Predicting Food Security in Sub-Saharan Africa with Machine Learning

Yujun Zhou, Kathy Baylis, Erin Lentz, and Hope Michelson
(with most excellent research support by Manny Kim)
ASSA Annual Meeting, January 2-4, 2021
on a monitor near you

Problem

- We need to identify food insecure populations in time to intervene
- But crises are rare events, tricky to predict

Opportunity

- Novel data sources and analytical methods

Objective

- Can we 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

  …that can be automatically updated, generalizable, scalable and cost-effective

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 3 years of LSMS data for Malawi, Tanzania and Uganda as ground truth; use first 2 years to predict most recent year

What we do

- Build ML models to predict cluster-level food security status for targeting, aid purposes in times of food shortage
- Use 3 years of LSMS data for Malawi, Tanzania and Uganda as ground truth; use first 2 years to predict most recent year
- 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 3 years of LSMS data for Malawi, Tanzania and Uganda as ground truth; use first 2 years to predict most recent year
- 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, cost-sensitive learning) 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 3 years of LSMS data for Malawi, Tanzania and Uganda as ground truth; use first 2 years to predict most recent year
- 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, cost-sensitive learning) to improve prediction performance
- Machine learning models ~ 70-83% probability of being right. Can capture 70% food insecure villages for 50-80% accuracy (90% recall for 40-60%)

What do we find?

- Build ML models to predict cluster-level food security status for targeting, aid purposes in times of food shortage
- Use 3 years of LSMS data for Malawi, Tanzania and Uganda as ground truth; use first 2 years to predict most recent year
- 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, cost-sensitive learning) to improve prediction performance
- Machine learning models ~ 70-83% probability of being right. Can capture 70% food insecure villages for 50-80% accuracy (90% recall for 40-60%)

Tradeoff between recall and precision that should be determined by policy use

Decisions, decisions

1. What do we target?

   Villages with > 20% food insecure households

2. How do we address rare events?

   Over-sampling and cost-sensitive learning

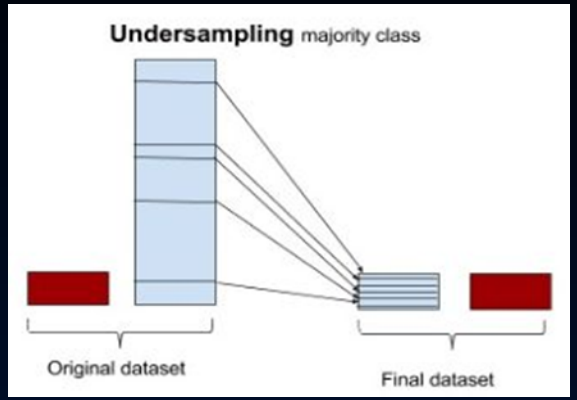
3. What algorithm do we use? And how do we assess it?

   Tree-based methods and ROC curves (can target to policy objectives)

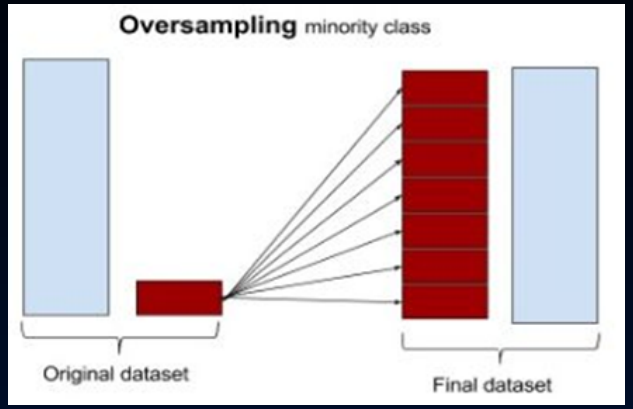
4. How do we split the data?

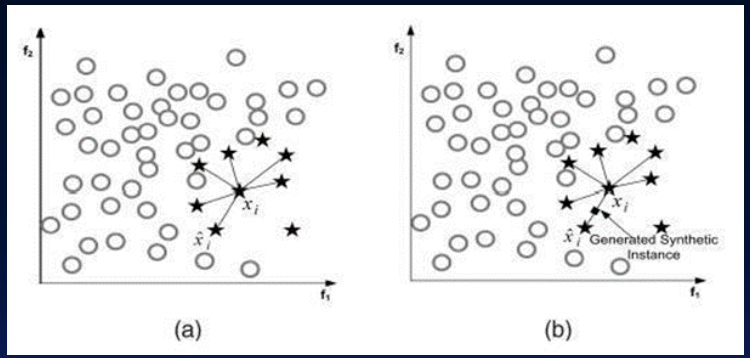
   Test on last year (policy relevant; minimizes issues of correlation between train and test set)

Methods: Sampling design



SMOTE





Models

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

vs

Random Forests and Gradient Boosting
Data split: year split (cross-validated)
Data segmentation : by country
Down/over sampling: None, Oversampling
SMOTE, ADASYN

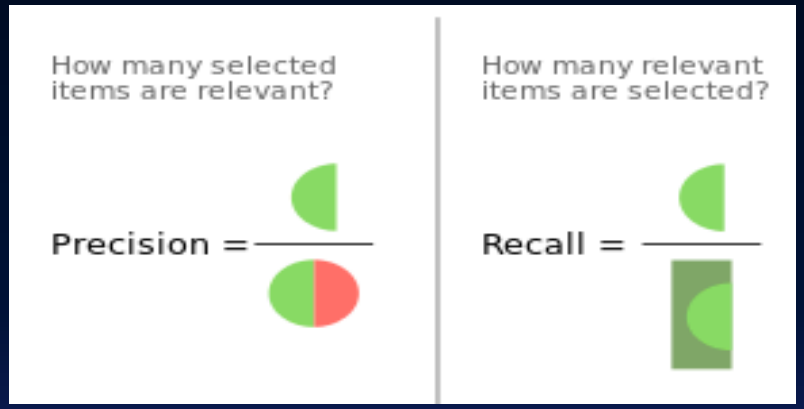
Variable groups ("Features")

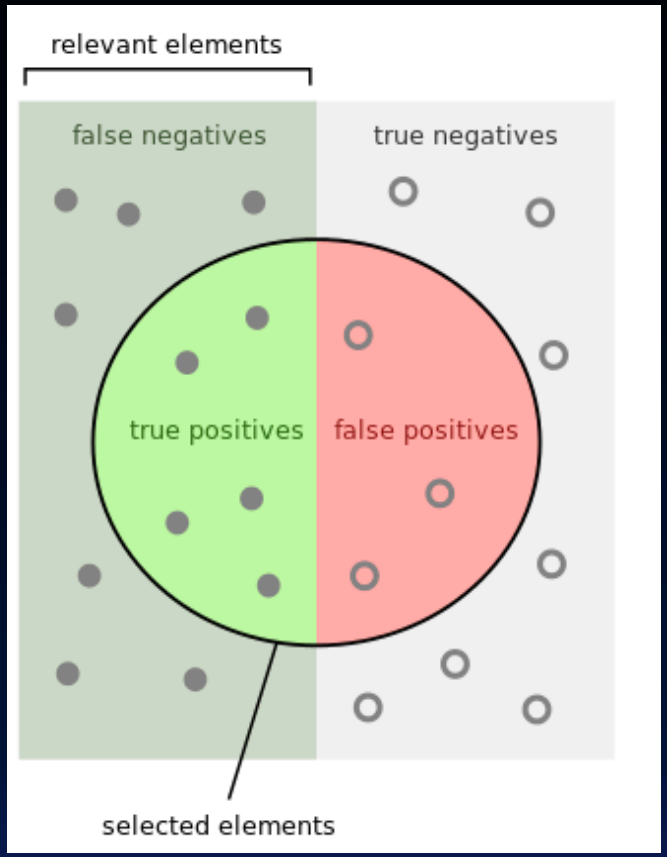
Market: food prices, market thinness
Asset: cellphone ownership, floor/roof material, asset index
Weather: dry spells, average temperature, growing degree days, heating degree days, total rain, start date of rains
Location: elevation, distance to road, urban/rural
At village, district and regional level
Month and region fixed effects

Results Metrics

1. Recall (are we getting all the insecure households ?)
2. Precision (are we mistaking secure households as insecure?)
3. Overall categorical accuracy

…tradeoffs… Ideally set metric based on use





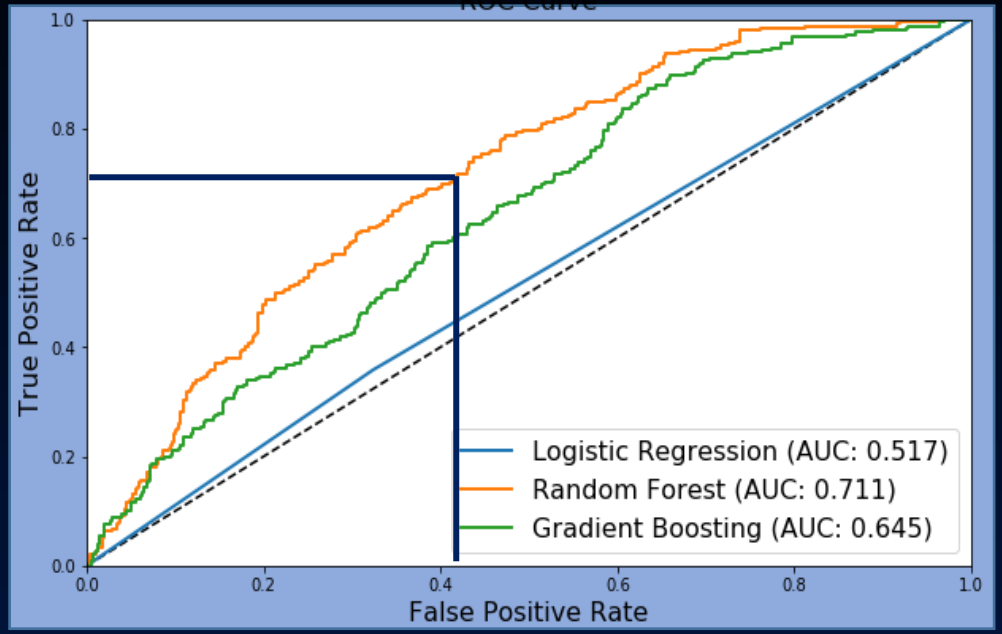
Metrics: Receiving Operator Characteristic - Area Under the Curve (ROC-AUC)

Maps true positive rate vs false positive rate as one varies the cutoff level used to determine classification

The larger the area between the curve, the better the model

1 = perfect
0.5 = 50/50 no better than random
0 = perfectly wrong

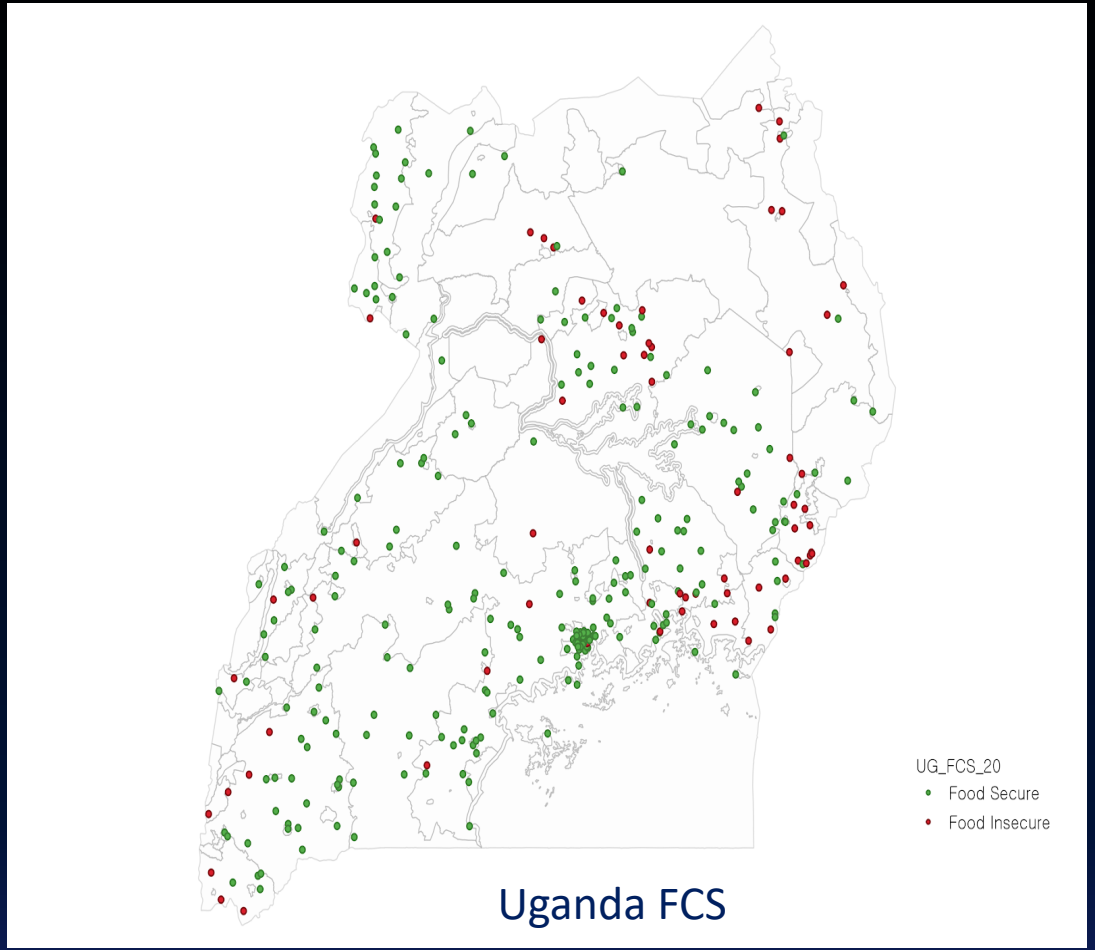
For any chosen positive rate, can determine the % of false positives
Alternatively, could chose the acceptable ratio



Malawi FCS

Data

- LSMS survey data as ground truth
- Uganda/Tanzania/Malawi
- Categorized at 20% of households in village food insecure using Food Consumption Score (FCS), reduced Coping Strategies Index (rCSI)
- Three different rounds with broad spatial coverage



Uganda FCS

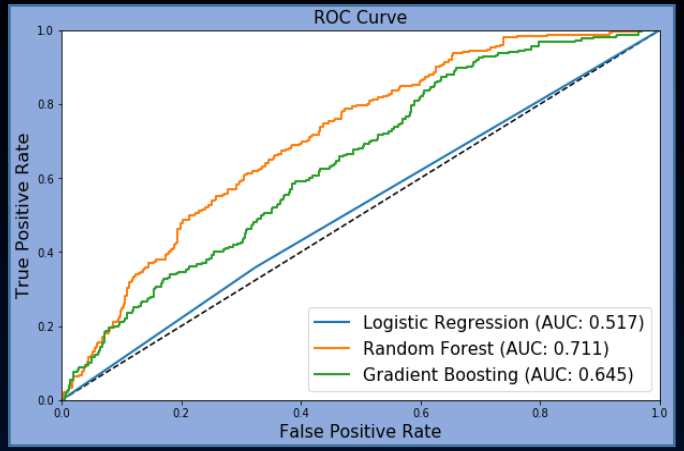
Zhou, Baylis, Lentz, and Michelson

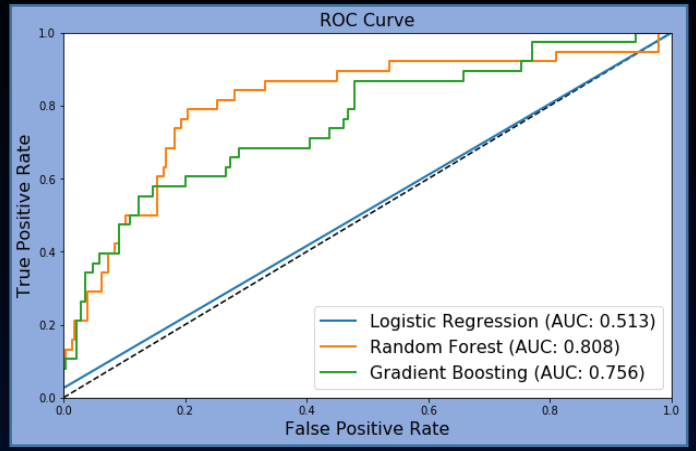
Results

Malawi

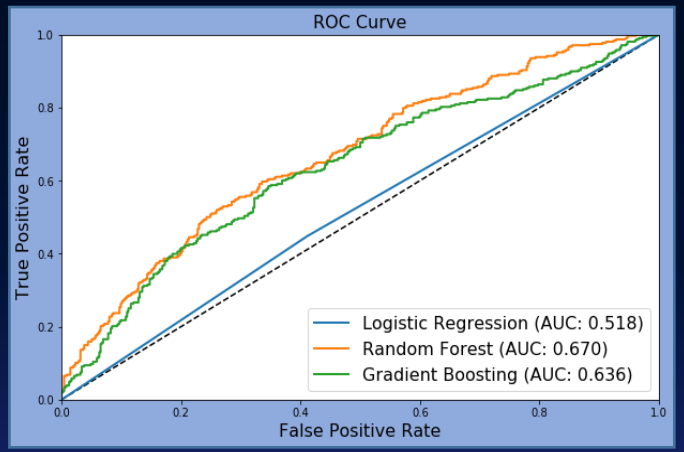
Tanzania

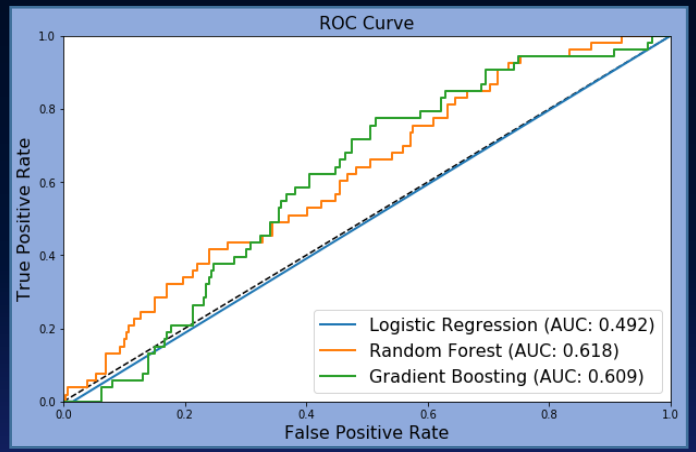
FCS





rCSI

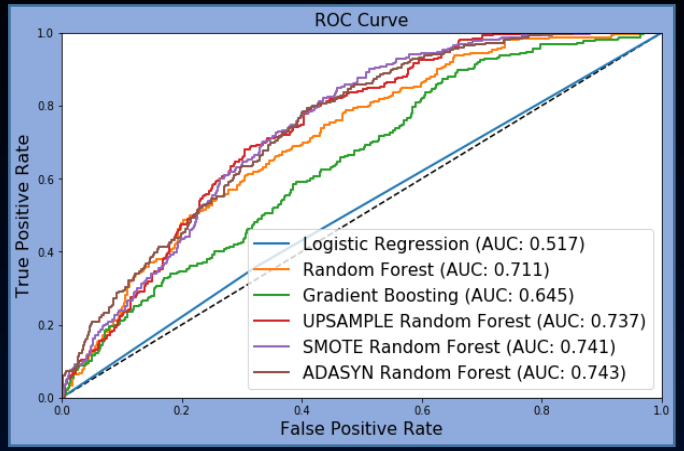


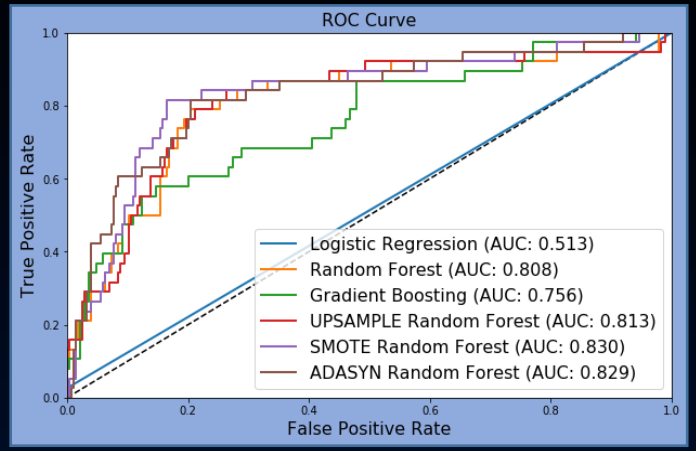


Malawi

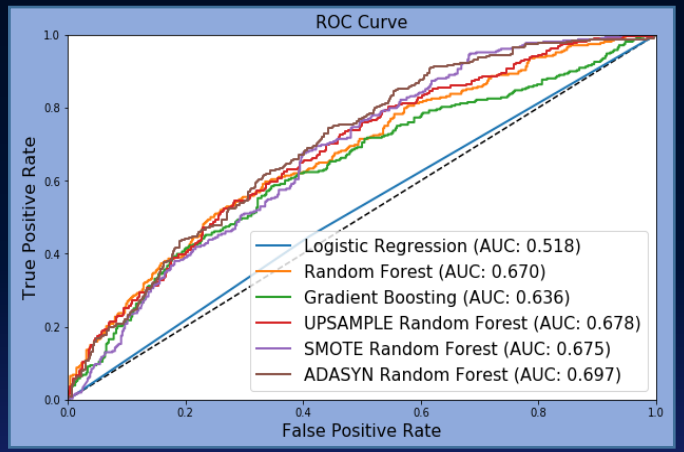
Tanzania

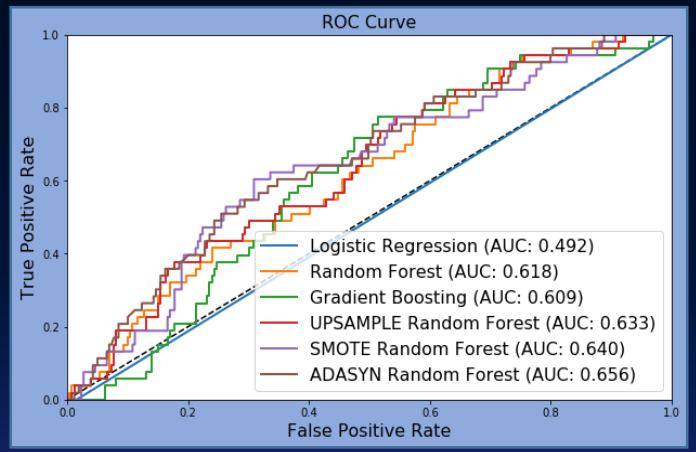
FCS





rCSI





In table format… Baseline vs ML algorithms at 70% Recall (year split)

Country	Food Security Measure	Precision Baseline vs *ML* *(worst to best)*	Precision Baseline vs *ML* with oversampling	Accuracy Baseline vs *ML*	Accuracy Baseline vs *ML* with oversampling
Malawi 2010/11, 2013 to predict 2015/16	FCS	0.32 *0.39-0.46*	0.32 *0.48-0.50*	0.32 *0.55-0.64*	0.32 *0.66-0.68*
	rCSI	0.39 *0.48*	0.39 *0.50-0.52*	0.39 *0.58*	0.39 *0.61-0.63*
Tanzania 2010/11, 2012/13 to predict 2014/15	FCS	0.12 *0.19-0.35*	0.12 *0.36-0.40*	0.12 *0.15-0.19*	0.12 *0.81-0.84*
	rCSI	0.17 *0.20-0.24*	0.17 *0.22*	0.17 *0.21*	0.17 *0.52-0.53*
Uganda 2010/11 to predict 2012	FCS	0.26 *0.26-0.27*	0.26 *0.26-0.29*	0.49 *0.22-0.24*	0.49 *0.48-0.55*

In table format… Baseline vs ML algorithms at 70% Recall (year split)

Country	Food Security Measure	Precision Baseline vs *ML (worst to best)*	Precision Baseline vs *ML* with oversampling	Accuracy Baseline vs *ML*	Accuracy Baseline vs *ML* with oversampling
Malawi 2010/11, 2013 to predict 2015/16	FCS	0.32 *0.39-0.46*	0.32 *0.48-0.50*	**0.32** ***0.55-0.64***	**0.32** ***0.66-0.68***
	rCSI	0.39 *0.48*	0.39 *0.50-0.52*	**0.39** ***0.58***	**0.39** ***0.61-0.63***
Tanzania 2010/11, 2012/13 to predict 2014/15	FCS	0.12 *0.19-0.35*	0.12 *0.36-0.40*	**0.12** ***0.15-0.19***	**0.12** ***0.81-0.84***
	rCSI	0.17 *0.20-0.24*	0.17 *0.22*	**0.17** ***0.21***	**0.17** ***0.52-0.53***
Uganda 2010/11 to predict 2012	FCS	0.26 *0.26-0.27*	0.26 *0.26-0.29*	**0.49** ***0.22-0.24***	**0.49** ***0.48-0.55***

ML is more accurate for given recall rate; oversampling seems to help

Baseline vs ML algorithms at **90% Recall**

Country	Food Security Measure	Precision Baseline vs *ML (worst to best)*	Precision Baseline vs ML with oversampling	Accuracy Baseline vs ML	Accuracy Baseline vs ML with oversampling
Malawi 2010/11, 2013 to predict 2015/16	FCS	0.32 *0.39-0.40*	0.32 *0.48-0.50*	0.32 0.52-0.54	0.32 *0.57-0.60*
	rCSI	0.39 *0.40-0.43*	0.39 *0.50-0.52*	0.39 0.44-0.50	0.39 *0.51-0.59*
Tanzania 2010/11, 2012/13 to predict 2014/15	FCS	0.12 *0.15-0.19*	0.12 *0.36-0.40*	0.12 0.60-0.80	0.12 *0.46-0.56*
	rCSI	0.17 *0.21*	0.17 *0.22*	0.17 *0.39-0.40*	0.17 *0.34-0.38*
Uganda 2010/11 to predict 2012	FCS	0.26 *0.22-0.24*	0.26 *0.26-0.29*	0.49 *0.26-0.33*	0.49 *0.34-0.40*

Again, oversampling seems to help. In some cases, higher recall threshold not too costly

Peek inside the black box…

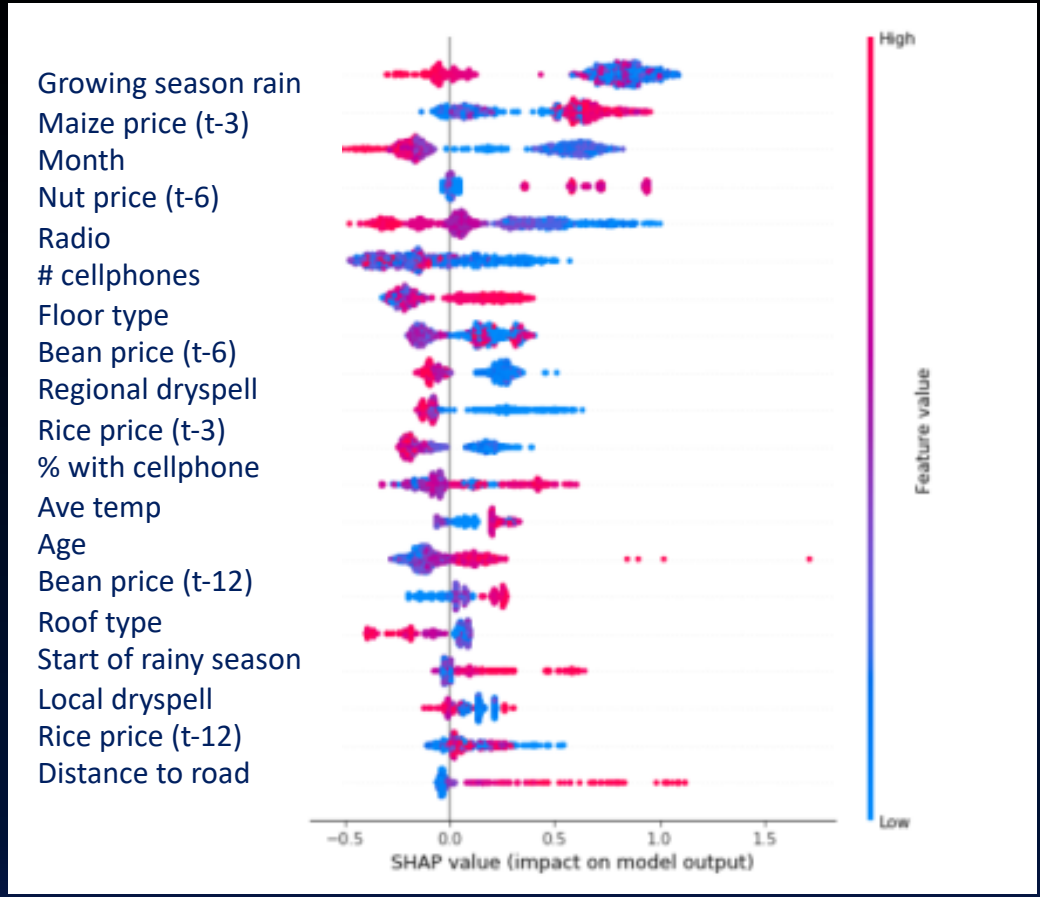
Regress output against IPC Zone and month x year fixed effects: explain < ½ variation

Shapley Values

Check that they make sense

(weather, prices and assets all important

Compare oversampled models to regular models

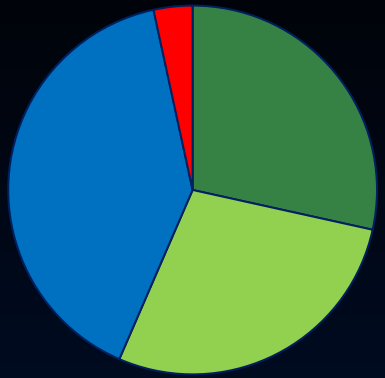


Gradient Boosted
Malawi FCS

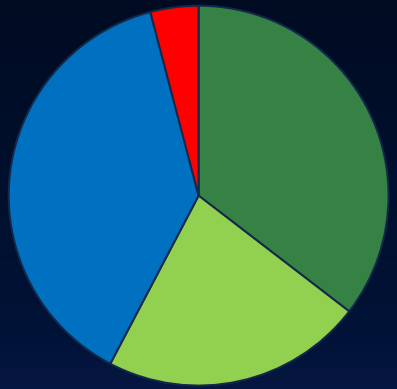
Error Analysis
(ADASYN Random Forest 90% Recall)

Malawi

FCS

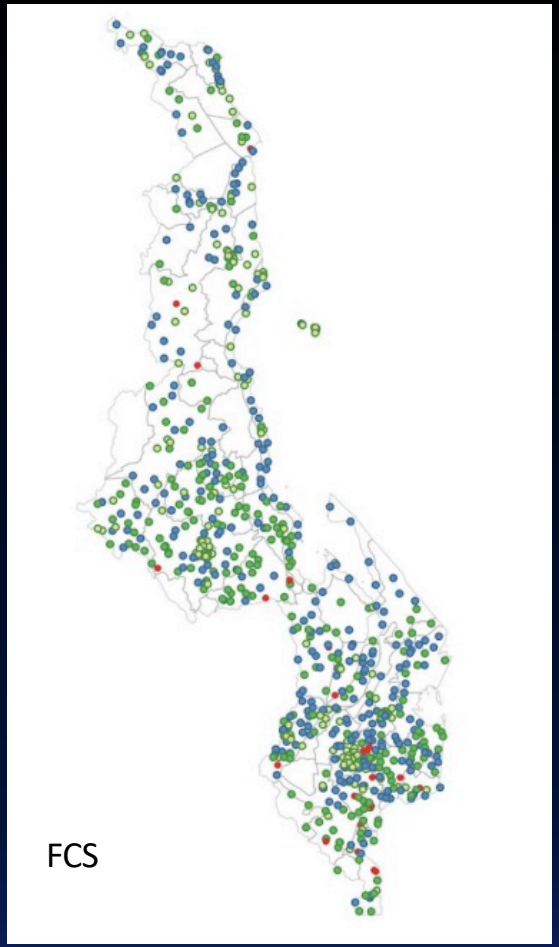


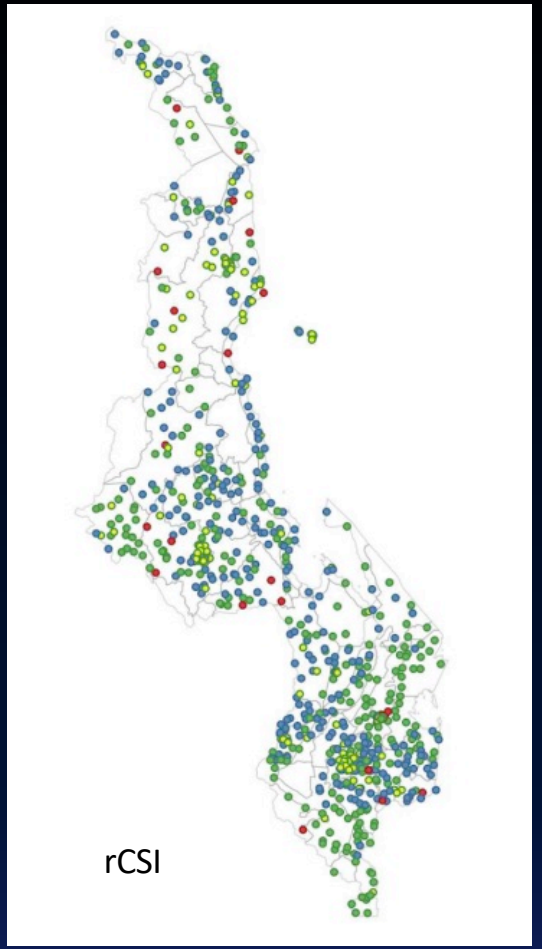
rCSI



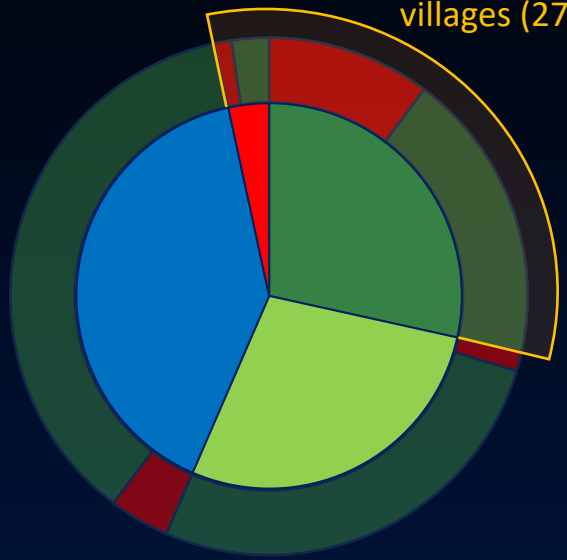
True Positive True Negative

False Positive False Negative

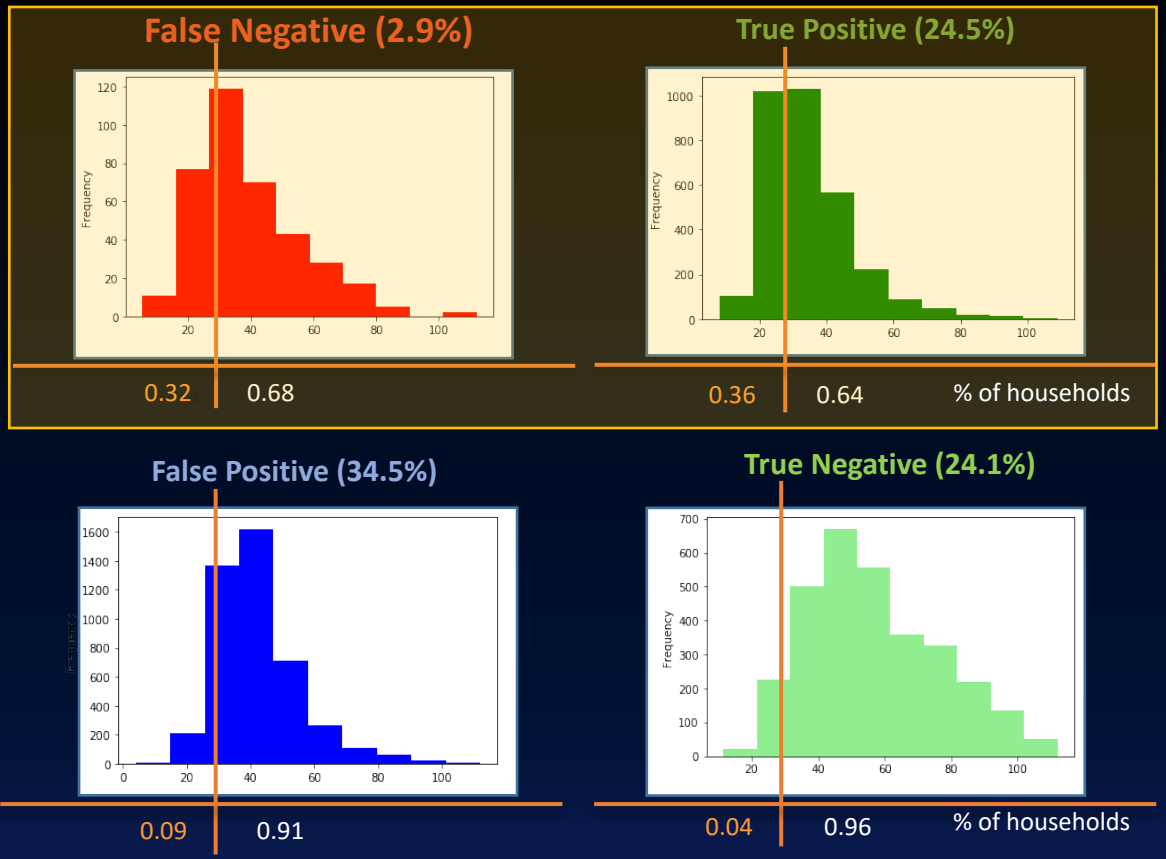




Malawi FCS by Household



Inner circle: villages
Outer ring: households



Lentz, Michelson, Baylis and Zhou

Summary of Results

- Machine learning models substantially improve over logit
- Oversampling methods improve both recall and accuracy
- Prices, weather and assets measures all contribute to predictions in sensible ways
- Error analysis suggest that households in mis-classified villages are closer to the food security cut-off
- We do not find systematic differences in precision over time or location* (more to do here)

…lots of caveats…

- At high recall (90%), our accuracy is not great (40-60%)
- Would not have picked up COVID-related food insecurity (other than through price movements)
- Not tested in a conflict setting
- Although picks up some changes over time, most variation is driven over space

Conclusions: maybe?

1. Proof of Concept

Machine learning and data techniques can improve food insecurity forecasts

2. Tune models to use

Are the outputs being used to target more information gathering? If so – how costly is information gathering?

Are the outputs being used to trigger or distribute aid?