



UN Youth Hackathon 2022

Presentation by Mayo Team

TABLE OF CONTENTS

01

INTRODUCTION

Mayo Team
Background and problem
statement
Scope and Objective

02

KNOWING THE DATA

Data collection
Preprocessing
EDA and feature selection

03

MODEL SELECTION

Describing the models that are
used in this research

04

EVALUATION

Machine learning performance by
objectives

05

CONCLUSION AND SUGGESTION

Project's conclusion and
suggestion for further use



INTRODUCTION



Timotius Marselo

timotiusmarselo@gmail.com

Data Analyst at UMN Consulting



Mario Caesar Kristantoputra

caesarmario87@yahoo.com

Data Engineer and Analytics at
FinAccel



Yoel

yoelcoding@gmail.com

Data Analyst at CIMB Niaga

BACKGROUND

Malawi is one of the world's least developed countries and is ranked 174 out of 189 countries according to the 2020 Human Development Index. It has about 16 million people, 53% living under the national poverty line and 90% living on less than \$2 per day.

Some of Malawi's main challenges are low economic growth and also commodity price increase. The increase in food prices and energy prices will determine Malawi's food insecurities and livelihood.

This project hopes to help alleviate some of Malawi's challenges, by classifying and ranking Malawi's cities based on their food insecurity level, in hope that government can prioritize the city with the most severe food insecurity level.



SCOPE AND OBJECTIVE



OBJECTIVE

- **Classify Malawi households** that are prone to food insecurities
- **Rank the Malawi cities** based on their food insecurity level



SCOPE

- **Households in Malawi** on year 2013, 2016, and 2019



KNOWING THE DATA

DATA COLLECTION



Dataset: Food Prices data for Malawi

Source: World Food Programme Price Database, updated by Humanitarian Data Exchange

Data description: Monthly data containing the price of food (such as maize, rice, and beans) and energy (such as diesel, petrol, and kerosene) in Malawi. Energy prices are available at the national level, while food prices are available at the local market level.

Link: <https://data.humdata.org/dataset/wfp-food-prices-for-malawi>

DATA COLLECTION



Dataset: Malawi Integrated Household Panel Survey 2010-2013-2016-2019 (Long-Term Panel, 102 EAs)

Source: World Bank Microdata Library

Data description: Household surveys conducted by the National Statistics Office of Malawi and the World Bank cover question regarding household, individual, agriculture, fishery, and community. For each year, the survey data are separated into multiple modules with household ID as the primary key.

DATA PREPROCESSING

1. **Convert data type.**
2. **Select important columns.**
3. **Check for null values.**
4. Make a function to **calculate the moving averages (3, 6, and 12 months) of each food (groundnuts, maize, rice, and beans) price** in each city for the years 2013, 2016, and 2019 (according to the date of data collection).
5. There are many cities with unavailable/ inconsistent monthly food price data, so we need to **filter some of the moving averages** (e.g. for the 6-month moving average, at least 4 monthly data should be available for that particular food and city).
6. **Merge all food price moving averages** for the same year.
7. Repeat step 4-6 to **calculate the moving averages of each food price in each region** (Northern, Central, and Southern). These moving averages will be used if the city-level moving averages are not available.
8. Make a function to **calculate the moving averages (3, 6, and 12 months) of fuel (diesel, kerosene, and petrol) at the national level.**

	Beans_MA3	Groundnuts_MA3	Maize_MA3	Rice_MA3	Beans_MA6	Groundnuts_MA6	Maize_MA6	Rice_MA6	Beans_MA12	Groundnuts_MA12	Maize_MA12	Rice_MA12
city												
Balaka	0.318033	0.285367	0.074533	0.325733	0.330320	0.314020	0.100075	0.295260	0.335400	0.323700	0.082322	0.316910
Blantyre City	0.339367	0.274567	0.086033	0.263033	0.360567	0.368583	0.099020	0.325133	0.340733	0.358092	0.087927	0.320683
Blantyre	0.421067	NaN	0.093467	0.336767	0.391280	NaN	0.113017	0.349960	0.348636	NaN	0.091375	0.320945
Chikwawa	NaN	NaN	0.089133	NaN	NaN	NaN	0.104800	NaN	NaN	NaN	0.091767	NaN
Chiradzulu	0.356750	0.405750	0.084167	0.311750	0.336275	0.361300	0.111825	0.307000	0.337244	0.359212	0.093267	0.307412
Chitipa	0.255211	0.269867	0.076033	0.334711	0.262400	0.278994	0.091489	0.334103	0.240508	0.262368	0.069165	0.312549
Dedza	0.290789	0.333172	0.076461	0.321606	0.282644	0.329286	0.087831	0.333383	0.294327	0.335015	0.076838	0.319274
Dowa	0.300117	0.317708	0.099189	0.334392	0.324485	0.366824	0.103911	0.354835	0.325710	0.349083	0.078908	0.326083
Karonga	0.370000	0.368300	0.080633	0.307550	0.347775	0.378600	0.104350	0.342340	0.349380	0.414900	0.092508	0.344027
Kasungu	0.390838	0.301221	0.072909	0.334536	0.363584	0.352440	0.087429	0.343805	0.357935	0.321178	0.068144	0.312550
Lilongwe	0.256544	0.286244	0.093878	0.309517	0.272906	0.327589	0.102239	0.326625	0.274186	0.281557	0.075715	0.312521

Figure 1. Example of price moving averages for some cities in 2013.

9. **Choose the survey modules** related to our objective for year 2013, 2016 and 2019 (Food security , Household Identification, Housing, Household Roster, Education, Durable Goods, and Filter Questions for Agriculture and Fishery).
10. **Handling missing values.**
11. **Process relevant columns for each module.** Convert the data to binary values (0, 1) if necessary.
 - **Food security:** calculate the weighted average for coping strategy questionnaire, which will be used to make our target variable (see next page for details)
 - **Household Identification:** urban/rural, city code, city name
 - **Housing:** type of construction materials, electricity, cell phone ownership
 - **Household Roster:** gender of the breadwinner
 - **Education:** calculate the average education per household
 - **Durable Goods:** ownership of assets such as radio, tv, bicycle, motorcycle, car, computer, bed, refrigerator, fan, lorry, minibus, kerosene stove
 - **Filter Questions for Agriculture and Fishery):** plot cultivation and livestock ownership

12. For each year, **merge all the processed data** into one dataset using "hhid" as the primary key.
13. **Merge the food price and fuel price moving averages data** with the dataset obtained in the previous step.
14. **Validate** the merged dataset.

FOOD SECURITY TARGET VARIABLE

```
[40] 1 weighted_fs_16['total_score'].quantile(.6)

7.0

[41] 1 fs_indicator_16 = weighted_fs_16[['y3_hhid', 'total_score']]
2 fs_indicator_16.columns = ['y3_hhid', 'fs_score']
3 fs_indicator_16['food_insecure'] = np.nan
4
5 def binarize_fs(fs_score):
6     if fs_score >= 7:
7         return 1
8     else:
9         return 0
10
11 fs_indicator_16['food_insecure'] = fs_indicator_16['fs_score'].apply(binarize_fs)
```

Figure 2. Configuring Food Security Target Variable

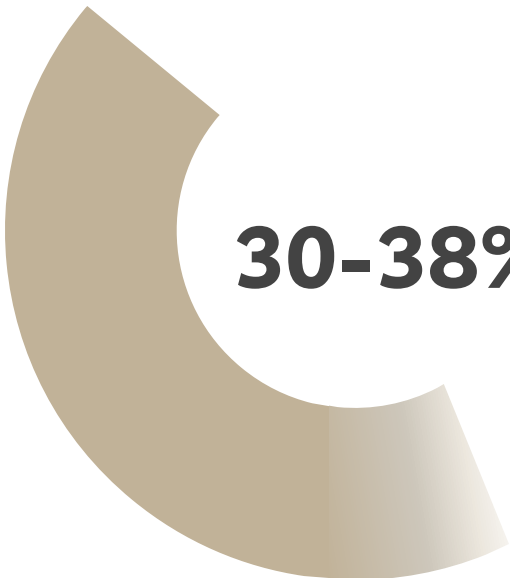
To make the target variable, we **convert the RCSI into a binary variable**. For each year, we **calculate the 60% percentile of the weighted average and use it as a cut-off value** to classify which households are food insecure (~40% of households are classified as insecure for each year). We use percentile instead of an absolute value because RCSI is subject to significant seasonal variations, so the values between different years are not comparable.

In the past 7 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	Severity Weight
Rely on less preferred and less expensive foods?	1
Limit portion size at mealtimes?	1
Reduce number of meals eaten in a day?	2
Restrict consumption by adults in order for small children to eat?	2
Borrow food, or rely on help from a friend or relative?	2

Table 1. Household CSI Index Weight
(The Coping Strategies Index Field Methods Manual Second Edition, January 2008)

RCSI (Reduced Coping Strategy Index)

The Reduced Coping Strategies Index (RCSI) is **a proxy indicator of household food insecurity**. It considers both the frequency and severity of five pre-selected coping strategies that the household used in the seven days prior to the survey. It is a simplified version of the full Coping Strategies Index indicator.



30-38%

Food security is generally defined as the state “**when all people at all times have both physical and economic access to sufficient food to meet their dietary needs** for a productive and healthy life” (USAID 1995).

According to the chronic food insecurity analysis conducted by the Malawi Vulnerability Assessment Committee (MVAC) and its partners in February 2022, **33 percent** of Malawi’s rural population (approximately 5.4 million people) are facing moderate (IPC CFI Level 3) to severe (IPC CFI Level 4) chronic food insecurity, with 12 percent of the country (1.9 million people) facing the highest level (IPC CFI Level 4 – Severe). ^[1]

[1] <https://malawi.un.org/sites/default/files/2022-05/IPC%20Malawi%20Chronic%20Food%20Insecurity%20Report%20May%202022.pdf>

EDA is done on the training dataset, which consists of the 2013 and 2016 datasets. The test dataset (2019 dataset) is not included in EDA to avoid data leakage.

Data dictionary:

- ***rural*** of 1 indicates rural residence (as opposed to urban).
- ***permanent_cons*** of 1 indicates that permanent construction materials (as opposed to tradition or semi-permanent) are used for the dwelling.
- ***elect*** of 1 indicates that a household has electricity working in their dwelling.
- ***male_head*** of 1 indicates that the breadwinner of the household is male (as opposed to female).
- ***educ_mean*** indicates the education mean of the household (calculated from ordinal variable).
- ***cellphone, radio, tv, bicycle, motorcycle, car, computer, bed, refrigerator, fan, lorry, minibus, and kerosene_stove*** with a value of 1 indicates that the household own that particular asset.
- ***cultivate_land*** of 1 indicates that the household cultivate a plot in that year.
- ***livestock*** of 1 indicates that the household own any livestock in that year.
- **Variables ending with _MA3, _MA6, and _MA12** indicate 3, 6, and 12-month moving average for that particular item.
- ***food_insecure*** is the target variable, a value of 1 indicates the household is food insecure.

In our train dataset (2016 and 2017 dataset), 43.2% of the household classified as food insecure.

Household Food Security Classification (2013 and 2016)

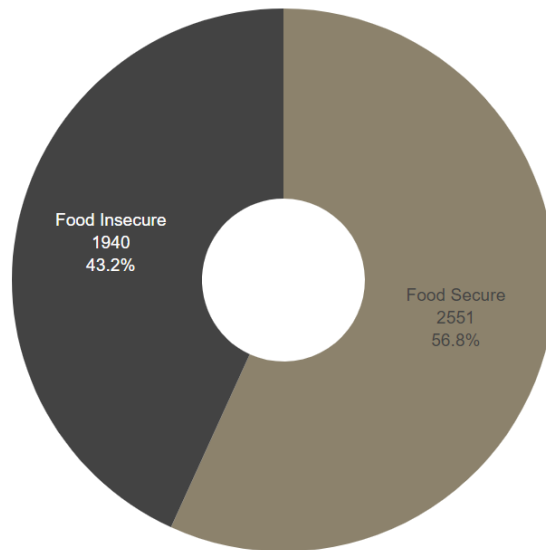


Figure 3. Household Food Security Classification

EDA AND FEATURE SELECTION

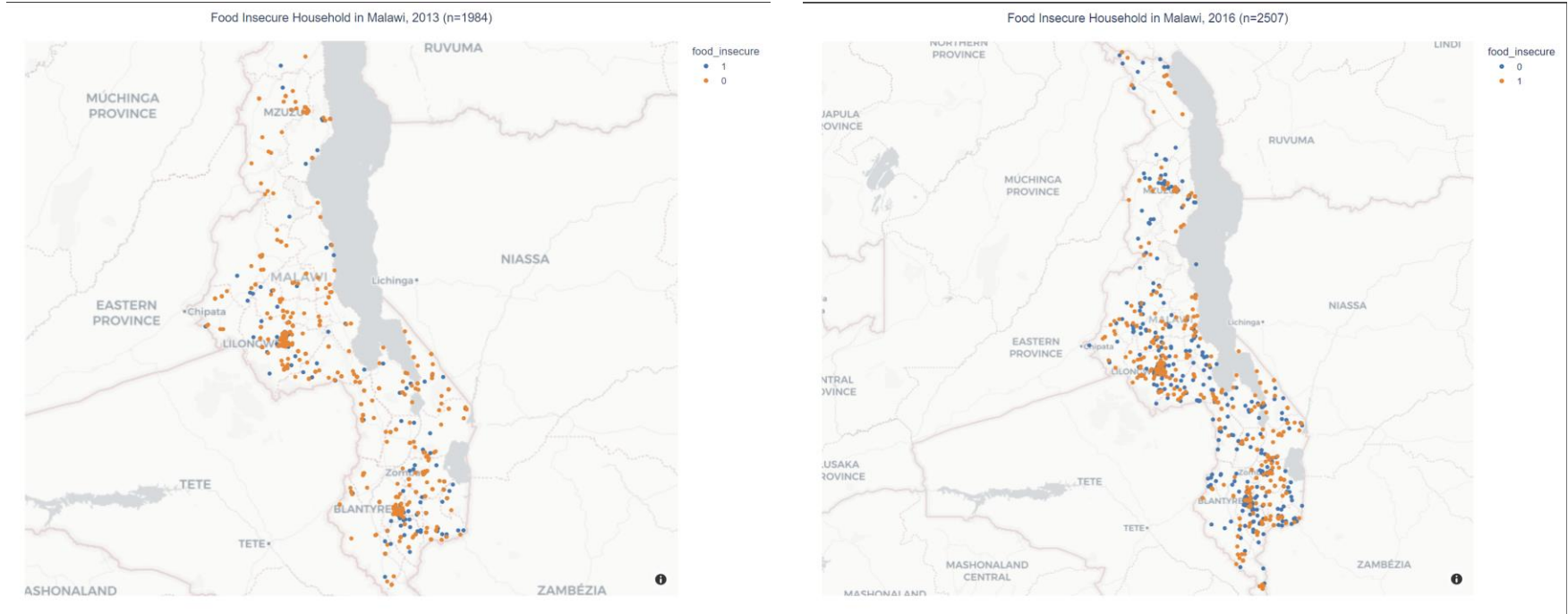


Figure 4. Household Insecurity Mapping 2013(Left) and 2016(Right)

EDA AND FEATURE SELECTION

Ownership of *motorcycle*, *car*, *lorry*, and *minibus* are quite low among household so they are combined into a new variable called *vehicle*.

Ownership of *kerosene_stove* and *computer* are low, so we will **not use** it as our feature.

rural	0.727900
permanent_cons	0.340459
elect	0.159653
cellphone	0.571810
male_head	0.758406
educ_mean	1.488930
radio	0.438210
tv	0.179693
bicycle	0.420396
motorcycle	0.019817
car	0.027165
computer	0.032509
bed	0.448675
refrigerator	0.083946
fan	0.066800
lorry	0.003785
minibus	0.002449
kerosene_stove	0.003785
cultivate_land	0.762636
livestock	0.474950

Figure 5. Feature Mean Values

EDA AND FEATURE SELECTION

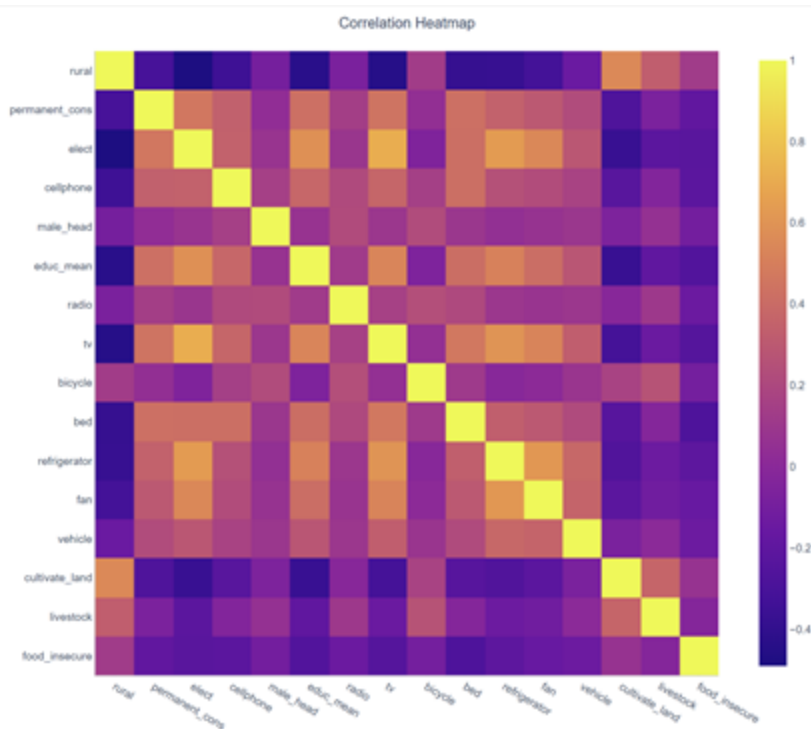


Figure 6. Correlation Heatmap of Features

tv, *refrigerator*, and *fan* have a medium-high correlation with *elect*, so we will **discard** them.

bicycle, *cultivate_land*, and *livestock* have a low correlation with our target variable (*food_insecure*), so we will **discard** them either.

EDA AND FEATURE SELECTION

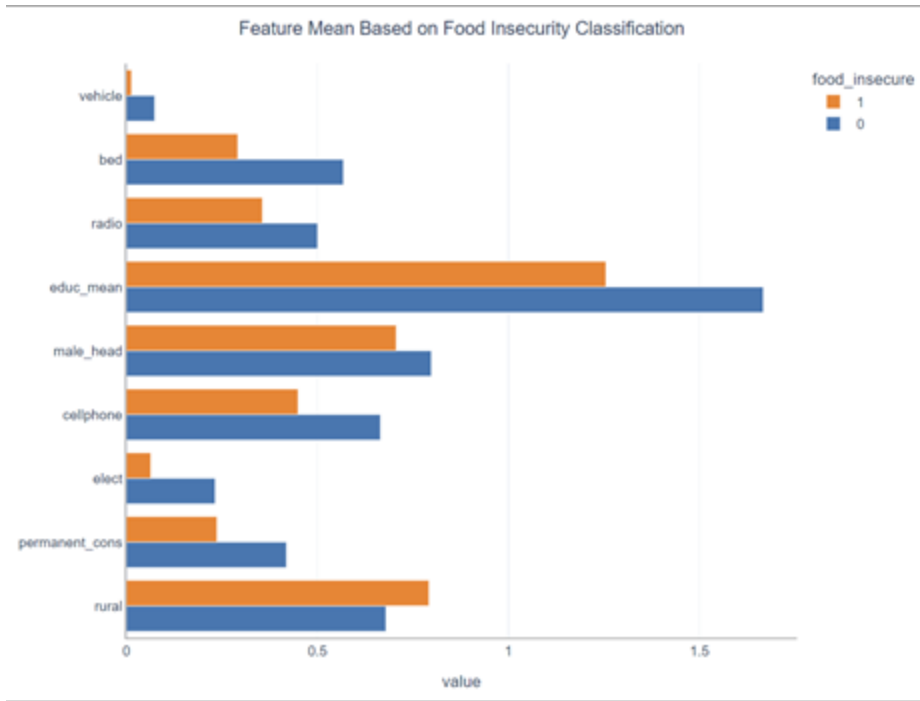



Figure 7. Mean values of features on *food_insecure* variable

Compared to food secure household, **the proportion of food insecure household who own vehicle, bed, radio, and cellphone are lower.** The food insecure household also **generally have lower education.** Regarding housing condition, food insecure household are **less likely to have permanent construction material and electricity** in their dwelling.



MODEL SELECTION

```

1 X_train = train_data[['rural', 'permanent_cons', 'elect', 'cellphone', 'radio', 'male_head',
2     'vehicle', 'bed', 'Petrol_MA3',
3     'Diesel_MA3', 'Kerosene_MA3', 'Petrol_MA6', 'Diesel_MA6',
4     'Kerosene_MA6', 'Petrol_MA12', 'Diesel_MA12', 'Kerosene_MA12',
5     'Beans_MA3', 'Groundnuts_MA3', 'Maize_MA3', 'Rice_MA3', 'Beans_MA6',
6     'Groundnuts_MA6', 'Maize_MA6', 'Rice_MA6', 'Beans_MA12',
7     'Groundnuts_MA12', 'Maize_MA12', 'Rice_MA12', 'educ_mean']]
8 Y_train = train_data["food_insecure"]
9
10
11 X_test = test_data[['rural', 'permanent_cons', 'elect', 'cellphone', 'radio', 'male_head',
12     'vehicle', 'bed', 'Petrol_MA3',
13     'Diesel_MA3', 'Kerosene_MA3', 'Petrol_MA6', 'Diesel_MA6',
14     'Kerosene_MA6', 'Petrol_MA12', 'Diesel_MA12', 'Kerosene_MA12',
15     'Beans_MA3', 'Groundnuts_MA3', 'Maize_MA3', 'Rice_MA3', 'Beans_MA6',
16     'Groundnuts_MA6', 'Maize_MA6', 'Rice_MA6', 'Beans_MA12',
17     'Groundnuts_MA12', 'Maize_MA12', 'Rice_MA12', 'educ_mean']]
18 Y_test = test_data['food_insecure']
19
20 scaler = StandardScaler()
21 scaled_X_train = scaler.fit_transform(X_train.loc[:, 'Petrol_MA3':])
22 scaled_X_train = np.concatenate([X_train.loc[:, : "bed"].values, scaled_X_train], axis=1)
23 scaled_X_test = scaler.transform(X_test.loc[:, 'Petrol_MA3':])
24 scaled_X_test = np.concatenate([X_test.loc[:, : "bed"].values, scaled_X_test], axis=1)

```

Figure 8. Creating train, test, and standardizing data

Besides the features we have selected before, we will also include the food and energy price moving averages as our features. The continuous variable are first standardized to transform the data into a standard format.


```

1 from sklearn.metrics import f1_score
2
3 xgb = XGBClassifier(random_state=42)
4 gnb = GaussianNB()
5 dtc = DecisionTreeClassifier(random_state=42)
6 knn = KNeighborsClassifier()
7
8
9 models_df = pd.DataFrame()
10 models_df['Model'] = ['XGBoost', 'Decision Tree', 'GaussianNB', 'KNeighbors']
11 models_df.index = [xgb, dtc, gnb, knn]
12
13 for model in models_df.index:
14     model.fit(scaled_X_train, Y_train)
15     Y_pred = model.predict(scaled_X_test)
16     models_df.loc[model, 'f1_score'] = f1_score(Y_pred, Y_test)

```

```

1 models_df.sort_values(by="f1_score", ascending=False)

```

	Model	f1_score	
	GaussianNB()	GaussianNB	0.673285
	XGBClassifier(random_state=42)	XGBoost	0.634670
	DecisionTreeClassifier(random_state=42)	Decision Tree	0.590332
	KNeighborsClassifier()	KNeighbors	0.571429

Figure 9. Training Machine Learning Models

MODEL SELECTION

We train **4 classifier models** (XGBoost, Gaussian Naive Bayes, Decision Tree, and K-Nearest Neighbors) on our test dataset (2013 and 2016) and use them to predict the food insecure households in 2019.

To ensure the balance between precision and recall, we use the **f1-score** for the food insecure class as our scoring method.

Gaussian Naive Bayes and XGBoost have the highest f1 score, so we will further evaluate them.



EVALUATION

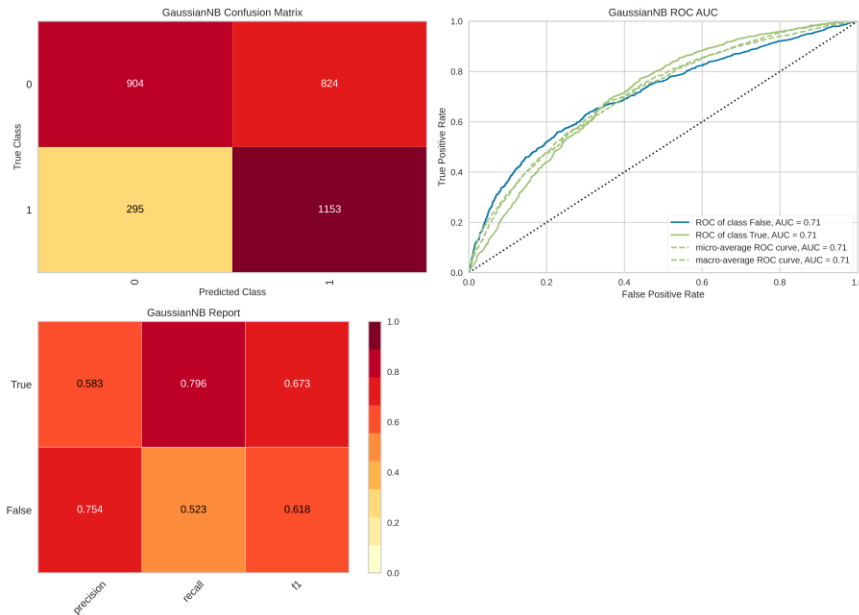


Figure 10. GaussianNB Model Evaluation

GaussianNB model has a **recall of 0.796** and **precision of 0.583** for the food insecure class. This means out of the 1448 food insecure households, **the model have predicted 79.6% of them as food insecure**; and out of the 1977 households predicted as insecure, **58.3% of the households are truly food insecure**.

The **ROC-AUC** for food insecure class is **0.71** which means the diagnostic ability of the model is **acceptable**.

Classification of food insecure households with XGBoost

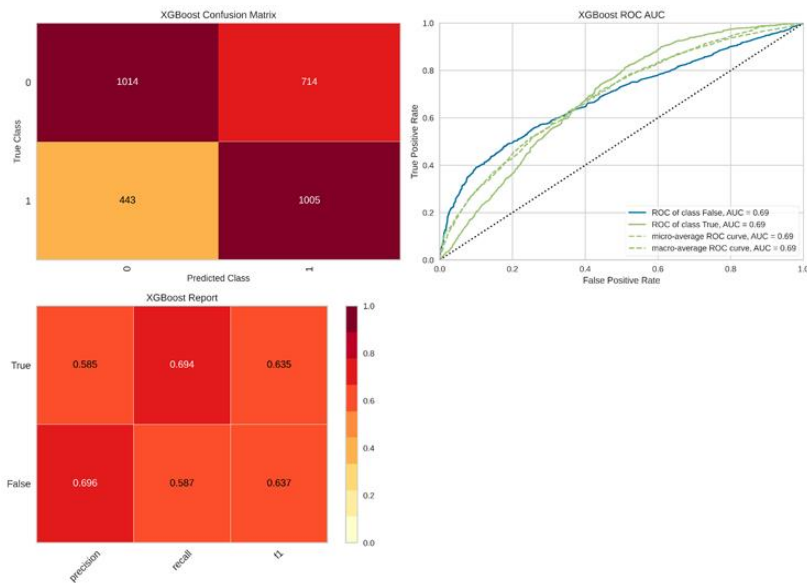


Figure 11. XGBoost Model Evaluation

XGBoost model has a **recall of 0.694** and **precision of 0.585** for the food insecure class. This means out of the 1448 food insecure households, **the model have predicted 69.4% of them as food insecure**; and out of the 1719 households predicted as insecure, **58.5% of the households are truly food insecure**.

The **ROC-AUC** for food insecure class is **0.69** which means the diagnostic ability of the model is **acceptable**.

MODEL EVALUATION OBJECTIVE 1

We want to rank the cities based on the proportion of the food insecure households so policymakers can prioritize their (limited) resources on the more impacted cities. First we input the predictions from the **XGBoost** and **GaussianNB** model into the 2019 dataset. We then **calculate the mean of *food_insecure* and the predictions for each city**, which indicates the proportion of food insecure household in each city. Last **we compare the food insecurity ranking across cities** from the ground truth and models prediction by using Spearman's rank correlation.

```
[26] 1 xgb = XGBClassifier(random_state=42)
     2 xgb.fit(scaled_X_train,Y_train)
     3 Y_pred_xgb = xgb.predict(scaled_X_test)
     4 final_19["Y_pred_xgb"] = Y_pred_xgb
```

```
[27] 1 gnb = GaussianNB()
     2 gnb.fit(scaled_X_train,Y_train)
     3 Y_pred_gnb = gnb.predict(scaled_X_test)
     4 final_19['Y_pred_gnb'] = Y_pred_gnb
```

```
[37] 1 final_19_mean = final_19.groupby("city_name").mean()
     2 final_19_mean
```

Figure 12. Fitting and Predicting City-Level Data Script

Spearman's rank correlation is calculated to assess the relationship between the proportion of food insecure household for each city based on Gaussian NB prediction and the actual value. There is a moderate positive monotonic correlation between the two variables, $r(29) = .54$, $p = .002$.

Out of the 15 cities with the highest food insecurity level, the ranking derived from GaussianNB prediction have identified 11 of them correctly (73.3% accuracy).

```
1 pg.corr(final_19_mean.food_insecure, final_19_mean.Y_pred_gnb, method="spearman")
```

	n	r	CI95%	p-val	power
spearman	31	0.541587	[0.23, 0.75]	0.001652	0.901871

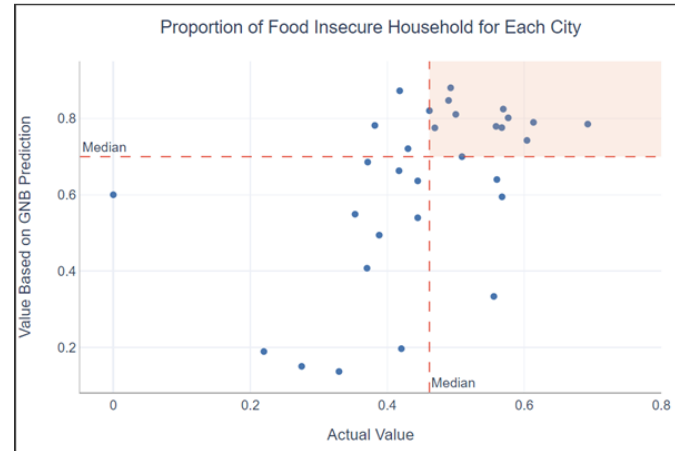


Figure 13. GaussianNB Evaluation on Malawi Cities Ranking

Spearman's rank correlation is calculated to assess the relationship between the proportion of food insecure household for each city based on XGBoost prediction and the actual value. There is a moderate positive monotonic correlation between the two variables, $r(29) = .41$, $p = .02$.

Out of the 15 cities with the highest food insecurity level, the ranking derived from XGBoost prediction have identified 11 of them correctly (73.3% accuracy).

```
1 pg.corr(final_19_mean.food_insecure, final_19_mean.Y_pred_xgb, method="spearman")
```

	n	r	CI95%	p-val	power
spearman	31	0.408912	[0.06, 0.67]	0.022368	0.644298

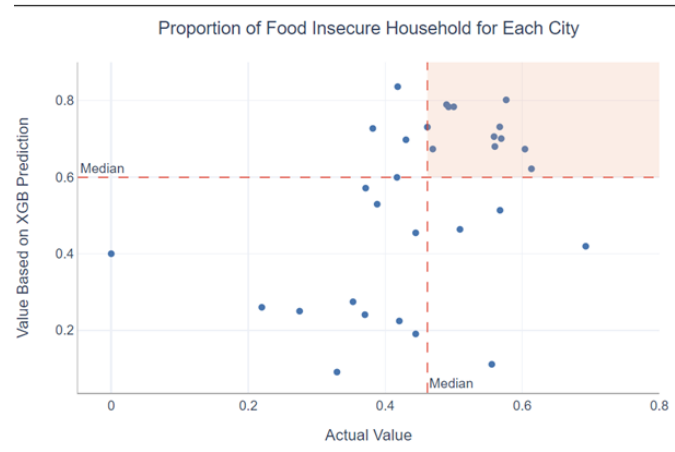


Figure 14. XGBoost Evaluation on Malawi Cities Ranking

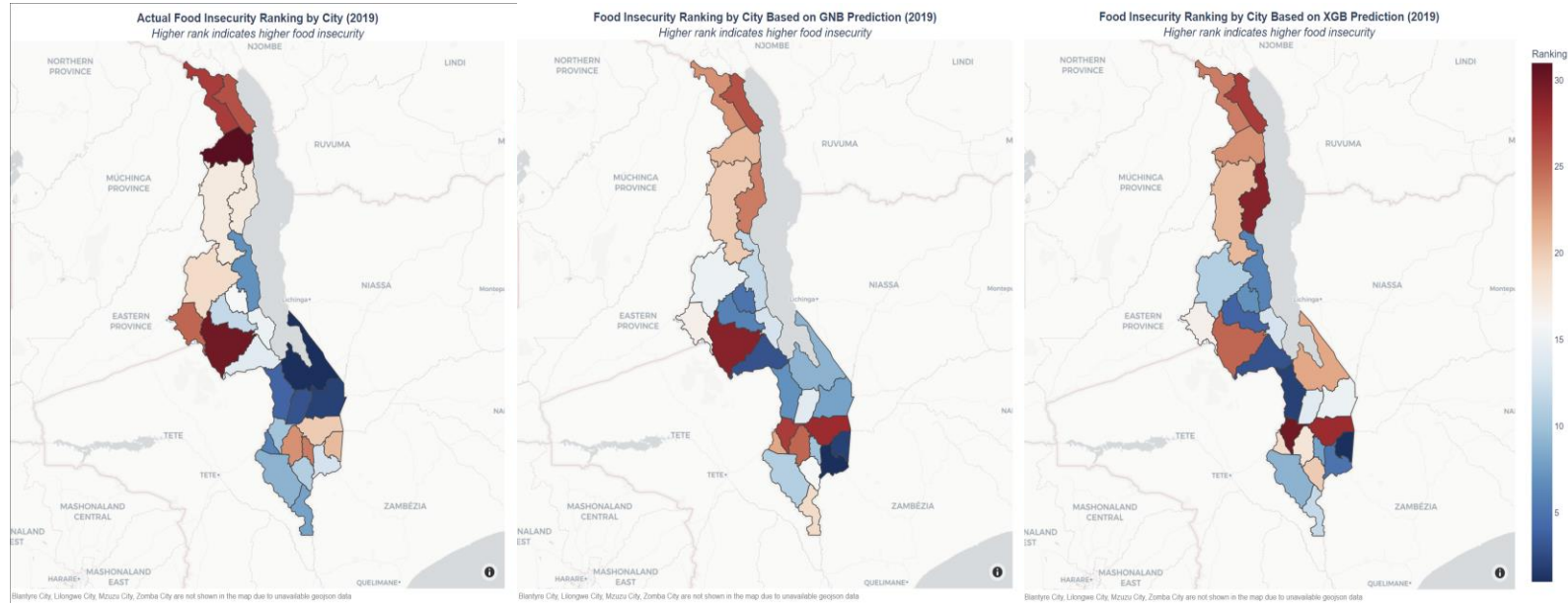


Figure 15. Food Insecurity Ranking Visualizations



CONCLUSION AND SUGGESTION

- In this project, we have **select the two Malawi public datasets**, namely food prices data and integrated household panel survey datasets. We also performed **data preprocessing, calculating the moving average, and merging** those datasets.
- We also have **built prediction models** to predict food insecure Malawi households based on their assets, housing, demographic, energy price, and regional food price. From the 4 models that we built, we found that the best models are **XGBoost with 0.69 ROC-AUC score** and **GaussianNB with 0.71 ROC-AUC score**.
- Finally, we have **designed a method to predict the ranking of Malawi cities** based on their food insecurity level.

1. The energy price data is only available at national-level, hence we used the same value for fuel price moving average for all households on our model. For this case, we suggest to use city-level energy price data in order to improve the model.
2. Similarly, the food price data is not available consistently for many cities, hence we used region-level food price moving averages for those cities. The model may improve if the food price data at city-level are updated consistently for all cities.
3. The model performance can be improved by including more years on the training dataset, performing more rigorous feature selection/ engineering and hyperparameter tuning.
4. The model created can be applied to other regions as well, by adjusting the food and energy items accordingly.

THANK YOU

For your attention!

Presentation by Mayo Team
2022

