Using machine learning models to analyze standardized household survey data collected in 21 countries.

Authors: Ana Boeriu, Christina De Cesaris, Mary-Francis LaPorte

I. Abstract

RHoMIS, a massive standardized survey effort in global agriculture, has allowed scientists to consider the association between household factors and poverty/farm productivity. Before now, predictive methods have never been used to study a 21-country survey of household farm productivity. In this study, we fit parametric (linear regression, ridge regression, LASSO) and non-parametric (Random Forest) machine learning models to analyze indicator variables. We find that geography are good predictors of Market Orientation and Percent Poverty Index, but the top specific predictors change in each model. Furthermore Random Forest is the best model overall. These findings open the door to the applications of predictive testing on agricultural survey data.

II. Introduction

RHoMIS survey procedures allow for international agriculture researchers to apply standardized methodologies to generate a globally comparable picture of household and farm demographics. To this date, machine learning methodologies have never been fit to predict indicator values based on survey data. Applying predictive models may allow scientists to determine the cause of the trends identified through the machine learning results. Additionally, once a predictive model is deemed properly fit and versatile for RHoMIS datasets, machine learning could be used to aid in farm consulting.

The data used in this study was previously published in an article entitled “The Rural Household Multiple Indicator Survey (RHoMIS) data of 13,310 farm households in 21 countries”, in *Scientific Data* (a Nature publication) in 2019 (doi: https://doi.org/10.7910/DVN/9M6EHS) The data is all publicly accessible through the Harvard Dataverse. In the experiment they conducted, the researchers surveyed over 13,000 families in 21 countries using a standardized survey approach. By standardizing both the questions, possible responses, and time to survey, they assert that the data reported by each household is advantageously reliable, compared to other surveying methods. They collected dozens of variables about each family, and calculated around 50 indicators based on the collected information. In this study, we utilize a selection of the raw variables and indicators (that are not confounding with the response variables of PPI and MO). For a full list of the variables used in our study, see the clean.csv dataset under the data directory in the project github repository. For a full description of all variables and their meaning, see “Raw data code book.tab” and “Explanation_of_Calculations_and_Outputs.tab” from the Harvard Dataverse public data (doi: https://doi.org/10.7910/DVN/9M6EHS). Any variables significant in our study will be explained in the report.

The main research question for this project is: “how will different machine learning models perform in predicting the Poverty Probability Index (PPI) and Market Orientation (MO) based on other survey factors”. Furthermore, we aim to identify the most important coefficients in predicting PPI and MO. Poverty Probability Index is a widely-used indicator to determine if a household lives in what is considered poverty. The term PPI is standardized in meaning across disciplines because it is actually a registered trademark (Poverty Probability Index®) of the PPI alliance. The indicator takes the answers

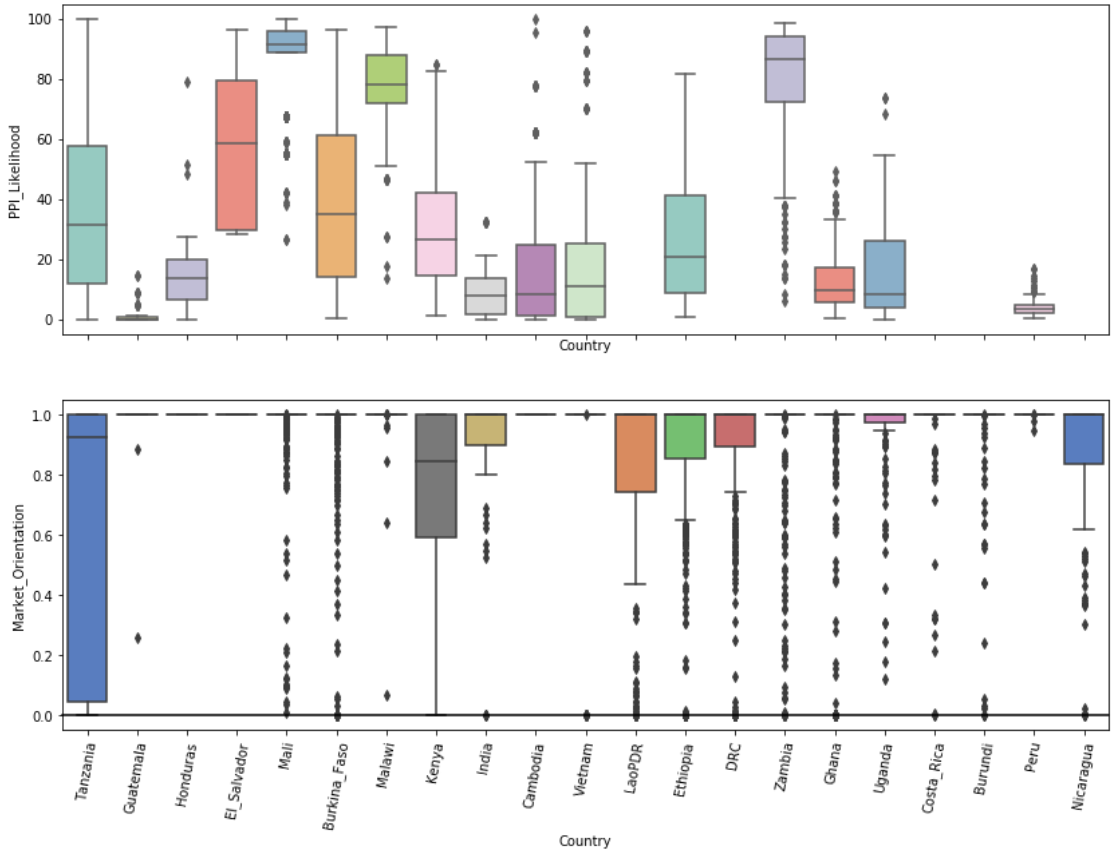
to a list of pre-determined questions that are known to be highly correlated to living in poverty, and thereby calculates the probability that the family being surveyed is in poverty based on the contents of their response. In this study, we did not calculate the PPI-- it was already determined by the authors and collectors of the original experiment and dataset. Market orientation is defined as the percentage of an agricultural good that has been produced, will be sold at the market. MO can be used to describe both crops and livestock. The MO was also included in the original dataset.

III. Exploratory Data Analysis

The first steps for this project were exploring the variables and cleaning the data for analysis. First, the correlation between the numerical predictors was plotted to gain a better understanding of how different features were linearly related. The correlation plot combined with intuition from the data cleaning process led us to choose PPI_Likelihood and Market_Orientation as the predictors for our models.

We then produced a box plot of our response variables against the countries still present in the data after cleaning. The boxplot for PPI_Likelihood against countries indicated that there was high variability between countries. This variability was not present in the boxplot relating Market_Orientation to countries. This exploratory insight presented a potential set back, as it implied that PPI_Likelihood was country specific and therefore the models would primarily predict PPI_Likelihood using the country's name and not the other features as we intended.

Figure 1: Boxplots: countries vs responses (ppi, mo)



IV. **Data cleaning**: (see /laporpe/STA_208/data/more_data_cleaning.ipynb)

The first step for data cleaning was to choose the variables to include in the study using brief background information gained from the “Raw data code book” and “Explanation_of_Calculations_and_Outputs.tab”. The goal of this was to omit anything that was a) redundant with any of the response variables b) irrelevant to our study scientifically or, rarely, c) where the meaning of the values were unclear. We did this by hand, as a group, manually discussing each of the ~200 variables, and reducing them to the ~40 to be used in the final model. The result of this were two files (included on the github repo) entitled “RHoMIS_Full_Data.csv” and “data/RHoMIS_Indicators.csv”, which contained the raw variables and non-confounding indicators, respectively.

Upon further data exploration, we found a significant number of NA’s. Variables that had more than 30 percent of NAs were removed since imputing any amount greater than that would greatly skew our results. For all the numeric variables we imputed using the median since it is less affected by any skewness compared to the mean. A missingness category of na was created for categorical variables with NA values. We then removed all rows where our response variable was NA.

Furthermore, we also adjusted the Education column levels, as described in the next paragraph. All variable names were converted to lowercase, and the variable types (numeric, categorical) were verified and corrected if needed.

Head Education Level (or the highest level of education achieved by the head of the household) was recorded as being either “primary, secondary, post-secondary, literate, illiterate, or no school”. In some surveying contexts (i.e. groups of surveys administered by the same person), the survey administrator chose to record non-standard descriptions of the head education level. Examples of such non-standard descriptions were describing the religion of the parochial school but not the education level, recording the education level in the context of the language that was spoken at the school, or using locally-recognized literacy ratings without converting them to the standardized responses. As these non-standard descriptors are not clearly defined, it was impossible to map them to one of the standardized responses. As a result, the rows with non-standard head education levels were dropped. This did not affect the majority of rows for any country, and was primarily an issue in Burkina Faso and Mali.

Next, we manually added the continent corresponding to each country in the cleaning file. In downstream analyses, we used this coding to compare the results, including countries but not continents, with those containing continents but not countries.

A. Introduce potential data problems as well as state solutions

V. Model Methods

Given the complexity of the data, we decided to train both parametric and nonparametric models and evaluate feature importance accordingly. The optimally tuned linear parametric models--OLS, Ridge, and LASSO regression--tended to have lower mean squared error values than their nonparametric RandomForrestRegressor counterpart. The RandomForrest feature importances are based on the mean decrease in impurity which occurs when a tree splits on a specific feature. The

feature importance scores are not indicative of how a particular feature relates to the response variable, but rather its significance in determining the splits across all estimators in the RandomForrest. For the parametric models, feature importance was extracted from the values of the model coefficients which allowed us to investigate the relationship between the response variables and predictors.

The linear models were trained using the cleaned dataset and either the PPI_Likelihood or Market_Orientation variables as the response. It was often the case that the dominating features were those representing the different countries present in our dataset. For example, in the PPI_Likelihood models, the countries in Africa would overshadow the importance of other features. This issue was handled by replacing the country column with the continent to which each country belonged. Unsurprisingly, the continent Africa was found a significant predictor in the new model, but other features which were previously subdued emerged.

PERHAPS HERE WE PLACE results (top five cofficents) with countries, then with continents for one of the models as a visual example.

Separate country-specific RandomForrest models were run on subsets of the data to gain a better understanding of which features were the most influential in predicting the PPI_Likelihood for each country. Continent-specific models were also run for comparison. The RandomForestRegressor was chosen for the subsetted data because the RandomForest model performed best on the complete dataset.

VI. Cross Validation (perhaps move this until after results section)

Before we are able to fit the respective models we performed leave one out cross validation to obtain the optimal parameters based on the criterion of mean square prediction error. This allows us to optimize the performance of each model and prevent overfitting the data?????? All models are fit without intercept since we do not have baseline categories.

Ordinary Least Squares was performed on the cleaned dataset. The analysis was run first with just the numerical data, then including the categorical variables as dummy variables. The MSE values for OLS are reported in Table 5. Standardization was completed using a custom function that performed a z-score transformation. The results were reported using the statsmodels.api library. These results are included in the LinearModel.ipynb notebook.

A. Ridge Regression

Figure 3: Optimal Alpha for y=Market Orientation

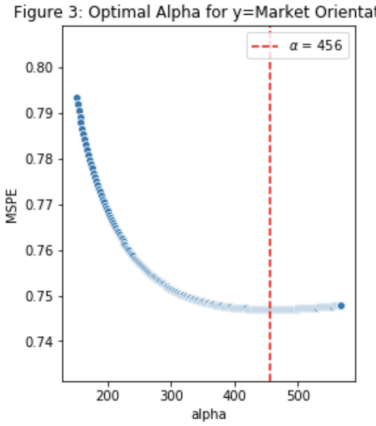
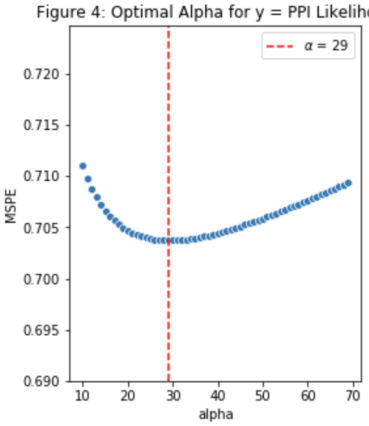


Figure 4: Optimal Alpha for y = PPI Likelihood



Ridge regression decreases model complexity and reduces the variance by using an additional regularization parameter ,α, that controls the size of our beta coefficients. As our α increases, the smaller our coefficients will be. From figures 3 and 4 we can see that the using Market orientation as our response requires a larger α which could underfit the model.

B. LASSO
C. Random Forest

VII. Results

The ordinary least squares model was calculated on numeric values from the dataset only. This is because the MSE was over 10 orders of magnitude smaller when fitting with just the numeric values. These findings were replicated by fitting a ridge regression with no ridges, confirming that this was not an error with data processing.

Ridge Regression Results

Ridge Regression did significantly better than ordinary least squares regression but not as well as LASSO or Random forest regression. As seen in Table in the appendix, Country, year, livestock holdings, crop and household types were all important predictors for market orientation. This makes sense since MO is affected economically by year and country. Similarly for PPI as expected, continent year, education level, and number of household members are all important predictors. Not surprisingly, people with a small edu level, who live in africa will tend to have a high PPI compared to others.

LASSO

The LASSO Regression model performed slightly better than the Ridge Regression model when predicting both MO and PPI. It produced similar results to the Ridge model regarding the predictors associated with the highest coefficient values. Many of the predictors deemed significant by the Ridge model concerning MO overlapped with those of the LASSO model, and those which did not overlap tended frequently belonged to the same category with the exception of Male Gender Control which was found important in the LASSO MO model but not the Ridge MO model. Regarding the PPI model, the LASSO regression predictors consisted of a mix of significant predictors from the RandomForrest model and the Ridge Model.

1. Random Forest

Tables to include in the Text:

a. Table with all the most important features of each model

b. Plots of feature importances

Random Forest Results
Country-Level Feature Importance

VIII. Discussion

The Ordinary Least Squares model had the worst MSE of all of the models. This is because the linear prediction is not sufficient for this dataset. The variables that were the most significant for OLS were not the same as the variables found as significant for the other three models. The importance of gender control (whether male or female) was the same. This is likely because the values were standardized, and female control is calculated as (1- percentage male control). To further understand these variables, a different processing procedure would be needed. Livestock production ended up being an important predictor of market orientation. This is likely due to the fact that The value of a particular type of cattle is correlated to its desirability in a market. Food availability and the contents of the diet were important predictors of PPI.

Ridge Regression did not fit as well as the Random Forest and LASSO.

A. LASSO and Random forest significance convo
B. Difference between LASSO using countries and continents as variables
C.

Some variables that were found as potentially significant (near the top 10) were likely confounding predictors. One explanatory examples is polygamy. Polygamy was a significant predictor of PPI, but this, as always, does not imply causation. Primarily in just a few countries, so could be unintentionally correlating with poverty in those nations. It is important to note as this is a cultural aspect of some of these households, so it is important to interpret these models without making assumptions about causality.

In this study, we treated the years as categorical variables. Conceptually similar to the geographic separation of countries being a categorical variable, the temporal separation of each year was treated as its own environment with myriad factors having the potential to change. One large

assumption that this treatment requires is that the survey answers are independent between years (to fulfil the independence assumption). This means that we are assuming that the families being surveyed in one year are distinct from those being surveyed the next. Alternatively, this could be true if the conditions were independent enough from year to year in one family, that the assumption could be met. We are unable to know if the families surveyed are consistent over the years because the survey data is de-identified. In summary: we assume sample independence.

Rare crops: the type of crop ended up being a major predictor in every case-- but the individual crops identified were not major global food crops. This suggests that the crops being significant were ones grown in very few households. This means that the significance may be in error, from the small sample sizes for these crops. Figure X in the appendix demonstrates the frequency of crop types on farms to depict this trend. In future analyses with this data, the crop frequency should be taken into account. One method for doing this would be to only include those which more than 20 farmers grew. Another factor to consider is the geographic distribution of crops. The crops might be a confounding predictor with country, so further analysis needs to be done.

To see how much PPI is affected by each predictor, we can look at the ridge regression model
Our ridge regression model has the form

$$Y_{PPI} = 0.02179X_{crop,count} + 0.3866X_{HH,members} - 0.00037X_{land\ owned} + \cdots + 0.6828X_{Africa} - 0.92944X_{central\ an}$$

Generally when we interpret categorical variables, we always compare the current factor level to the baseline factor level. However, because our model does not have baseline categorical levels, our interpretation will change slightly. In our model shown above we are interested in how much the PPI changes for those living in Africa versus those in Central America. We look at the difference between the two sub models for Africa and Central America. Thus, comparing Central America to Africa, the PPI would decrease by 1.6122 units on average, holding all other variables constant.
A general interpretation of a numeric variable is when X_1 increases by 1 unit, Y increases by β_1 units on average holding all other variables constant. Thus interpreting the variable land owned: When a person buys a hectare of land, the PPI decreases by 0.00037 units on average holding all other variables constant.

IX. Conclusion

A. Discuss results in relation to research question
B. Explain significance of results
C. Future directions
    1. Looking at “by year” interactions
D. next

Link to Github Repo: https://github.com/laporpe/STA_208

Appendix:

Table 1: Top 9 Coefficients (or Feature Importance Score for Random Forest) for Market Orientation for each model. Note: Country included, not continent. OLS was calculated on numeric values only.

	OLS		Ridge Regression		LASSO		Random Forest	
1	Livestock production	-2.21228e+13	Country, Tanzania	-0.4320	Country Tanzania	-0.7046	Value livestock produced and consumed	0.4296
2	Value of livestock	2.01358e+13	Country, Kenya	-0.2481	Country Kenya	-0.4281	Farm income	0.3309
3	Gender control male	1.56606e+13	Year, 2017	-0.2245	Year 2017	-0.3615	value_crop_consumed	0.1213
4	Gender control female	1.56606e+13	Country, Burkina Faso	0.1891	Livestock Holdings	-0.1435	Total income	0.1017
5	Livestock product sales	9.12637e+12	Livestock Holdings	-0.1817	Country Mali	0.1418	Value farm produce	0.0044
6	Value farm produce	9.06603e+11	Country, Mali	0.1458	Gender Male Control	-0.1359	Crop sales	0.0023
7	Crop sales	-6.91062e+11	Country, Laos	-0.1160	Country Ethiopia	-0.1331	Livestock product sales	0.0014
8	Total income	-2.06639e+11	Crop, Greengram	0.1073	Country Burkina Faso	0.1159	TVA (in USD)	0.0007
9	Off-farm income	1.32105e+11	Type of Household: Woman, single	0.0927	Crop Harvest na	0.0756	Value livestock production	0.0007
10	Farm income	-9.77692e+10	Year, 2016	0.0904	Crop Intercrop, na	0.0736	Value Crop produce	0.0006

Table 2: Top variable categories for Market Orientation (in order of appearance)

	OLS	Ridge Regression	LASSO	Random Forest
1	Livestock	Country	Country	Value of Livestock Consumed

	Production			
2	Gender Control	Year	Year	Income
3	Farm Production	Livestock Holdings	Livestock Holdings	Value of Crops Consumed
4	Crop sales	Crop	Gender Male Control	Crop Sales
5	Total income	Household Type	Crop	Livestock Sales

Table 3: Top 9 Coefficients (or Feature Importance Score for Random Forest) for PPI for each model. Note: Continent included, not country. OLS was calculated on numeric values only.

	OLS		**Ridge Regression**		**LASSO**		**Random Forest**	
1	Food availability	0.830114	Continent, Central america	-0.9359	Continent Africa	0.6607	Continent, Africa	0.0931
2	Household size	0.825778	Head Education Level, none	-0.8022	Year, 2015	0.4607	Number of Months Wild Food Consumption	0.0713
3	Total Value of daily activities	-0.741225	Continent, Africa	0.6756	Head Education Level, No school	0.3961	Livestock Holdings	0.0538
4	Household size (normalized)	-0.716008	Head Education level, no school	0.5045	Year 2017	0.3572	Household size male adult equiv	0.0519
5	Off-farm income	-0.150735	Year 2015	0.4944	Number of Months, Wild food consumption	0.124081	Year 2017	0.0449
6	Number of months food consuming wild food	0.1013	Year, 2018	-0.4888	Crop, millet	0.0871742	Food Self Sufficiency kCal, MAE, day	0.0435

7	Number of months food insecure	0.0973118	Crop, irish potato	-0.4335	Head Education Level, primary	0.0814309	Number of Household members	0.0382
8	Dietary diversity during a bad season	-0.068309	Number of household members	0.3781	Number of Household members	0.0690304	HDDS score, good season	0.0373
9	Crop count	0.0670102	Crop, teff	-0.3764	HouseholdType_couple	0.0597672	Year 2015	0.0325
10	Total income	-0.0583924	Crop, wheat	-0.334	HFIAS_status_severelyfi	0.0566652	HDDS score, bad season	0.0324

c. Table with MSE between models

Table 4: Top variable categories for PPI (in order of appearance)

	OLS	**Ridge Regression**	**LASSO**	**Random Forest**
1	Food availability	Continent	Continent	Continent
2	Household size	Head Education level	Year	Number of Months Wild Food was Consumed
3	Value of daily activities	Year	Head Education Level	Livestock Holdings
4	Off-farm income	Crop	Number of Months Wild Food was Consumed	Household Size Caloric Intake
5	Food insecurity/ Dietary diversity	Number of household members	Crop	Year

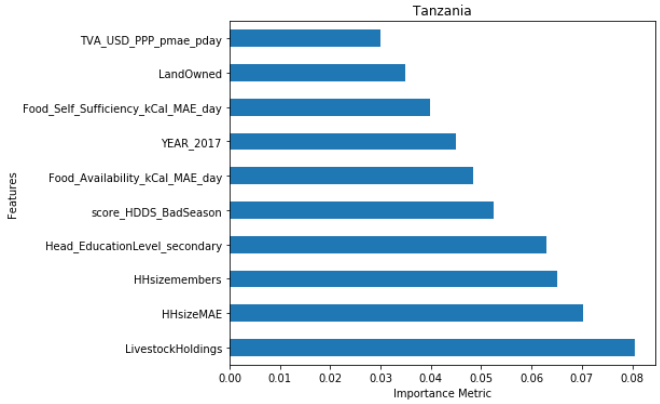
Table 5: Model evaluation criteria. Mean Squared Error. PPI uses only continent and not country, MO uses country but not continent.

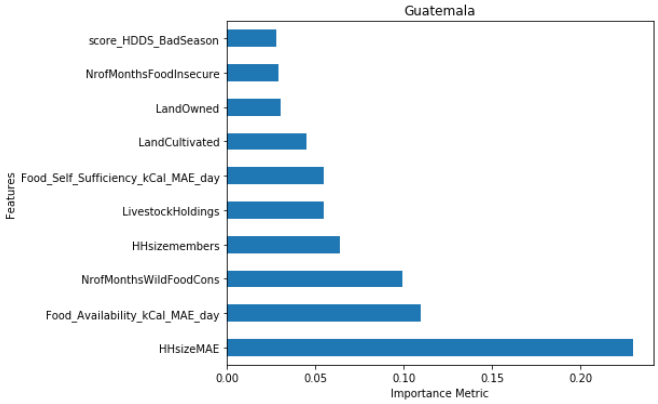
	OLS		Ridge Regression		LASSO		Random Forest	
	PPI	**MO**	**PPI**	**MO**	**PPI**	**MO**	**PPI**	**MO**
MSE	0.9815	3.2187	0.8451	0.7104	0.6969	0.6696	0.4778	0.0099
alpha			29	456	0.01	0.01		

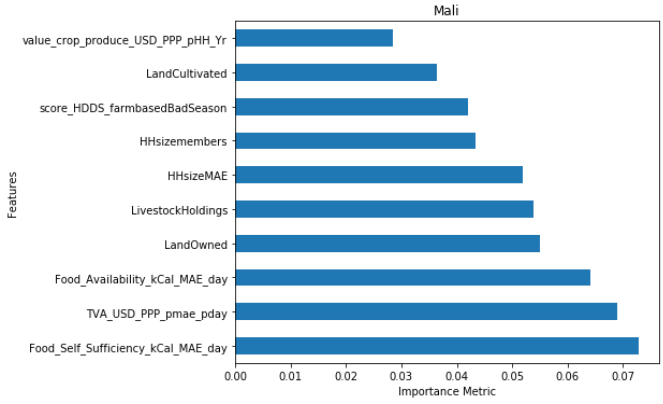
Graph checklist:
Has x and y axis clearly labeled with units if necessary
Tile or caption
(I'm wondering if pyplot has a viridis color palette equivalent to make choosing color contrasts easier, will look into this)

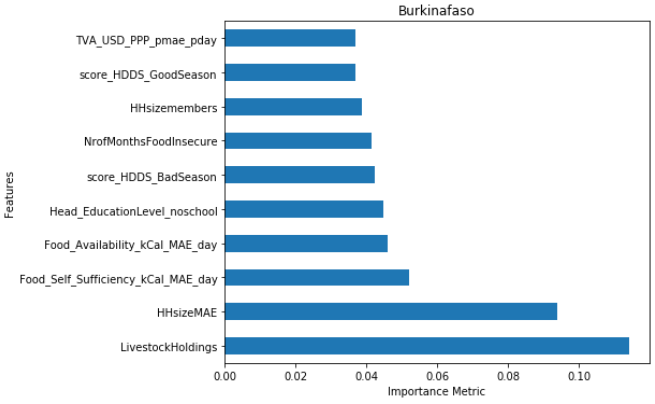
Focus on interpreting the results and not so much explaining the methods covered in class.

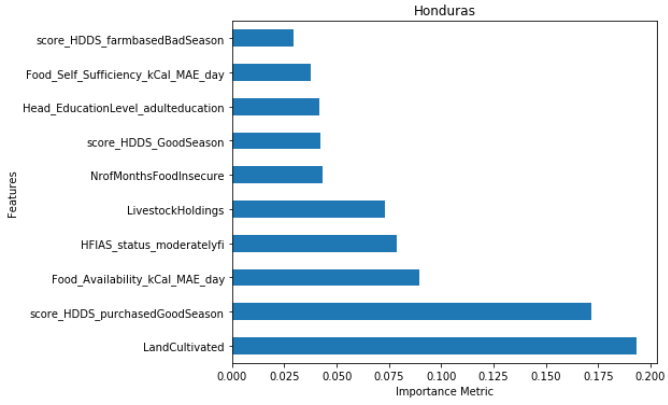
Figure X: PPI

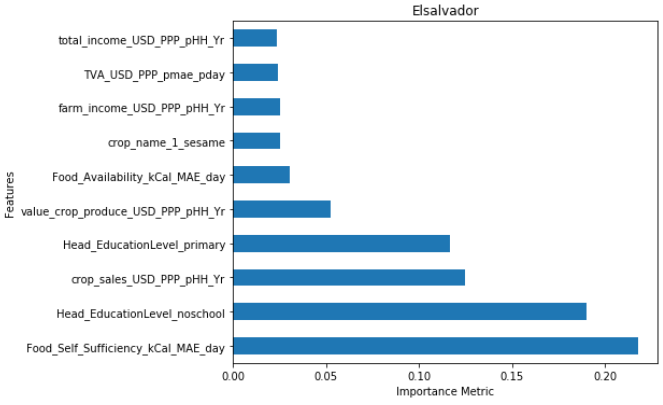

Tanzania
TVA_USD_PPP_pmae_pday
LandOwned
Food_Self_Sufficiency_kCal_MAE_day
YEAR_2017
Food_Availability_kCal_MAE_day
score_HDDS_BadSeason
Head_EducationLevel_secondary
HHsizemembers
HHsizeMAE
LivestockHoldings
Features
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
Importance Metric

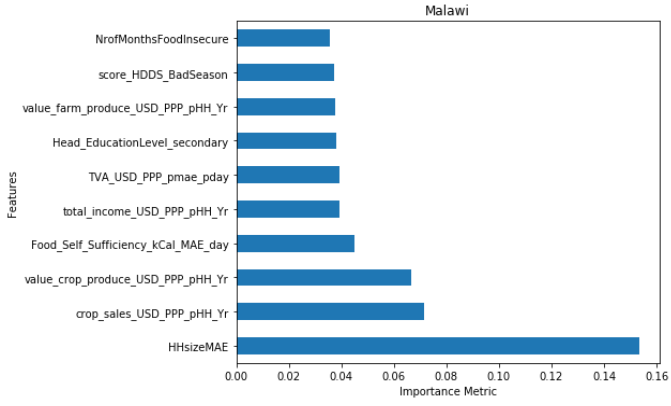

Guatemala
score_HDDS_BadSeason
NrofMonthsFoodInsecure
LandOwned
LandCultivated
Food_Self_Sufficiency_kCal_MAE_day
LivestockHoldings
HHsizemembers
NrofMonthsWildFoodCons
Food_Availability_kCal_MAE_day
HHsizeMAE
Features
0.00 0.05 0.10 0.15 0.20
Importance Metric

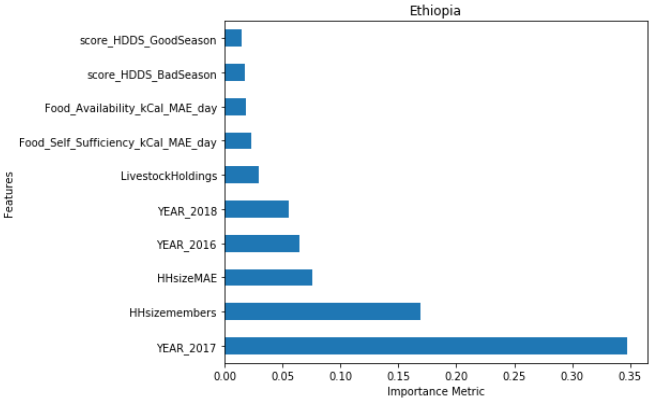

Mali
value_crop_produce_USD_PPP_pHH_Yr
LandCultivated
score_HDDS_farmbasedBadSeason
HHsizemembers
HHsizeMAE
LivestockHoldings
LandOwned
Food_Availability_kCal_MAE_day
TVA_USD_PPP_pmae_pday
Food_Self_Sufficiency_kCal_MAE_day
Features
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07
Importance Metric

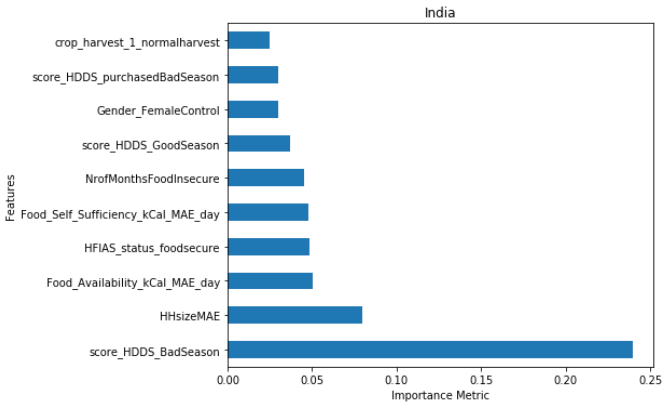

Burkinafaso
TVA_USD_PPP_pmae_pday
score_HDDS_GoodSeason
HHsizemembers
NrofMonthsFoodInsecure
score_HDDS_BadSeason
Head_EducationLevel_noschool
Food_Availability_kCal_MAE_day
Food_Self_Sufficiency_kCal_MAE_day
HHsizeMAE
LivestockHoldings
Features
0.00 0.02 0.04 0.06 0.08 0.10
Importance Metric

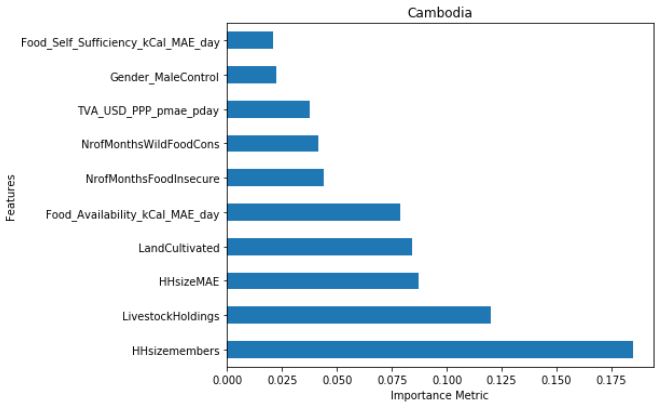

Honduras
score_HDDS_farmbasedBadSeason
Food_Self_Sufficiency_kCal_MAE_day
Head_EducationLevel_adulteducation
score_HDDS_GoodSeason
NrofMonthsFoodInsecure
LivestockHoldings
HFIAS_status_moderatelyfi
Food_Availability_kCal_MAE_day
score_HDDS_purchasedGoodSeason
LandCultivated
Features
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Importance Metric

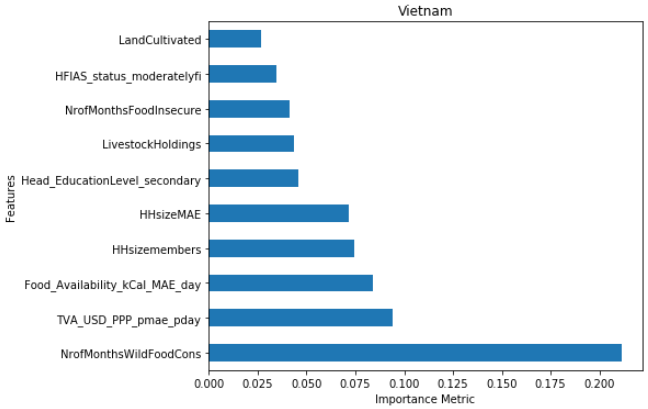

Elsalvador
total_income_USD_PPP_pHH_Yr
TVA_USD_PPP_pmae_pday
farm_income_USD_PPP_pHH_Yr
crop_name_1_sesame
Food_Availability_kCal_MAE_day
value_crop_produce_USD_PPP_pHH_Yr
Head_EducationLevel_primary
crop_sales_USD_PPP_pHH_Yr
Head_EducationLevel_noschool
Food_Self_Sufficiency_kCal_MAE_day
Features
0.00 0.05 0.10 0.15 0.20
Importance Metric

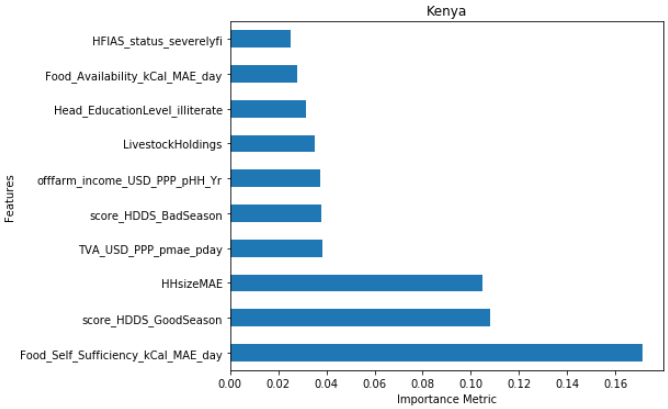

Malawi
NrofMonthsFoodInsecure
score_HDDS_BadSeason
value_farm_produce_USD_PPP_pHH_Yr
Head_EducationLevel_secondary
TVA_USD_PPP_pmae_pday
total_income_USD_PPP_pHH_Yr
Food_Self_Sufficiency_kCal_MAE_day
value_crop_produce_USD_PPP_pHH_Yr
crop_sales_USD_PPP_pHH_Yr
HHsizeMAE
Features
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
Importance Metric


Ethiopia
score_HDDS_GoodSeason
score_HDDS_BadSeason
Food_Availability_kCal_MAE_day
Food_Self_Sufficiency_kCal_MAE_day
LivestockHoldings
YEAR_2018
YEAR_2016
HHsizeMAE
HHsizemembers
YEAR_2017
Features
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
Importance Metric


India
crop_harvest_1_normalharvest
score_HDDS_purchasedBadSeason
Gender_FemaleControl
score_HDDS_GoodSeason
NrofMonthsFoodInsecure
Food_Self_Sufficiency_kCal_MAE_day
HFIAS_status_foodsecure
Food_Availability_kCal_MAE_day
HHsizeMAE
score_HDDS_BadSeason
Features
0.00
0.05
0.10
0.15
0.20
0.25
Importance Metric


Cambodia
Food_Self_Sufficiency_kCal_MAE_day
Gender_MaleControl
TVA_USD_PPP_pmae_pday
NrofMonthsWildFoodCons
NrofMonthsFoodInsecure
Food_Availability_kCal_MAE_day
LandCultivated
HHsizeMAE
LivestockHoldings
HHsizemembers
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance Metric


Vietnam
LandCultivated
HFIAS_status_moderatelyfi
NrofMonthsFoodInsecure
LivestockHoldings
Head_EducationLevel_secondary
HHsizeMAE
HHsizemembers
Food_Availability_kCal_MAE_day
TVA_USD_PPP_pmae_pday
NrofMonthsWildFoodCons
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance Metric


Kenya
HFIAS_status_severelyfi
Food_Availability_kCal_MAE_day
Head_EducationLevel_illiterate
LivestockHoldings
offfarm_income_USD_PPP_pHH_Yr
score_HDDS_BadSeason
TVA_USD_PPP_pmae_pday
HHsizeMAE
score_HDDS_GoodSeason
Food_Self_Sufficiency_kCal_MAE_day
Features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance Metric

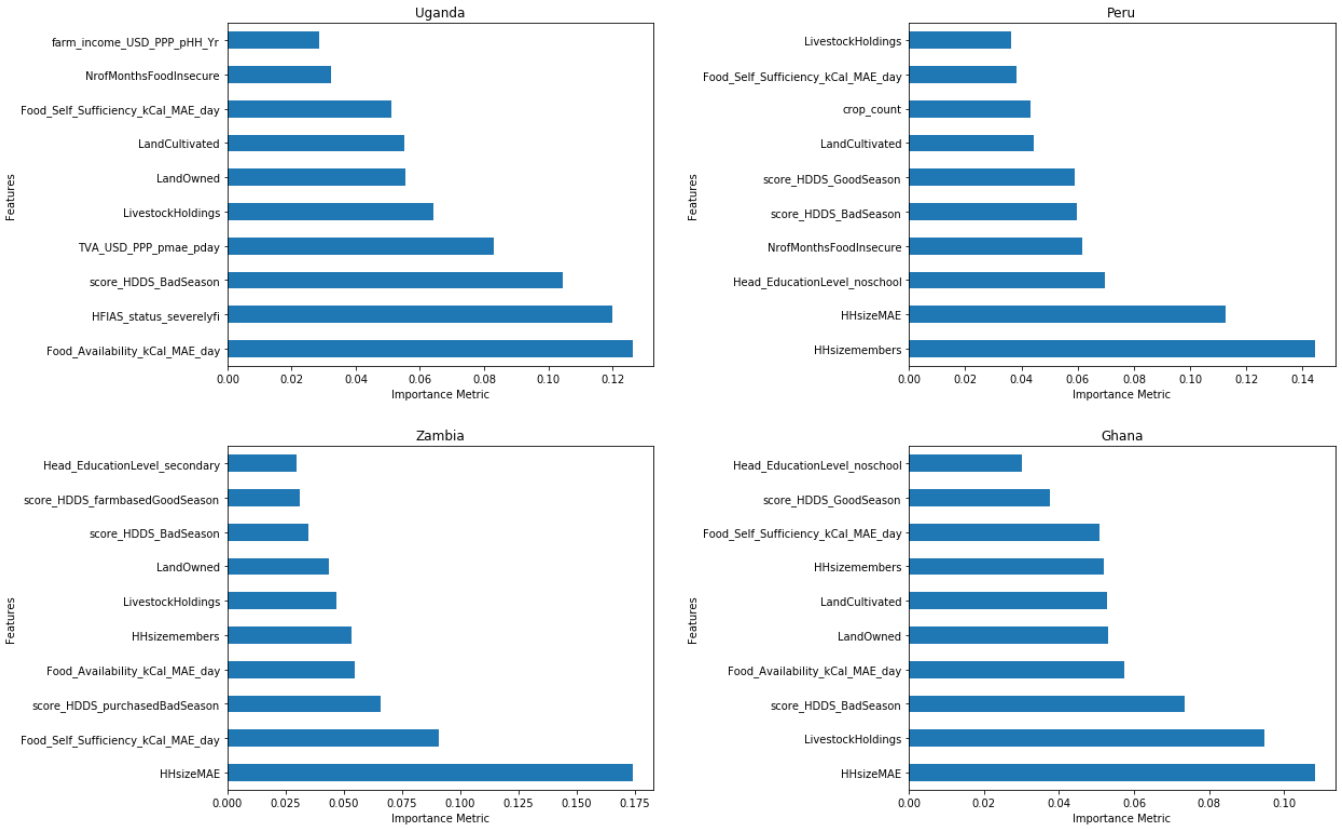

Uganda
farm_income_USD_PPP_pHH_Yr
NrofMonthsFoodInsecure
Food_Self_Sufficiency_kCal_MAE_day
LandCultivated
LandOwned
LivestockHoldings
TVA_USD_PPP_pmae_pday
score_HDDS_BadSeason
HFIAS_status_severelyfi
Food_Availability_kCal_MAE_day
Features
0.00 0.02 0.04 0.06 0.08 0.10 0.12
Importance Metric
Peru
LivestockHoldings
Food_Self_Sufficiency_kCal_MAE_day
crop_count
LandCultivated
score_HDDS_GoodSeason
score_HDDS_BadSeason
NrofMonthsFoodInsecure
Head_EducationLevel_noschool
HHsizeMAE
HHsizemembers
Features
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Importance Metric
Zambia
Head_EducationLevel_secondary
score_HDDS_farmbasedGoodSeason
score_HDDS_BadSeason
LandOwned
LivestockHoldings
HHsizemembers
Food_Availability_kCal_MAE_day
score_HDDS_purchasedBadSeason
Food_Self_Sufficiency_kCal_MAE_day
HHsizeMAE
Features
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
Importance Metric
Ghana
Head_EducationLevel_noschool
score_HDDS_GoodSeason
Food_Self_Sufficiency_kCal_MAE_day
HHsizemembers
LandCultivated
LandOwned
Food_Availability_kCal_MAE_day
score_HDDS_BadSeason
LivestockHoldings
HHsizeMAE
Features
0.00 0.02 0.04 0.06 0.08 0.10
Importance Metric

Figure: X+1

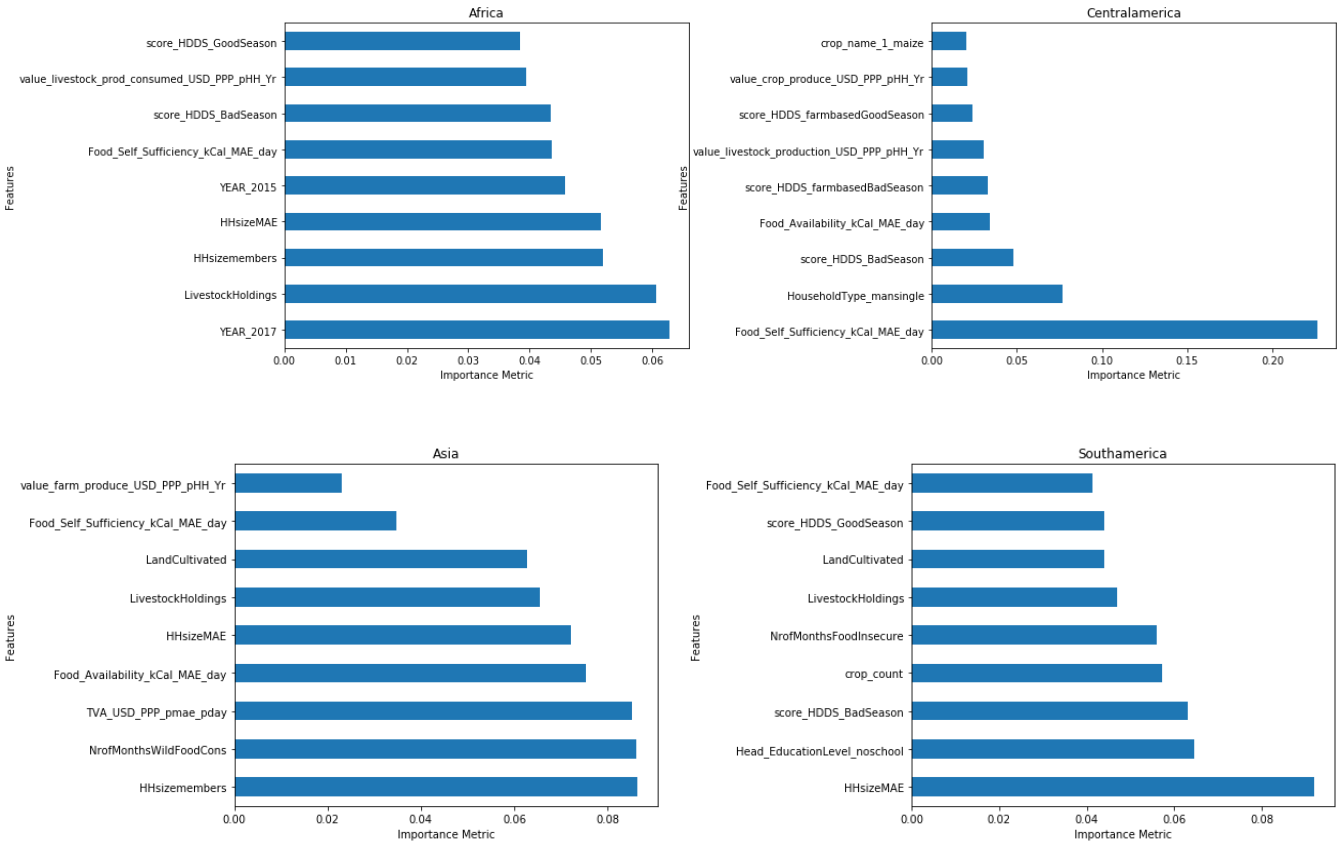

Africa
score_HDDS_GoodSeason
value_livestock_prod_consumed_USD_PPP_pHH_Yr
score_HDDS_BadSeason
Food_Self_Sufficiency_kCal_MAE_day
YEAR_2015
HHsizeMAE
HHsizemembers
LivestockHoldings
YEAR_2017
Features
0.00 0.01 0.02 0.03 0.04 0.05 0.06
Importance Metric
Centralamerica
crop_name_1_maize
value_crop_produce_USD_PPP_pHH_Yr
score_HDDS_farmbasedGoodSeason
value_livestock_production_USD_PPP_pHH_Yr
score_HDDS_farmbasedBadSeason
Food_Availability_kCal_MAE_day
score_HDDS_BadSeason
HouseholdType_mansingle
Food_Self_Sufficiency_kCal_MAE_day
Features
0.00 0.05 0.10 0.15 0.20
Importance Metric
Asia
value_farm_produce_USD_PPP_pHH_Yr
Food_Self_Sufficiency_kCal_MAE_day
LandCultivated
LivestockHoldings
HHsizeMAE
Food_Availability_kCal_MAE_day
TVA_USD_PPP_pmae_pday
NrofMonthsWildFoodCons
HHsizemembers
Features
0.00 0.02 0.04 0.06 0.08
Importance Metric
Southamerica
Food_Self_Sufficiency_kCal_MAE_day
score_HDDS_GoodSeason
LandCultivated
LivestockHoldings
NrofMonthsFoodInsecure
crop_count
score_HDDS_BadSeason
Head_EducationLevel_noschool
HHsizeMAE
Features
0.00 0.02 0.04 0.06 0.08
Importance Metric

Figure X+2. The frequencies of crop (crop_name_1) for all crops grown in more than 20 households. Any crops grown in less than 20 households are omitted (not to be confused with "other").