Harvard John A. Paulson
School of Engineering
and Applied Sciences

Mapping Nutritional Equity: Affordability of a Healthy Diet Around the World

Aseel Rawashdeh
Panthon Imemkamon
Omar Siddiqui
Talha Rehman
Mahmoud Elshafeei

Motivation

Problem Statement: We would like to predict the Percentage of the population unable to afford a healthy diet using a variety of different predictors of country-level data.

Motivation: These predictors primarily center around the agricultural output of a country, and the yield of multiple different types of crops. These crops can act as a proxy for the region but also have a bit more flexibility than simply geography because different regions may have similar agricultural yields.

Description of the data: We have collected our dataset from the FAOSTAT, which provides free access to food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available.

Data Cleaning

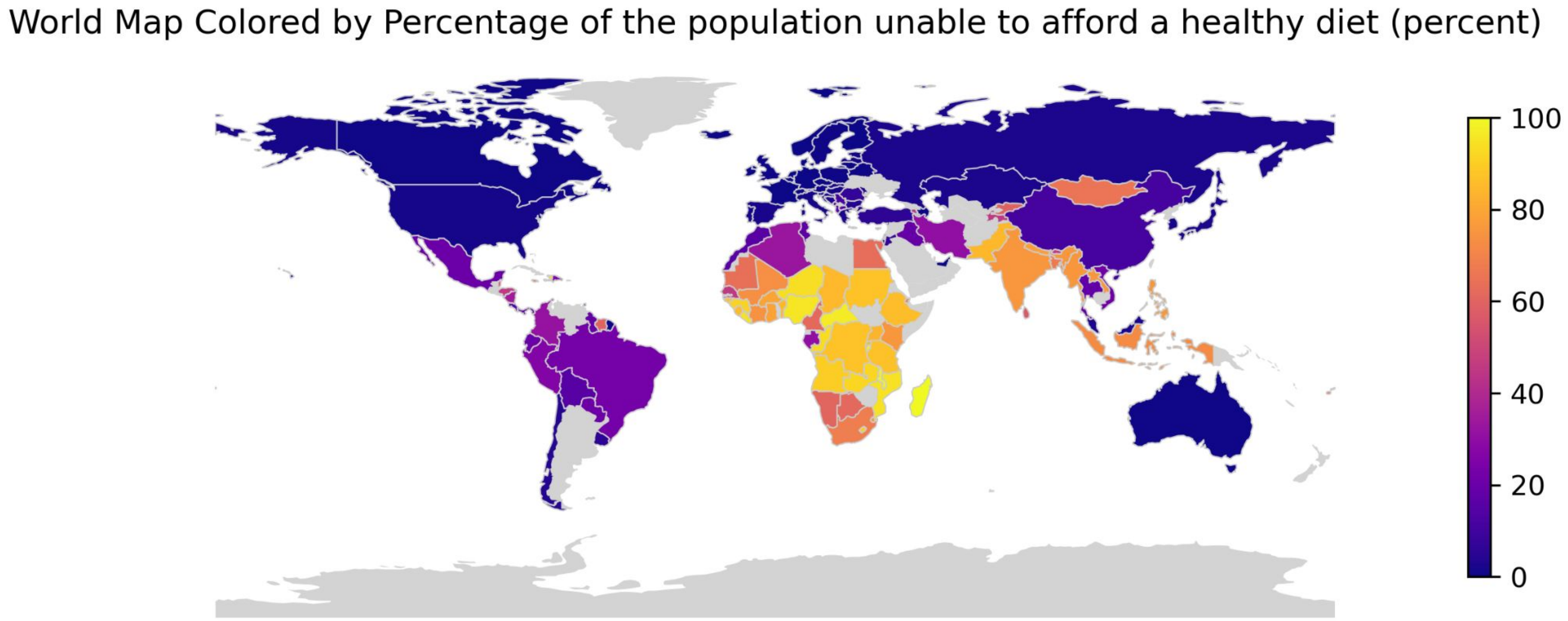
- The original dataset comes in separate files. We pivoted using Country as index and merge them into a single dataframe.
- Out of 238 countries in our original dataset, only 140 have the data for percentage of population unable to afford a healthy diet – our response variable – and so we dropped the rest.
- Columns with >80% null are dropped, the rest filled with zero.
- Ultimately we have
    - 140 rows of data
    - 218 columns of data (216 predictors, 1 response variable, 1 index(Country))

	Country	Forest land, Area	Temporary fallow, Area	Temporary meadows and pastures, Area	Temporary crops, Area	Cropland, Area per capita	Cropland, Area	Arable land, Area	Agriculture, Area	Land area, Area	Country area, Area	Percentage of the population unable to afford a healthy diet (percent)	Cost of a healthy diet (PPP dollar per person per day)	Agricultural land, Area	Other land, Area	Apples and products, Domestic supply quantity	Vegetables, other, Production	Vegetables, other, Domestic supply quantity	Oranges, Mandarines, Domestic supply quantity	Citrus, Other, Domestic supply quantity	Cocoa Beans and products, Domestic supply quantity	Grapes and products (excl wine), Domestic supply quantity	Fruits, other, Production	Fruits, other, Domestic supply quantity
0	Albania	788.900	179.300	219.70	200.900	0.241	687.530	599.90	1136.330	2740.0	2875.0	15.9	4.388	1136.330	814.770	114.0	906.0	851.0	59.0	1.0	4.0	214.0	162.0	172.0
1	Algeria	1958.333	2848.600	0.00	4682.000	0.193	8509.571	7530.60	41316.071	238174.1	238174.1	32.4	4.043	41316.071	194899.696	524.0	6379.0	6514.0	1467.0	1.0	43.0	644.0	960.0	994.0
2	Angola	66052.313	601.002	587.98	4184.019	0.165	5690.000	5373.00	45897.000	124670.0	124670.0	88.1	5.031	45897.000	12720.687	8.0	745.0	757.0	5.0	446.0	4.0	1.0	99.0	104.0
3	Armenia	328.260	216.266	0.00	227.154	0.180	503.720	443.42	2042.080	2847.0	2974.0	41.4	3.527	1674.820	476.660	101.0	528.0	543.0	32.0	0.0	7.0	235.0	258.0	234.0
4	Australia	134005.100	5552.678	0.00	25712.322	1.221	31650.000	31265.00	387265.000	769202.0	774122.0	0.7	2.437	363519.000	247931.900	501.0	1541.0	3218.0	381.0	12.0	72.0	1829.0	498.0	720.0

Exploratory Data Analysis

Response Variable: Percentage of Population unable to afford a healthy diet

20.7% is the median percentage of Population unable to afford a healthy diet,
but some countries especially in Africa/South Asia/Southeast Asia can have very high percentage of population unable to afford healthy diet



Exploratory Data Analysis

Top predictors for Percentage of Population unable to afford a healthy diet, listed by correlation coefficients are shown below

	Feature	Correlation with Response Variable
192	Sesame seed, Production	0.29
204	Plantains, Production	0.28
118	Plantains, Domestic supply quantity	0.27
180	Cassava and products, Production	0.27
10	Cost of a healthy diet (PPP dollar per person per day)	0.25
78	Cassava and products, Domestic supply quantity	0.25
84	Roots, Other, Domestic supply quantity	0.24
177	Roots, Other, Production	0.24
68	Beverages, Fermented, Domestic supply quantity	0.20
166	Beverages, Fermented, Production	0.20

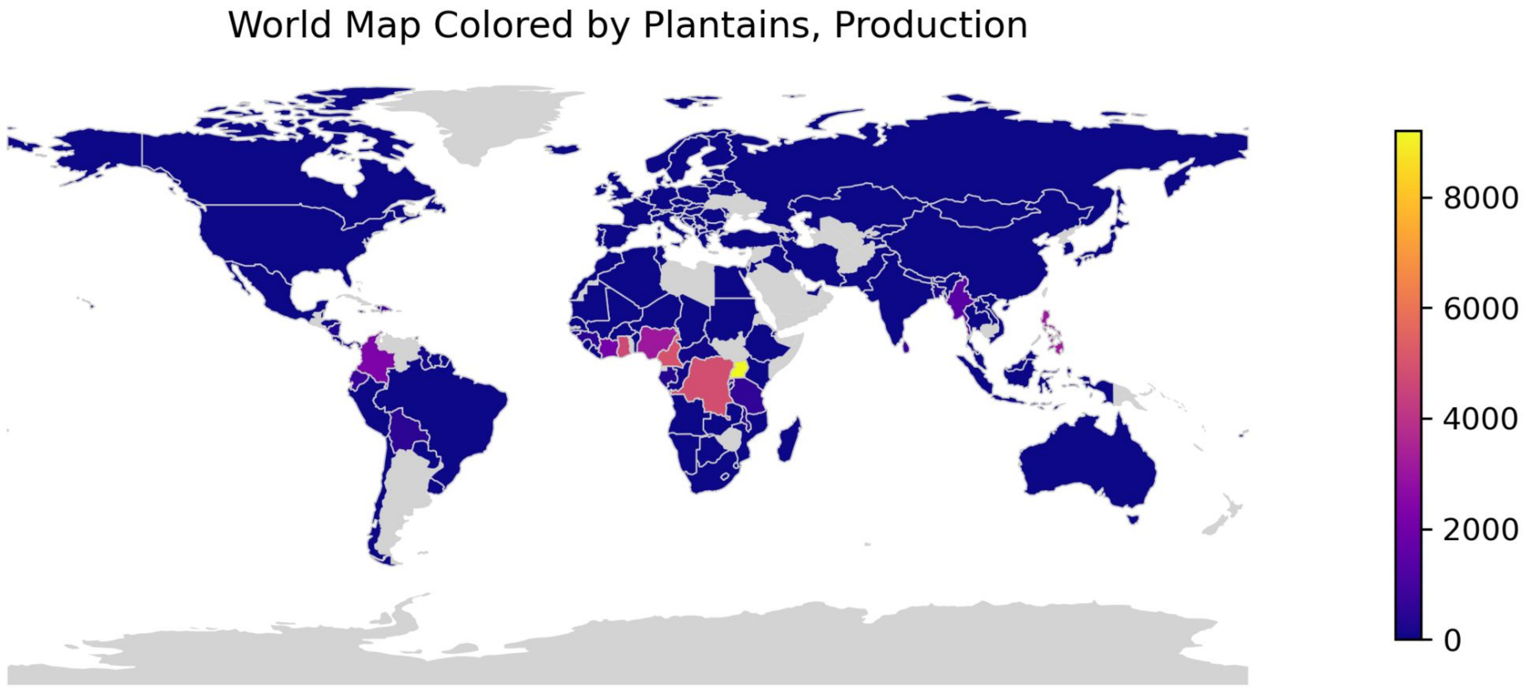
Top **positive** correlation
(associated with inability to afford healthy diet)

	Feature	Correlation with Response Variable
142	Sugar beet, Domestic supply quantity	-0.26
188	Sugar beet, Production	-0.26
77	Oats, Domestic supply quantity	-0.30
176	Oats, Production	-0.30
123	Cream, Domestic supply quantity	-0.31
70	Wine, Domestic supply quantity	-0.31
193	Cream, Production	-0.31
159	Barley and products, Production	-0.32
80	Barley and products, Domestic supply quantity	-0.34
85	Population, Domestic supply quantity	NaN

Top **negative** correlation
(associated with less inability to afford healthy diet)

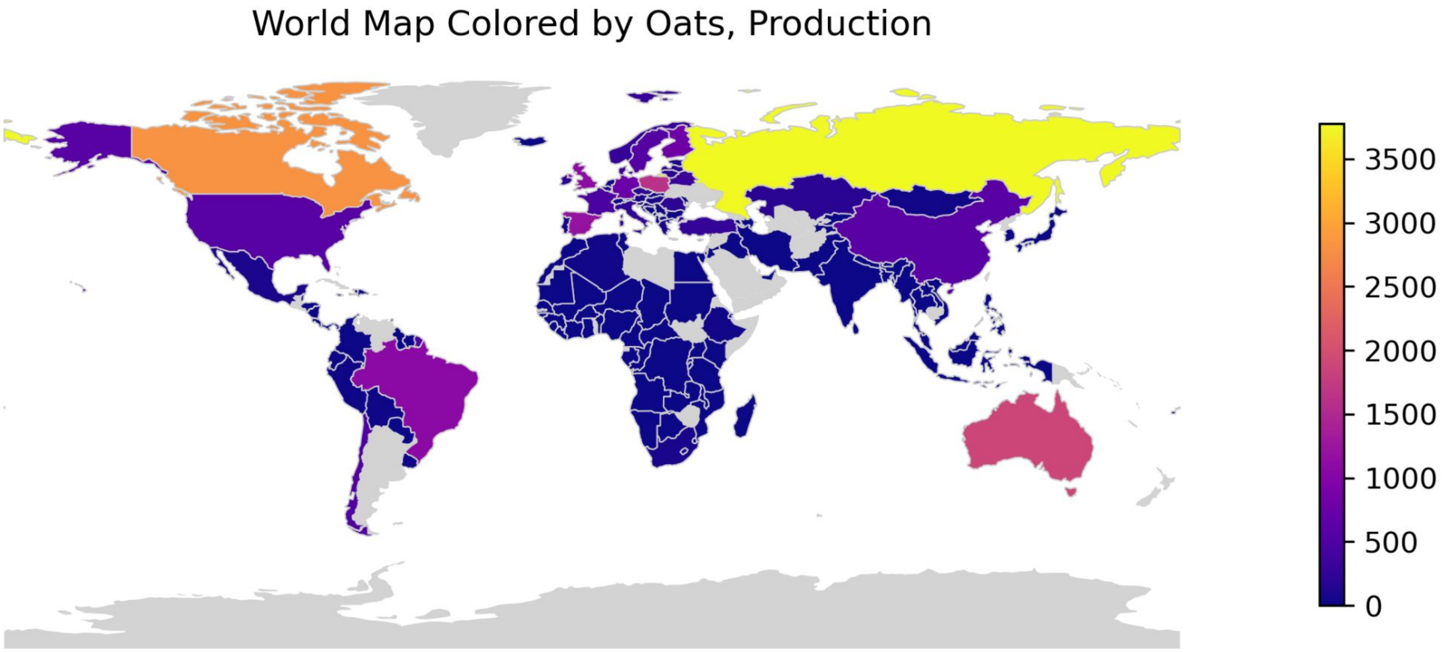
Exploratory Data Analysis

Plantain is associated with higher percentage of population unable to afford a healthy diet.
Coincidentally we can see that it is a warm-climate plant and is produced in Central America, Africa, and Southeast Asia



Exploratory Data Analysis

Oats are associated with lower percentage of population unable to afford a healthy diet.
Coincidentally we can see that it is a colder-climate plant and is produced in Europe, North America, and some South American countries



Main Challenge: High Dimensionality of Data

The main challenge of our project is having more predictors than we have data points.
This can lead to severe overfitting during modeling.

However, by plotting 2D PCA plot of the transpose of data (each feature now is a data point and the countries act as features instead), we can see that the majority of features are clumped together.

This infers that there might be many features that are similar and we might be able to reduce the dimension of this dataset.

Modeling: Baseline Linear Model

We divide the data into train-test set with 80-20 split, then we performed cross-validation for each of our model.

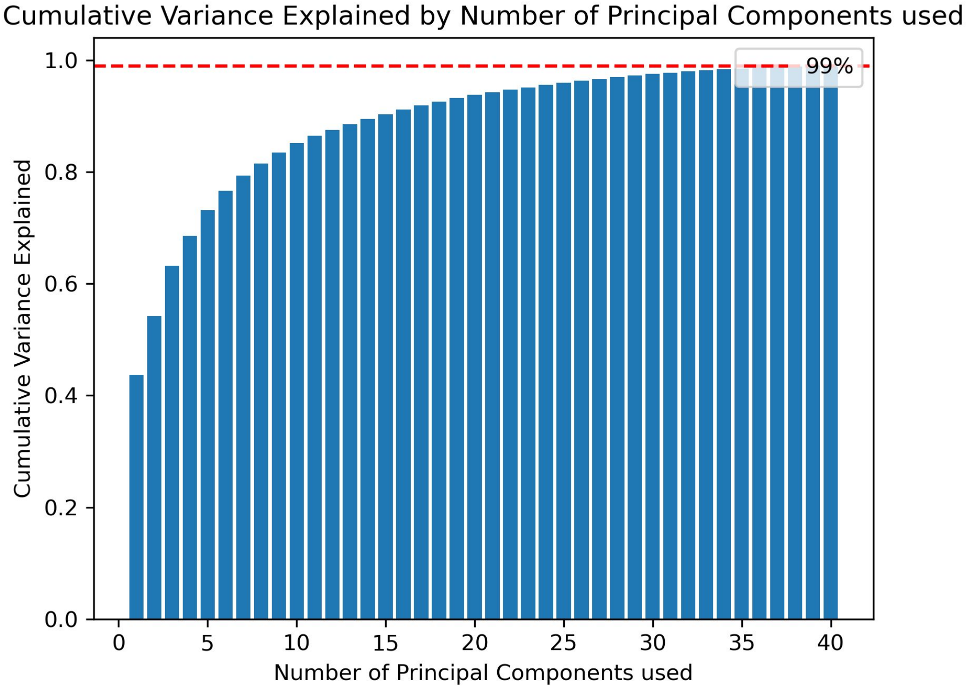
Our evaluation metric is **Mean Absolute Error(MAE)** as it is intuitive to interpret how far off the model is.

Baseline Model: Linear Regression

Linear Regression with all features
Training MAE 0.0
Validating MAE 35.96
Test MAE 54.76

Notice that our training MAE is 0 – or the model overfits perfectly.
This is because we have more dimensions than the number of data points.

Modeling: Preprocessing Features with PCA



Fortunately, PCA shows that we could retain 99%+ of the variance in the dataset with just the first 40 principal components.

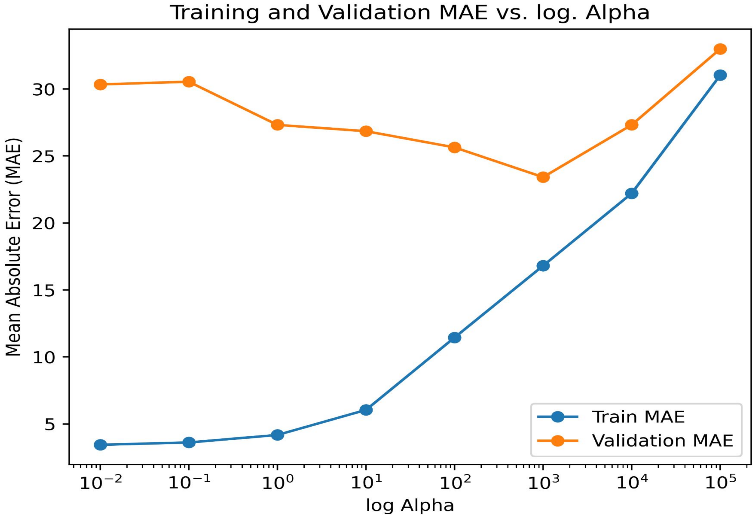
Using PCA to preprocess the dataset allows us to overfit less with Linear Regression model.

Linear Regression with **with PCA pre-processing**

Training MAE 14.5
Validating MAE 26.41
Test MAE 31.2

Modeling: Lasso Regularization

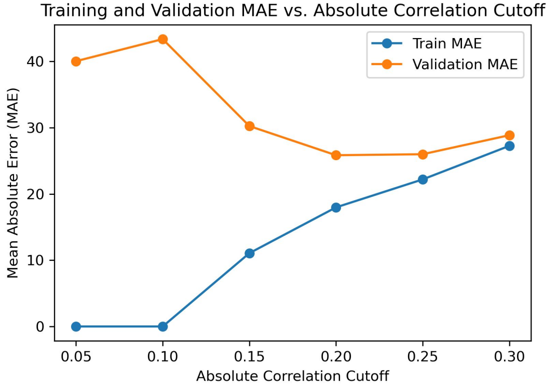
Alternatively, we can also use L1 regularization to reduce the number of features used in the model and lessen overfitting.
Our best lasso model utilizes just 34 features out of 216!



Lasso with best alpha = 1000
Estimated number of non-zero coefficients the Lasso model is using: 34
Training MAE 16.81
Validating MAE 23.42
Test MAE 30.72

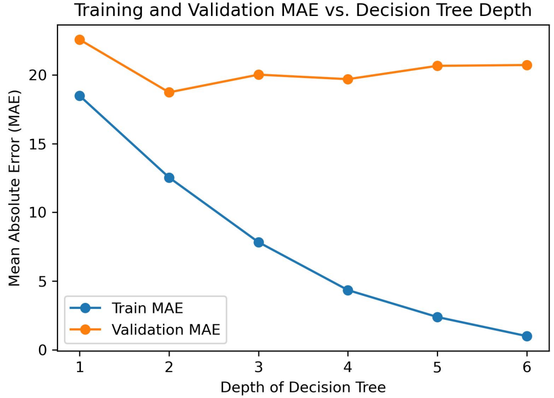
Modeling: Other Techniques

Linear Regression - Keeping only strong predictors (high correlation with response variable)



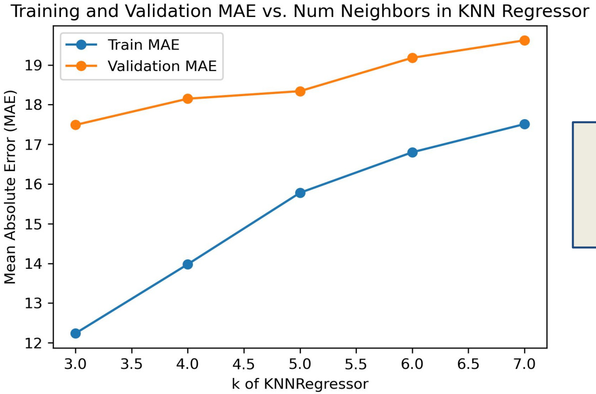
Training MAE 17.97
Validating MAE 25.84
Test MAE 33.62

Decision Tree – Limit Depth



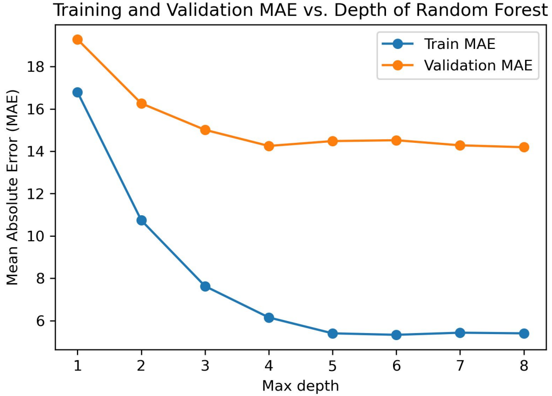
Best depth = 2
Training MAE 12.54
Validating MAE 18.74
Test MAE 17.19

KNN - varies K



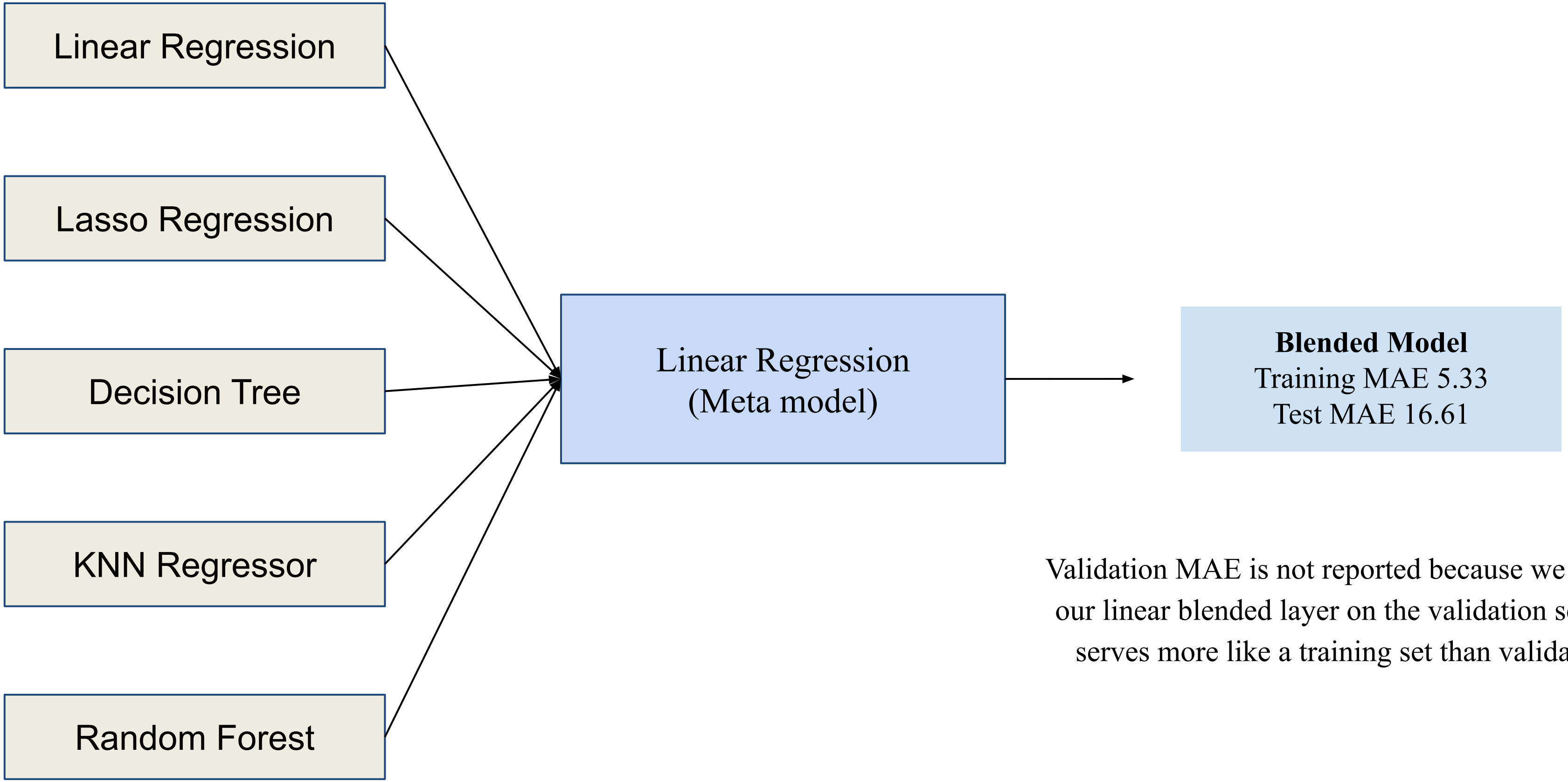
Best k = 3
Training MAE 12.24
Validating MAE 17.49
Test MAE 35.53

Random Forest - Limit Depth



Best depth = 8
Training MAE 5.4
Validating MAE 14.19
Test MAE 13.91

Modeling: Blending

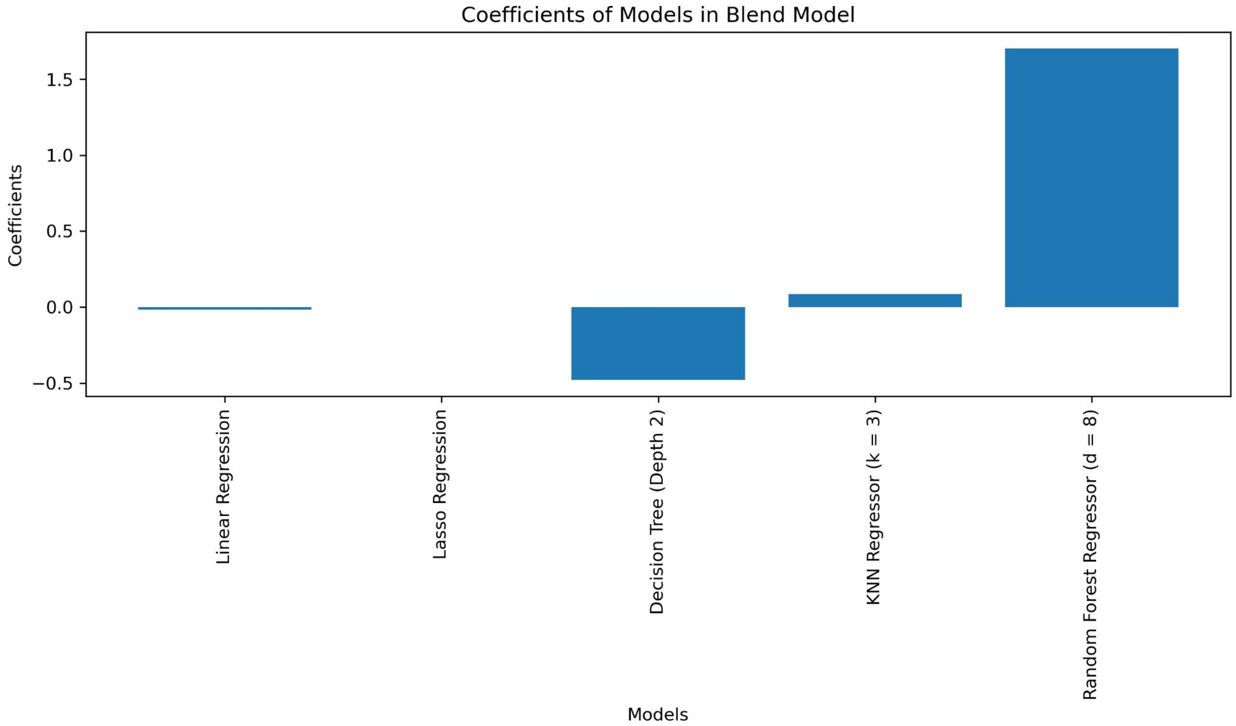


Validation MAE is not reported because we trained our linear blended layer on the validation set so it serves more like a training set than validation.

Modeling: Blending

Our Meta model is taking most of the prediction from Random Forest and adjusts it by a bit using the prediction from Decision Tree.

Blending performs worse than our pure Random Forest model. Blending technique might need more data for better tuning of the blended layer.



Results

Our Random Forest model is the best-performing model, with test-set MAE of 13.91

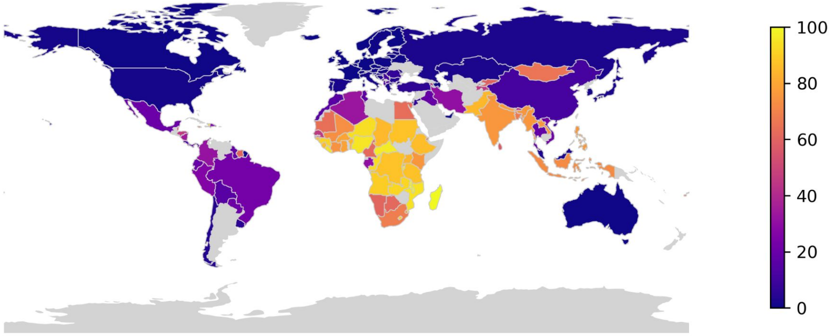
model	Train MAE	Val MAE	Test MAE
Basic Linear Regression	0.00	35.96	54.76
Linear Regression with PCA	14.50	26.41	31.20
LASSO Regression (Baseline Model)	16.81	23.42	30.72
Linear Regression with Manual Feature Selection (Correlation Analysis)	17.97	25.84	33.62
Decision Tree (Depth 2)	12.54	18.74	17.19
KNN Regressor (k = 5)	12.24	17.49	35.53
Random Forest Regressor (d = 8)	5.40	14.19	13.91
Blended Model	5.33	NaN	16.61

Conclusion

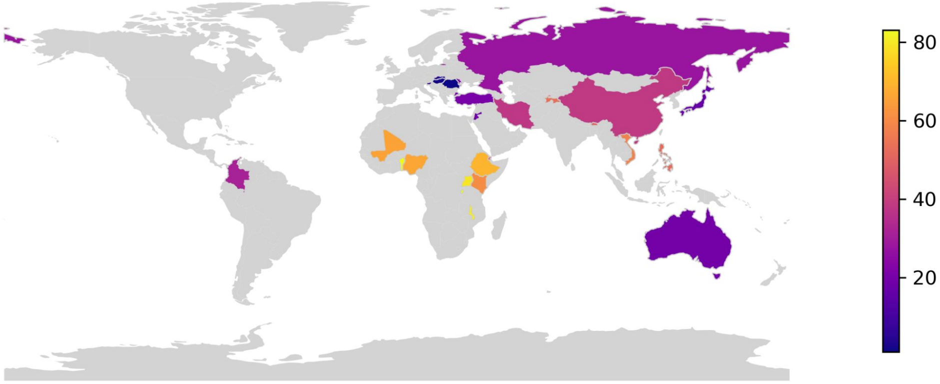
- We are able to make accurate predictions (MAE <15) using mostly just agricultural production data
- Predictions are made based on geographic location and regional economies, using crops as proxies
- Our model generally makes more conservative predictions that the observed data, over-predicting for wealthy countries with less developed neighbors

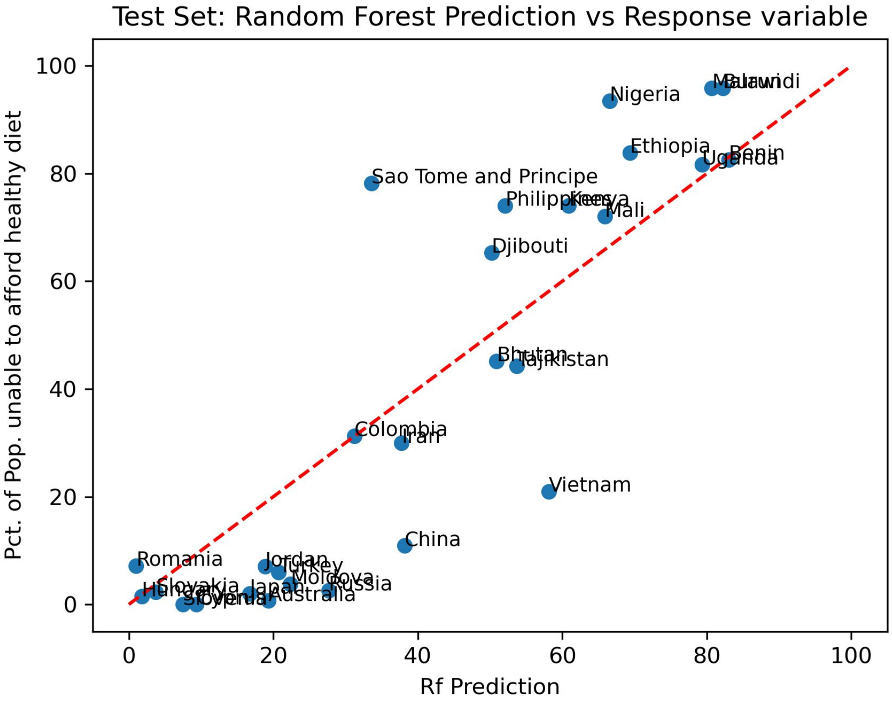
Conclusion: Test Set vs Predicted Map

World Map Colored by Percentage of the population unable to afford a healthy diet (percent)

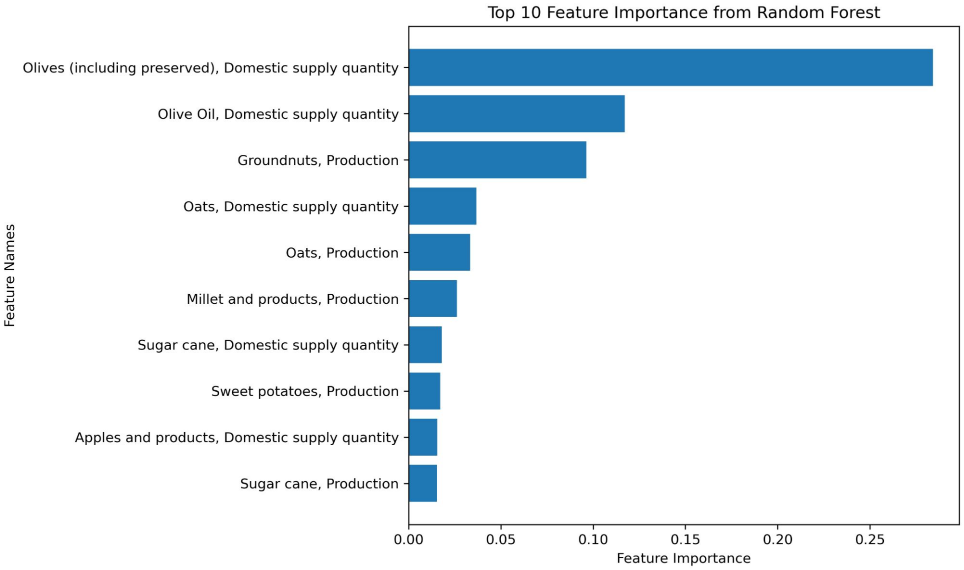


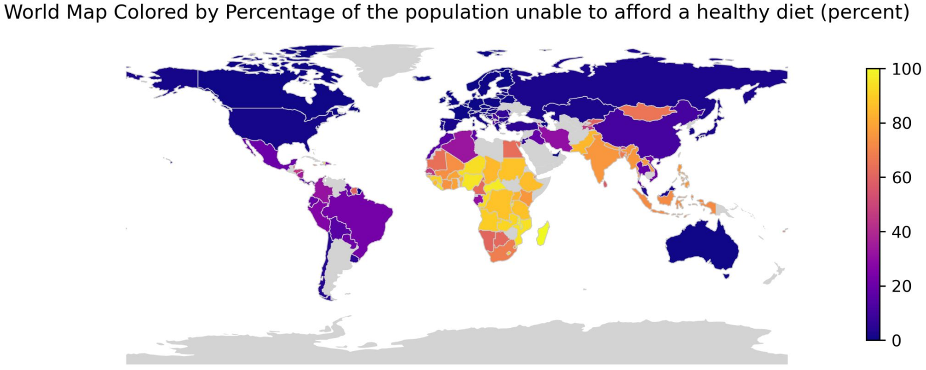
World Map Colored by rf_prediction

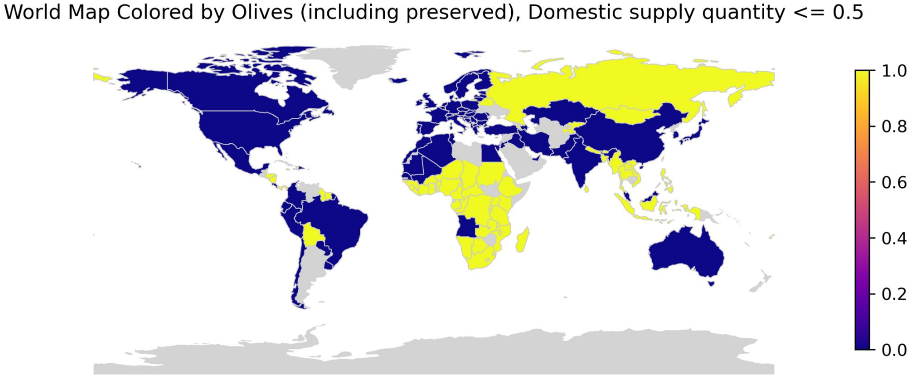




Conclusion: What Our Model is Really Learning







Future Directions

- FAOSTAT has many more domains of data, some of which might be an even better predictor e.g. food accessibility
- We would like to try dimensionality reduction and looking into more detailed feature selection in the future
- We in the future would also like to use KNN imputation instead of imputing zeros for missing datas
- Training a classification model that suggests effective policy for countries based on their crop production and yield