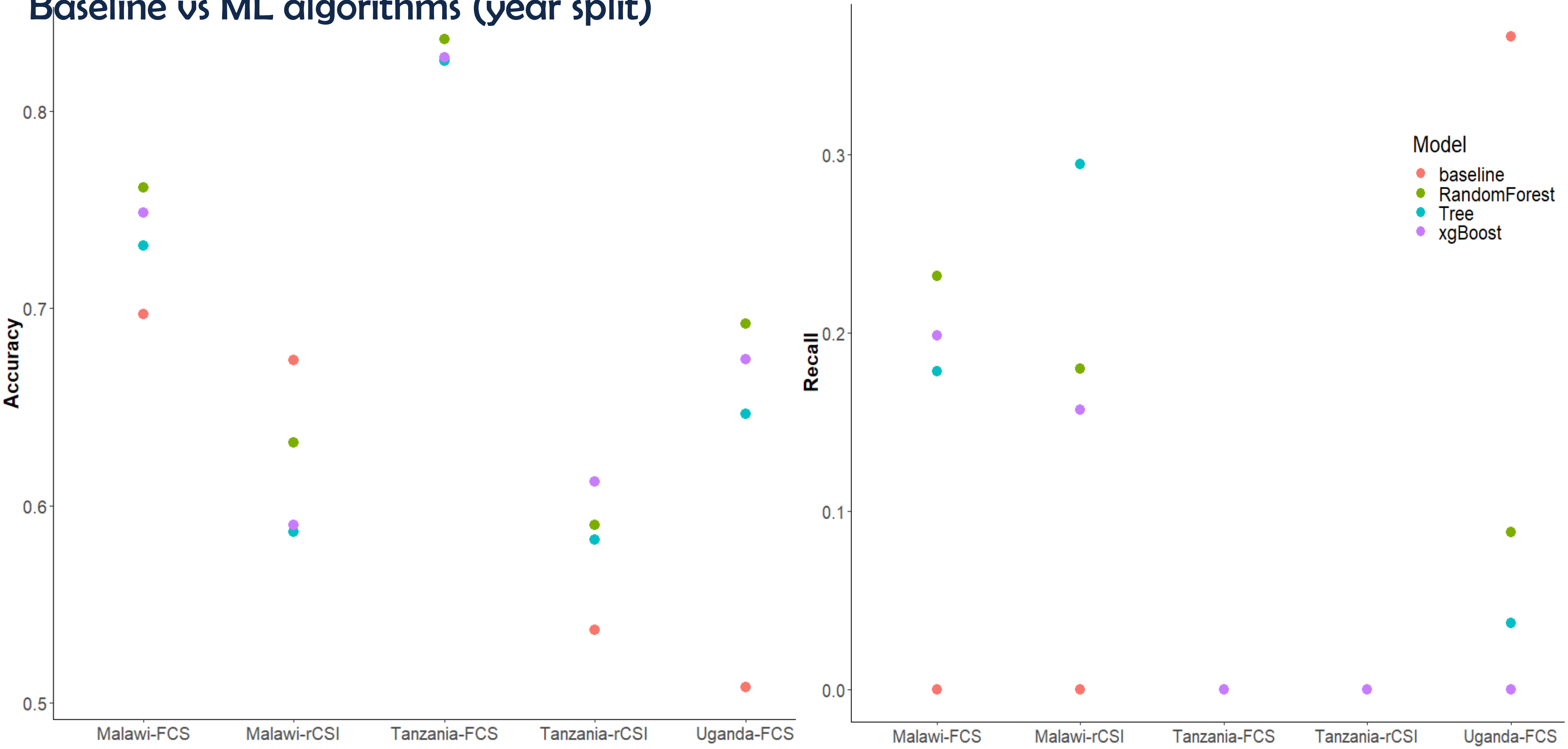
Similar accuracy, higher recall

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 2010/11 to predict 2012	FCS	0.67	**0.59-0.71**	0.36	0.33-0.36

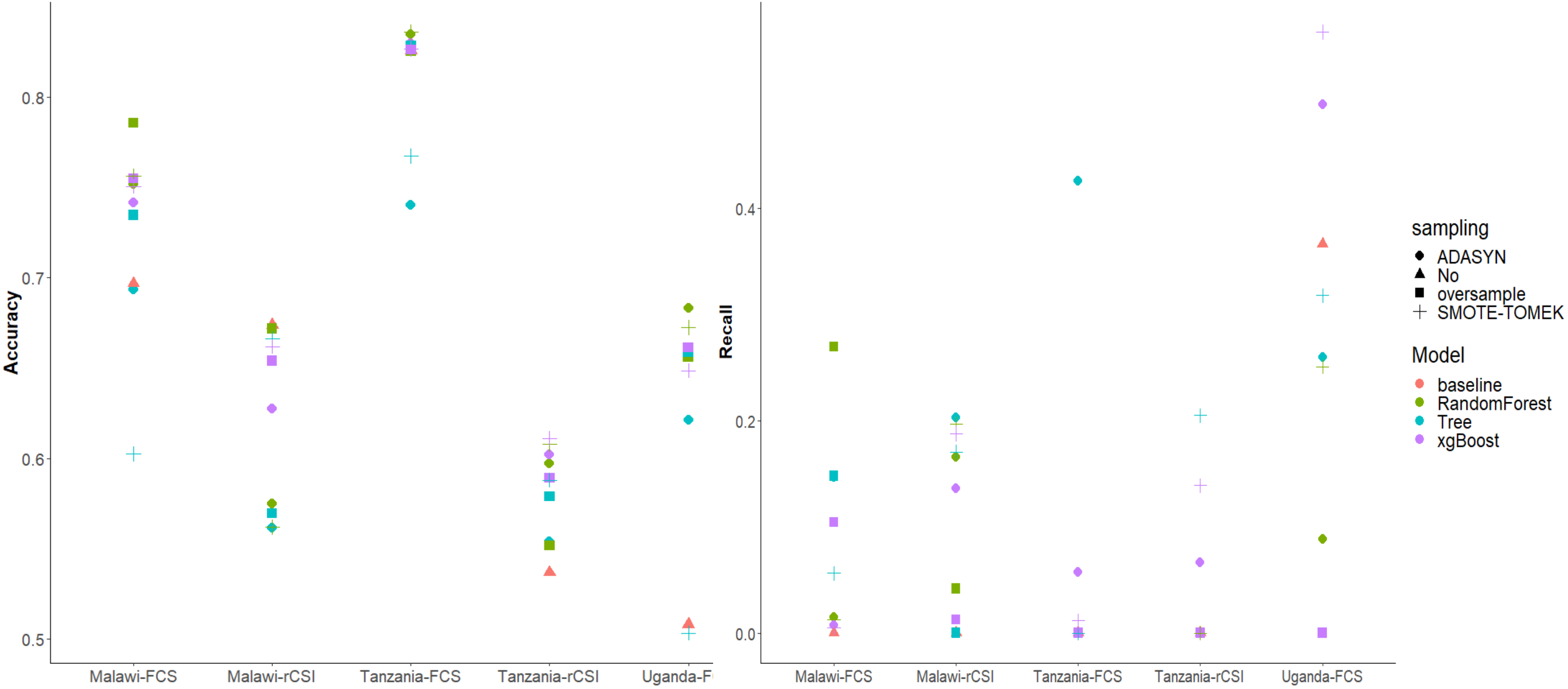
For most severe food security category with oversampling

Country	Food Security Measure	Overall Accuracy (baseline)	Overall Accuracy (ML)	Recall Rate Insecure category (baseline)	Recall Rate Insecure category (ML)
Malawi 2010/11, 2013 to predict 2015/16	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 2010/11, 2012/13 to predict 2014/15	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 2010/11 to predict 2012	FCS	0.51	0.62-0.68	0.37	0.00-0.57

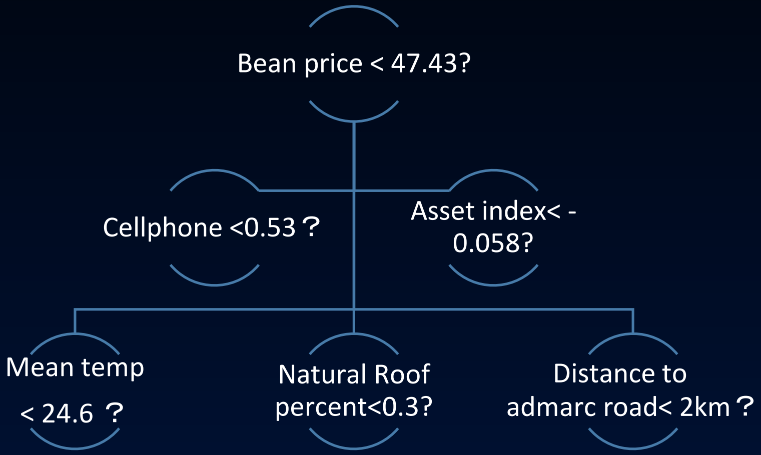
Baseline vs ML algorithms (year split)



Baseline vs ML algorithms with down/over sample technique



Feature Importance: Tree



Bean price < 47.43?

Cellphone <0.53？

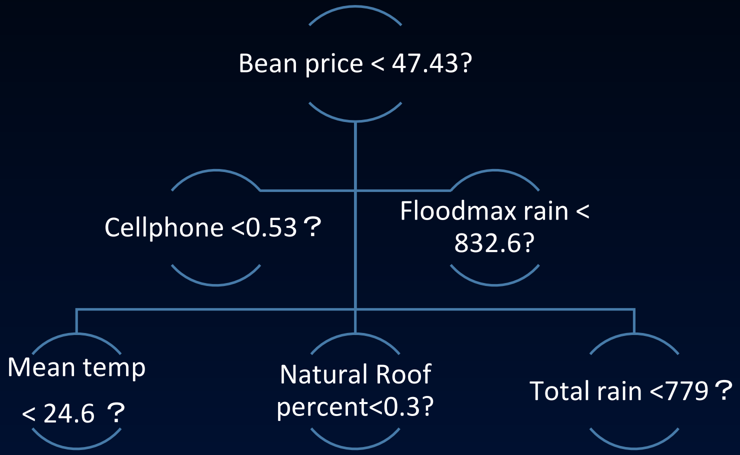
Asset index< -0.058?

Mean temp < 24.6 ？

Natural Roof percent<0.3?

Distance to admarc road< 2km？

Original



Bean price < 47.43?

Cellphone <0.53？

Floodmax rain < 832.6?

Mean temp < 24.6 ？

Natural Roof percent<0.3?

Total rain <779？

Oversample

Feature Importance: Random Forest

Variable	Importance	Std
# of cellphones	0.12	0.11
cellphone	0.09	0.10
Natural roof	0.05	0.06
Asset index	0.04	0.03
Month	0.04	0.02
Dirt Flood	0.03	0.06
Distance to popcenter	0.03	0.02
Distance to road	0.03	0.02
% ag land	0.03	0.02
Dry spell	0.03	0.02
Price of beans	0.03	0.02
Distance to ag market	0.03	0.02
High rains in flood zone	0.02	0.02
Maize market thinness	0.02	0.02

Original

Variable	Importance	Std
Natural roof	0.11	0.07
cellphone	0.09	0.10
Dirt floor	0.08	0.04
# of cellphones	0.05	0.10
Iron roof	0.04	0.06
When rains begin	0.04	0.01
Price beans	0.04	0.01
Dry spell	0.03	0.02
Nut market availability	0.03	0.01
Asset index	0.03	0.02
Age household head	0.03	0.02
Maize price	0.03	0.02
Distance to road	0.03	0.02

Oversample

Data Split

Split by year, by region or random

For different application purposes and different data structures

Feature importance: xgboost for FCS

Variable	Importance
Natural roof	0.11
Cellphone	0.09
Dirt floor	0.08
# cellphones	0.05
Iron roof	0.04
Start of the rainy season	0.04
Village bean price	0.04
district length of dryspell	0.03
District market nut avail.	0.03
Asset index	0.03
Age household head	0.03
Village maize price	0.03
Distance to road	0.03
Village maize availability	0.02

Malawi

Variable	Importance
Cellphone	0.11
# cellphones	0.09
Dirt floor	0.07
Iron roof	0.06
Asset index	0.06
Distance to popcenter	0.06
Start of rainy season	0.06
Maize price	0.06
Distance to road	0.06
Age household head	0.05
Region dummy	0.05
% Ag land	0.05
Region dummy	0.05

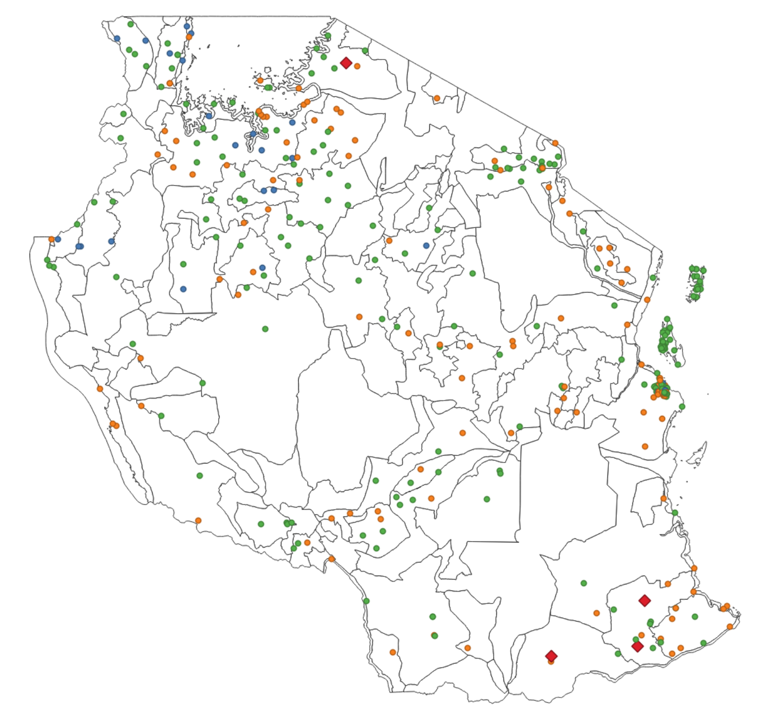
Tanzania

Variable	Importance
# cellphones	0.15
Iron roof	0.15
Region dummy	0.15
Distance to road	0.12
Dirt floor	0.12
Natural roof	0.12
Cellphone	0.12
Heavy rain in floodprone regions	0.08

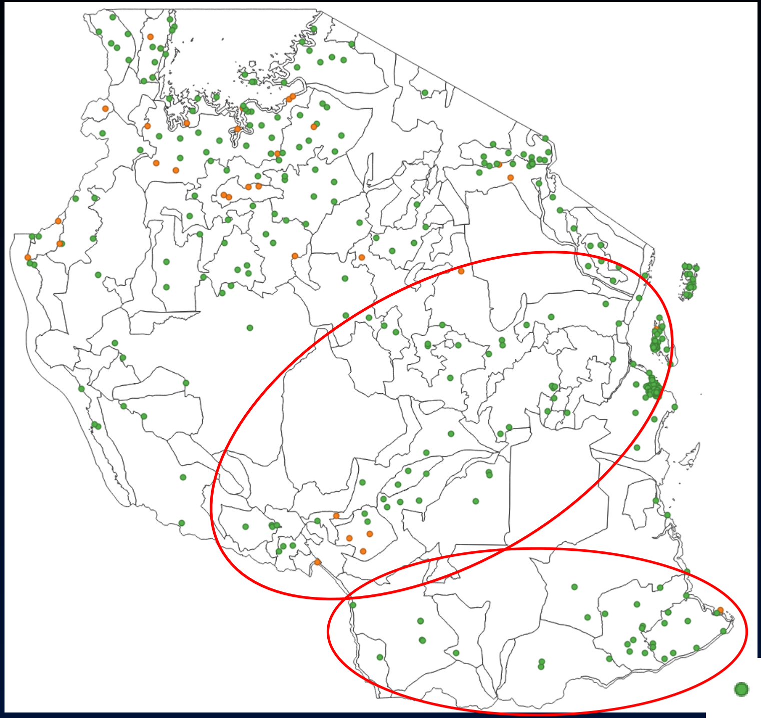
Uganda

Error Analysis

Tanzania rCSI



Baseline



ML + oversample

🟢 Corret

🟠 Overpredict by one

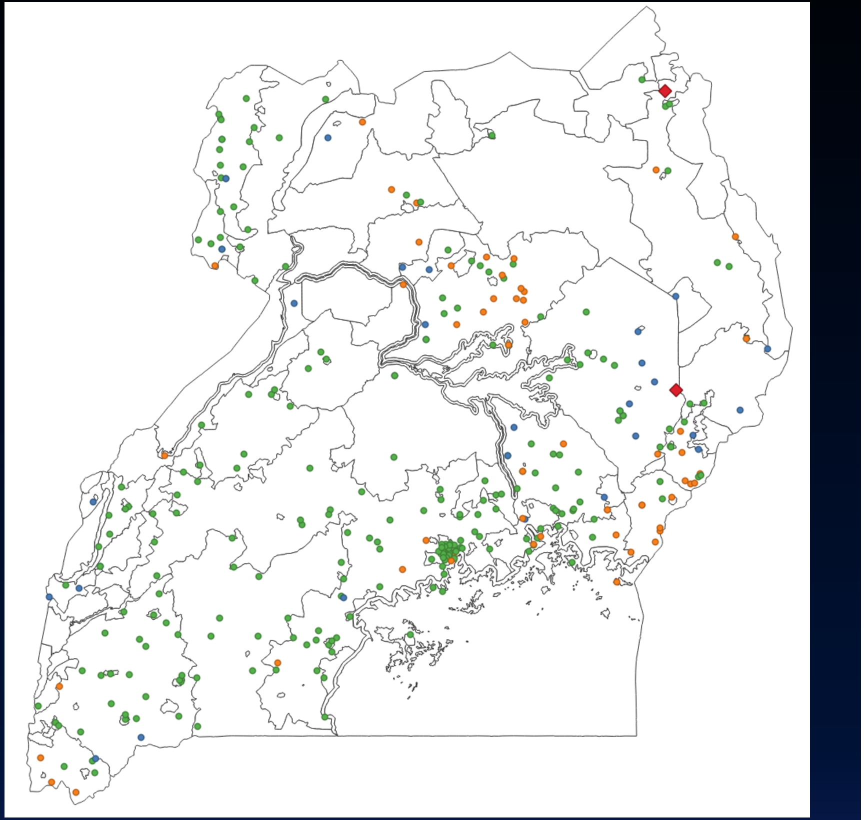
♦️ Overpredict by two

🔵 Underpredict by one

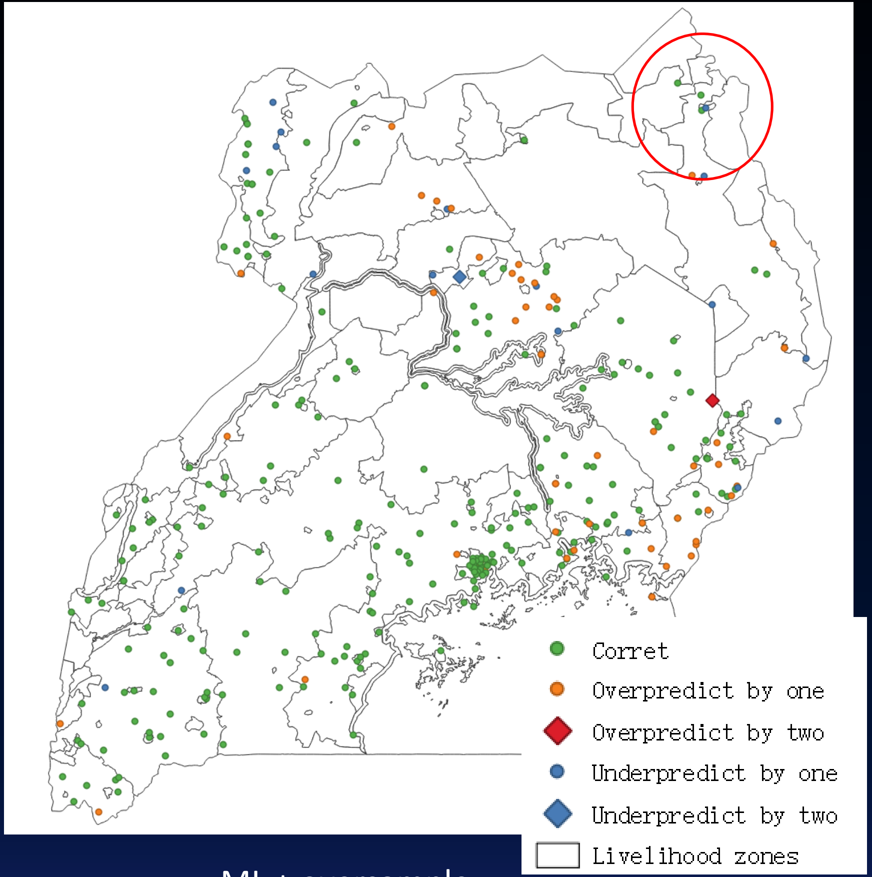
🔷 Underpredict by two

□ Livelihood zones

Uganda FCS

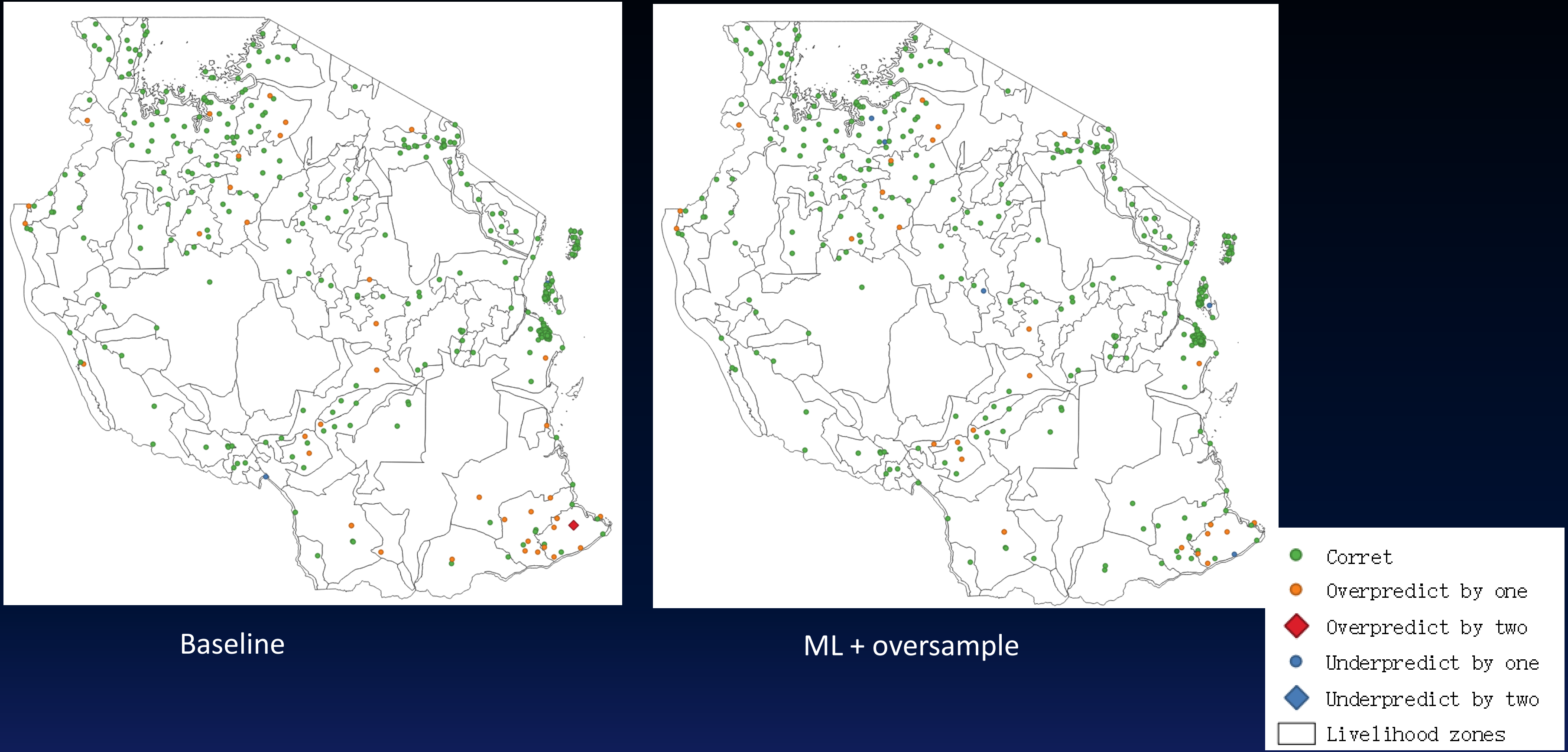


Baseline



ML + oversample

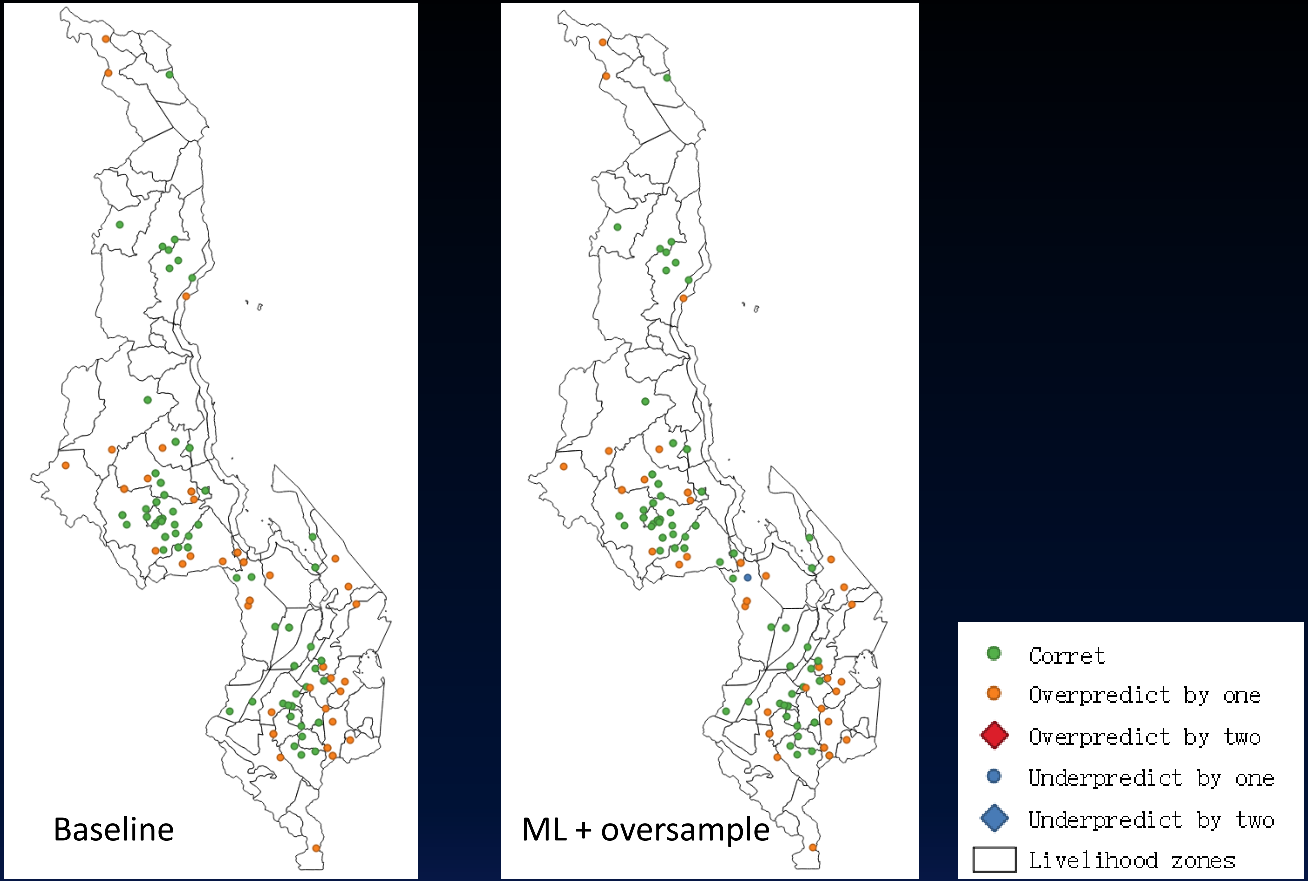
Tanzania FCS



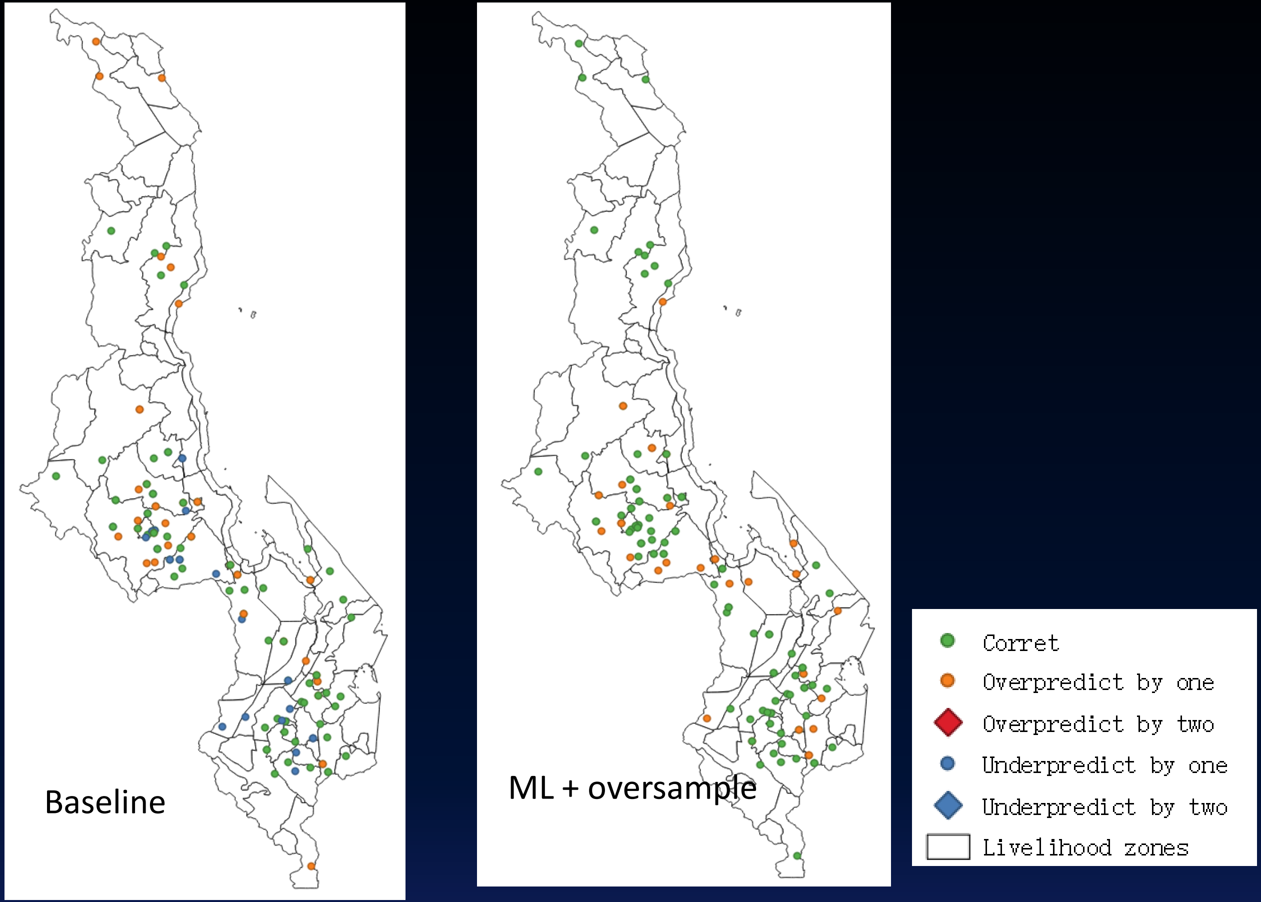
Baseline

ML + oversample

Malawi FCS


Baseline
ML + oversample
● Corret
● Overpredict by one
◆ Overpredict by two
● Underpredict by one
◆ Underpredict by two
□ Livelihood zones

Malawi rCSI


Baseline
ML + oversample
● Corret
● Overpredict by one
◆ Overpredict by two
● Underpredict by one
◆ Underpredict by two
□ Livelihood zones

Next steps

1. Error analysis and feature importance analysis
    - by region
    - by group
    - by month

2. Model generalization

   What happens when we directly apply models trained on one country/region to predict another

3. Model deploy and update

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

Conclusions: May be on to something…?

1. Combined with data techniques, machine learning methods not only improve prediction accuracy in general, but particularly of 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.