Using Machine Learning to Improve Treatment Targeting in Farmer Training

Dr. Joyce Lin, Sarah Ellwein, Giovani Thai

Department of Mathematics, California Polytechnic State University, San Luis Obispo

Introduction

Agricultural sustainability impacts smallholder farmers in developing countries with limited capacity to adapt [1]. The IFC/World Bank aims to provide climate-smart training for farmers to improve livelihood and agronomic productivity [2]. Machine learning (ML) can serve to efficiently distribute resources to target farmers for training.

We implement, evaluate, and compare machine learning algorithms—Random Forest, AdaBoost, XGBoost, Explainable Boosting Machine (EBM)—to determine farmer eligibility for financial management training based on food security.

Dataset Description

Interviews were conducted in the northwest Cambodian provinces of Battambang and Banteay Meanchey (Figure 1). Survey information was collected from smallholder farmers by IFC Advisory Services at the beginning (2018) and end (2022) of the intervention.
This was stored in **baseline** (2018) and **endline** (2022) datasets, with individual responses categorized below (Table 1):

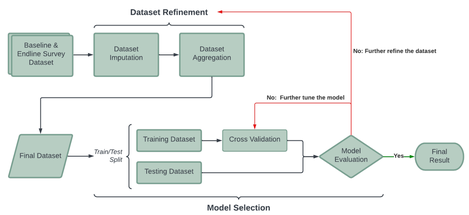
Fig. 1: Map of Cambodia by Surveyed Provinces.

Category	Examples
Household Demographics	Marital Status, Household Size, Years of Education
SWIFT	Rooms in House; Own fuel, vehicle, TV [3].
FIES	Lack of healthy/nutritious food in past 12 months, skipping meals [4].
FCS	Diversity of diet in last 7 days: fruits, vegetables, proteins, dairy [5].
Rice Production	Yield, sales, production costs last harvest
Agri Input Usage	Irrigation, fertilizer, pesticides used; livestock
Access to Training	Trained by IFC on land prep, harvesting, **financial management**

Tab. 1: Categories of column descriptions

Methodology

1. *Dataset Refinement*: Clean datasets, add features and merge baseline with endline
2. *Model Selection*: Train and evaluate each model tested. Rollback to prior steps depending on results.



Dataset Refinement

Success Metric: Food security is positively associated with agronomic productivity [6]. Therefore, improvements in **FIES** and **FCS** scores can serve as our proxy success metric. We added a feature to indicate if their FIES score **improved** from aggregated baseline after training.

Merging: Without a unique ID to directly merge the datasets, we instead pair **average baseline** district statistics with **individual endline** farmers (See Figure 2).

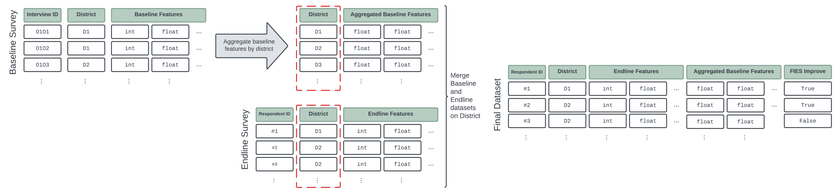


Fig. 2: Merging Process: Baseline dataset aggregated and merged with individual records in endline dataset.

Cleaned Data: After aggregation and cleaning, subset by farmers who received training for **financial management**, resulting in 102 records with 36 features.

Model Selection

Due to class imbalance, we applied SMOTE [7] on the training set and implemented a both a stratified 5-fold cross validation (CV) and stratified train/test split for each model (see Figure 3).

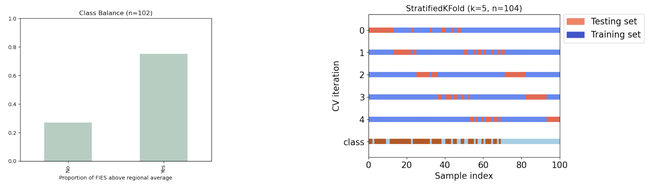


Fig. 3: (Left) Original frequency of labels based on FIES improvement. (Right) Stratified K-Fold CV on training set after SMOTE

Ensemble models is beneficial to reduce variance and overfitting. We implemented four ensemble models:

- **Random Forest**: ensemble learning algorithm that combines multiple decision trees to make predictions.
- **Adaptive Boosting (AdaBoost)**: boosting algorithm that iteratively combines weak learners into a strong learner. It assigns weights to training instances, focusing on those that are difficult to classify, and adjusts subsequent models to give more weight to misclassified samples [8].
- **XGBoost**: uses a gradient descent framework to sequentially build an ensemble of weak decision tree models, aiming to minimize the overall prediction error [9].
- **Explainable Boostin Machine (EBM)**: combines the ideas of boosting and generalized additive models to create a model that provides human-interpretable explanations for its predictions [10].

The following steps were taken to implement the models (see Results/Comparisons section for further elaboration on evaluation):

1. Apply 70:30 stratified train/test split, then apply SMOTE on training set.
2. For each algorithm:
   (a) Implement GridSearchCV [11] to hyperparameter tune each algorithm, in addition to using stratified 5-fold CV per iteration
   (b) Train the algorithm using the 70% training set and evaluate using the 30% testing set

Results/Comparisons

The World Bank IFC desires a model that is conservative in selecting candidates for training to save resources; therefore, we optimize and evaluate all models based on $F_{\frac{1}{2}}$ score, as it lends more weight to precision. Final results, in addition to training set evaluation, is shown in Table 2.

Model	Evaluation	Accuracy	$F_{\frac{1}{2}}$ Score
Random Forest	Training	100%	1.00
	Testing	64.5%	0.76
AdaBoost	Training	88%	0.88
	Testing	64.5%	0.76
XGBoost	Training	100%	1.00
	Testing	74%	0.84
EBM	Training	100%	1.00
	Testing	61.3%	0.72

Tab. 2: Evaluation results of training and testing datasets per model.

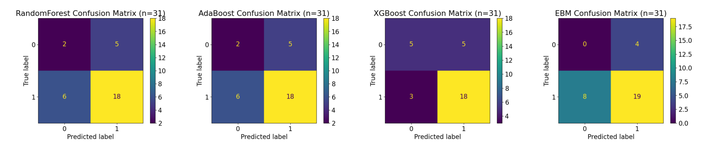


Fig. 4: Confusion matrices for each model on final test set.

XGBoost performed the best with a $F_{\frac{1}{2}}$ score of 0.84 and a relatively balanced confusion matrix (see Figure 4).
Therefore, we conclude that XGBoost is the best predictive model.

Future Work and References

Aggregation: Improve aggregation techniques by using **Cambodian census data** for more accurate 1-to-1 merging.

Trainings: Investigating more CSA trainings: **irrigation, fertilization, pesticide use**, etc. Develop a **multi-classification** model to see if farmers should be targeted for multiple trainings.

Remote Sensing: Incorporate satellite data, such as **crop vegetation index** and **soil quality**, to compare agronomic productivity with arable landscape.

COVID-19: How might we consider the effect of **COVID-19** on productivity in wet season 2021?

References: Refer to the QR Code to see our works cited.

Acknowledgements: Finally, we would like to thank the **BEACoN Research Mentoring Program** for funding our work and the **IFC/World Bank** for supplying us with data and supplementary materials.