Predicting Crop Yield

Team 8: Cassidy Gasteiger, Raymond Li, Suraj Shourie, Asadullah Syed, Angela Zheng

Agenda

- Problem Description
- Exploratory Data Analysis
- Modeling
- Modeling Analysis
- Conclusions and Recommendations

Problem Description

- *Question: How can we predict crop yield annually?*
- Food system experts must be able to accurately predict crop yield based on pre-harvest conditions to feed a growing world population
- We expect environmental conditions like rainfall, pests, and climate to be highly predictive of yield
- Our objective is to examine other socioeconomic metrics like GDP or fertility rates (wealth-linked metrics) to determine if they may shed light on differences in yield between geographies

Data Description 1/2

CROP YIELD DATA

The primary crop yield data was obtained from Kaggle.
The original data set consisted of 27,228 observations of yield for 10 unique crops from 98 countries across a 23-year period (1990-2013)

Variable Name	Type	Description
Area	Qualitative	Country in question (eg, USA, UK) --> 98 unique countries included in the dataset
Item	Qualitative	Type of Crop (eg, Maize, Potato) --> 10 unique crop types included in the dataset
Year	Quantitative	Year of produce (1990 - 2013)
Yield	Quantitative	Yield for each type of crop measured in hectogram yield per hectare
Average annual rainfall	Quantitative	Annual recorded rainfall for the given country in the given year measured in mm
Pesticides	Quantitative	Total Amount of pesticides used in Crops per year measured in tonnes
Average Temperature	Quantitative	Annual Average temperature of the country measured in degrees Celsius

[Kaggle Crop Yield Data Link]

Data Description 2/2

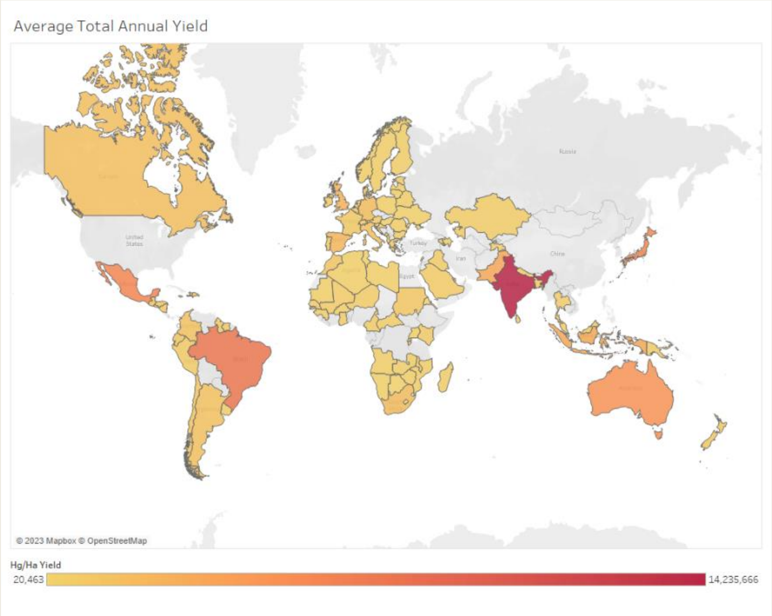
SOCIOECONOMIC AND LAND USE VARIABLES FROM WORLD BANK DATA

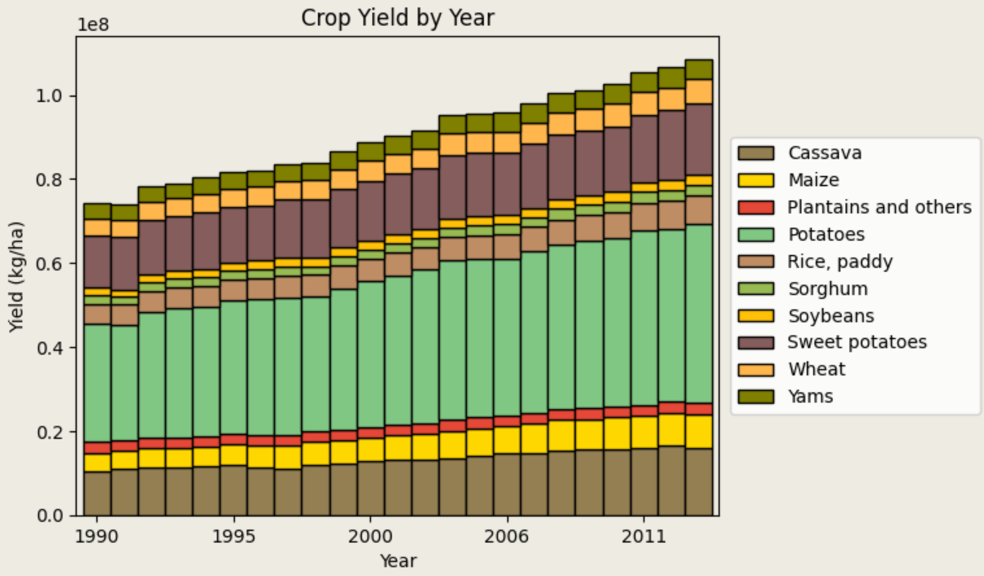
To further explore the correlation and effect of socio-economic and land variables on crop yield, data was gathered from the world bank data and merged with our primary data set.

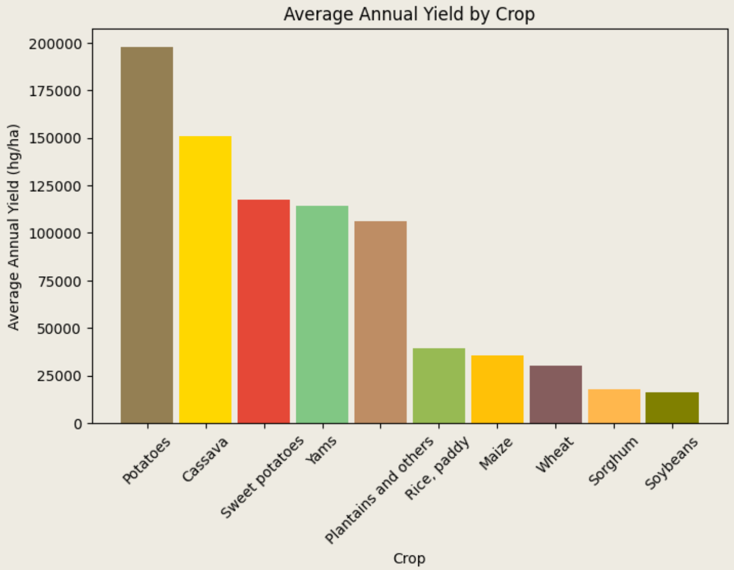
- Birth rate, crude (per 1,000 people) – quantitative
- Population, female (% of total population) – quantitative
- Land area (sq. km) – quantitative
- Urban population (% of total population) – quantitative
- Labor force, total – quantitative
- Fertility rate, total (births per woman) – quantitative
- Life expectancy at birth, total (years) – quantitative
- Survival to age 65, female (% of cohort) – quantitative
- Mortality rate, infant (per 1,000 live births) – quantitative
- Balance of trade (exports – imports as % of GDP) – quantitative
- Population in labor force (% of total population) – quantitative
- Agricultural land (% of total land area) – quantitative
- Forest land (% of total land area) – quantitative
- Net migration (% of total population – quantitative
- Patent applications (per 1,000 people) – quantitative

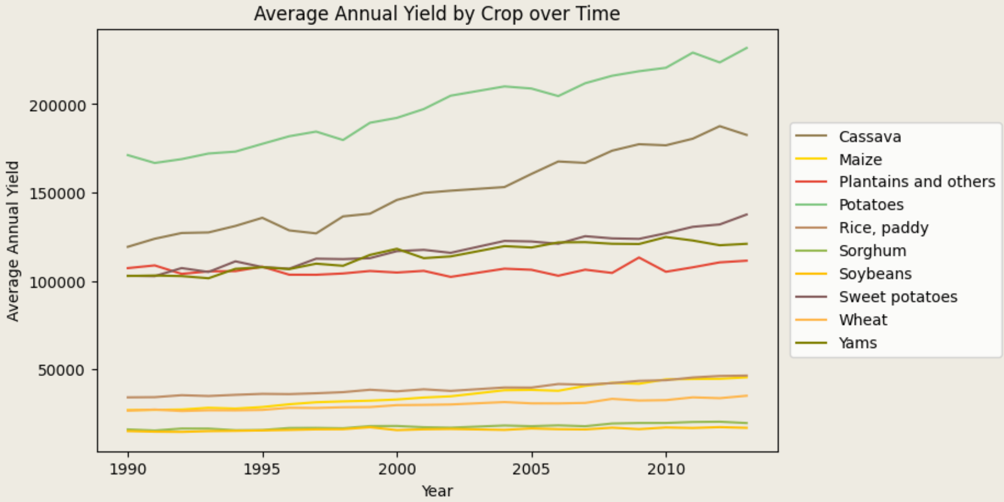
World Bank Data Link

EDA: Understanding Crop Yield









Key Insights

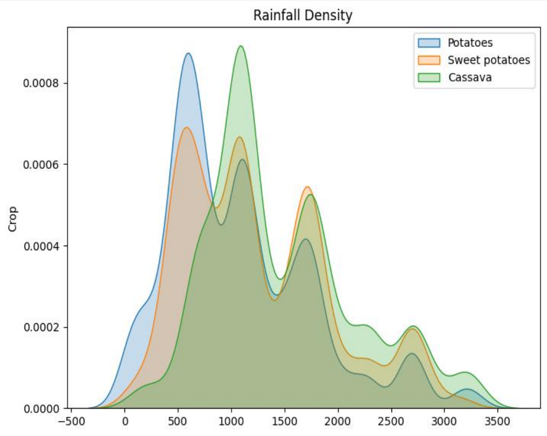
- Potatoes are by far the most-produced crop, followed by cassava and sweet potatoes
- Large countries like India, Brazil, and Mexico produce the most
- Total yield has increased over time
- Most crops show a steady average annual yield over time, with a few increasing trends

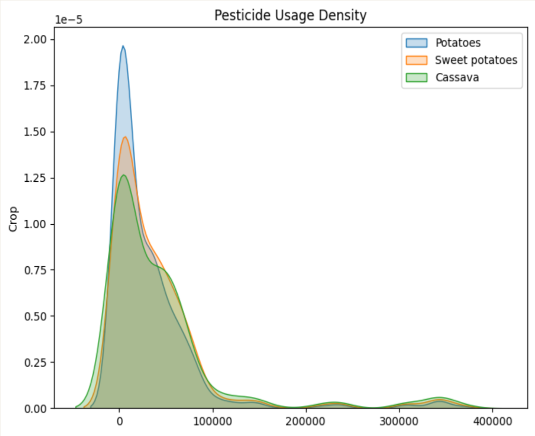
EDA: Exploring the Explanatory Variables

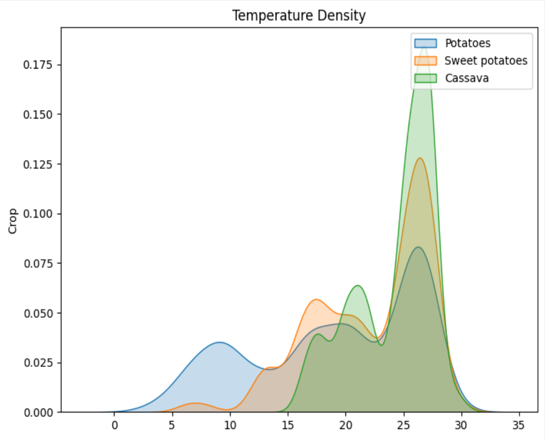
Data Cleaning and Pre-processing Predictor variables

- Total of 20 variables selected for baseline model.
- The correlation between these variables were explored through density plots and correlation matrix

Density Plots for the Top 3 Yielding Crops







Key Insights

- Pesticide density is similar across crops, but rainfall and temperature see significant differences.

Co-variance Matrix for All the Quantitative Variables

Key Insights

- No variables are highly correlated with yield
- Poverty metrics like high fertility and infant mortality are negatively correlated

Baseline: Multiple Linear Regression

- Performed multiple linear regression with 20 variables
- One-hot encoded the two qualitative variables (Country and Crop)

Model Performance on Validation Set	
R2	75.75%
RMSE	50866.34
MAPE	83.13%

Model Diagnostics

Highest 5 Cook's Distances:

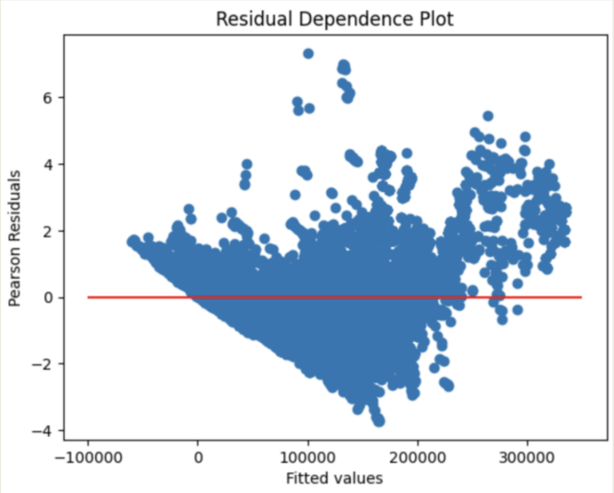
- .00659183
- .00636564
- .00546848
- .00457081
- .0043098

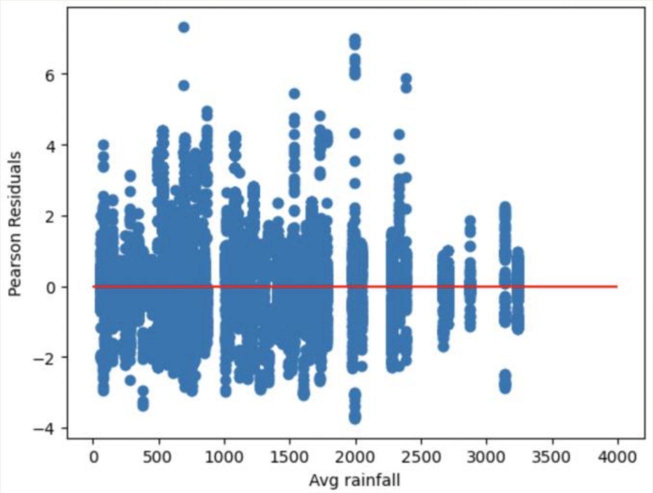
Durbin-Watson Test:
Statistic: 1.20555
Upper bound: 1.53
Lower bound: 1.47
Test failed

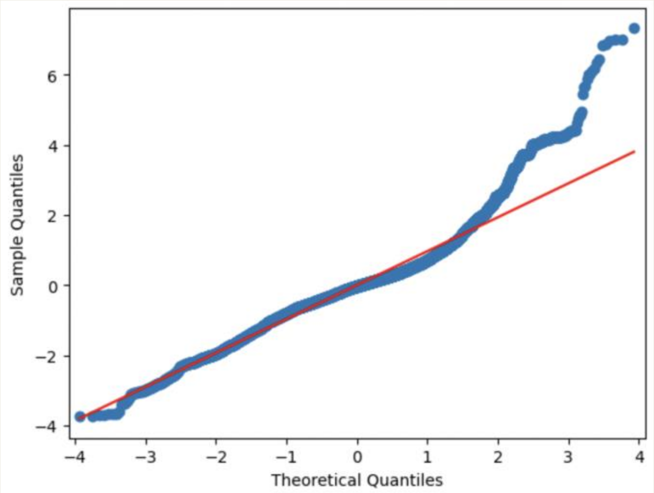
Baseline Model

Why does the baseline model need improvement?

- Insignificant regression variables → 55 out 127 Variables have p-values > 0.05
- Autocorrelation is an issue, as our model failed the Durbin-Watson test
- Residual Assumptions for MLR do not hold → i.i.d., constant variance, and normality assumptions do not hold





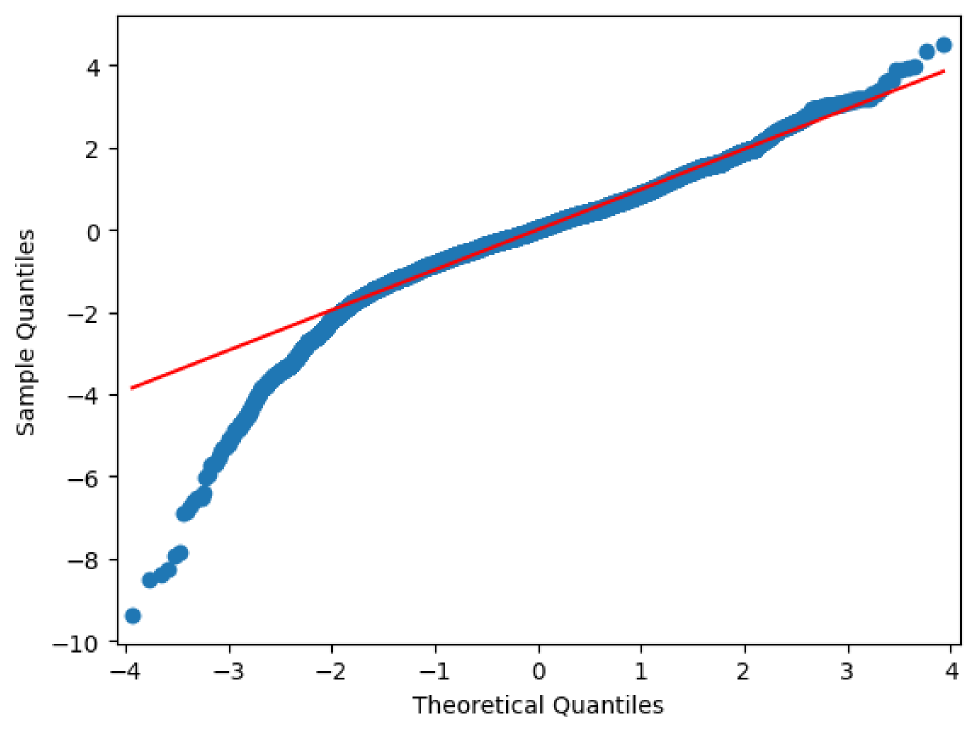


Model 2: Reduced and transformed model

- Re-ran the MLR with only significant variables and a Box-Cox transformation of y variable:

Model Performance on Validation Set	
R2	81.3%
RMSE	52720
MAPE	44.83%

The Box-Cox transformation led to significant improvement in model performance and goodness-of-fit metrics, although we still observed heavy tails in the qq-plot

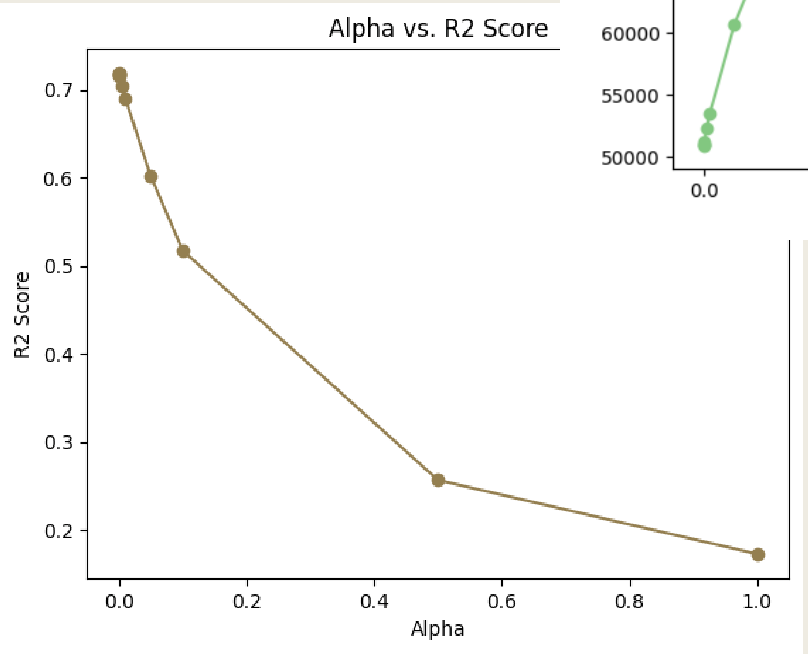


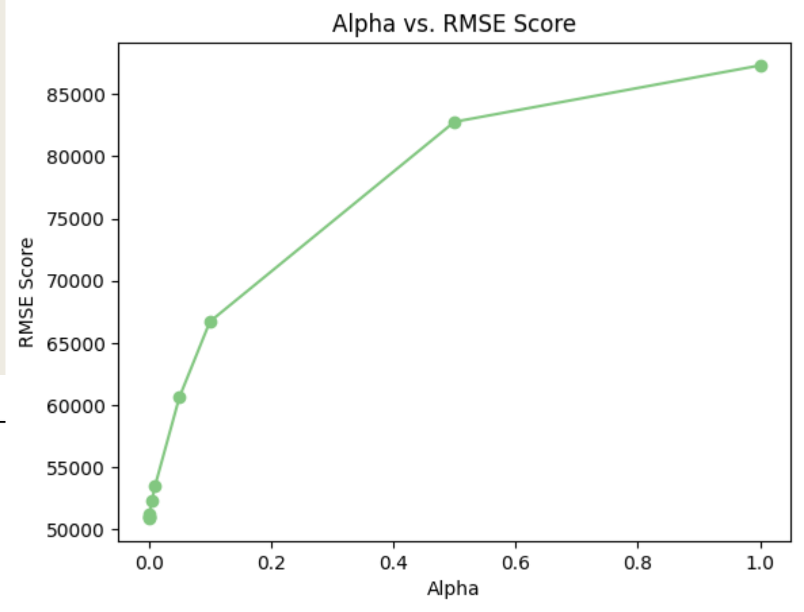
Model 3: Regularized Regression

- Running a regularized regression model on all variables to conduct feature selection:

Model Performance on Validation Set	
R2	73.1%
RMSE	49808.41
MAPE	39.82%

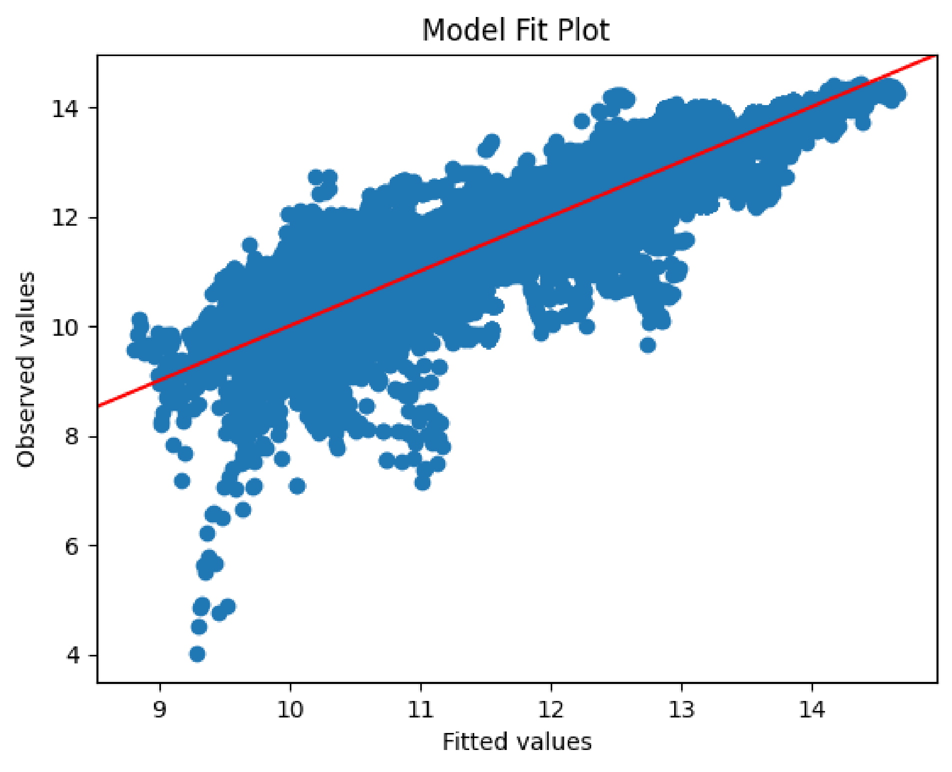
Selected Hyperparameters after Tuning
α = .001
L1 regularization weight = 0.1

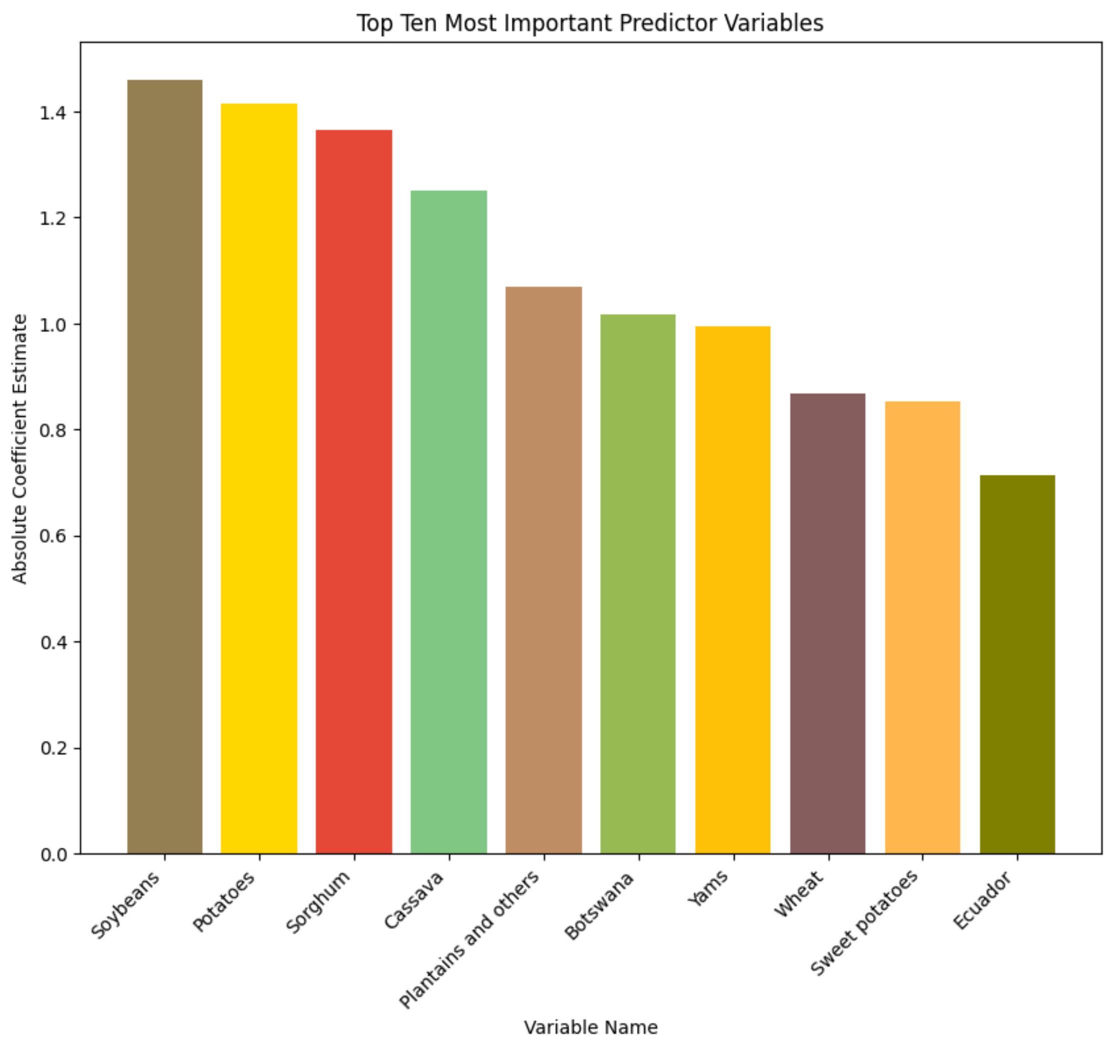




Model 3: Goodness of Fit

- Type of crop is the most important feature
- The regularized, transformed model shows improved goodness of fit

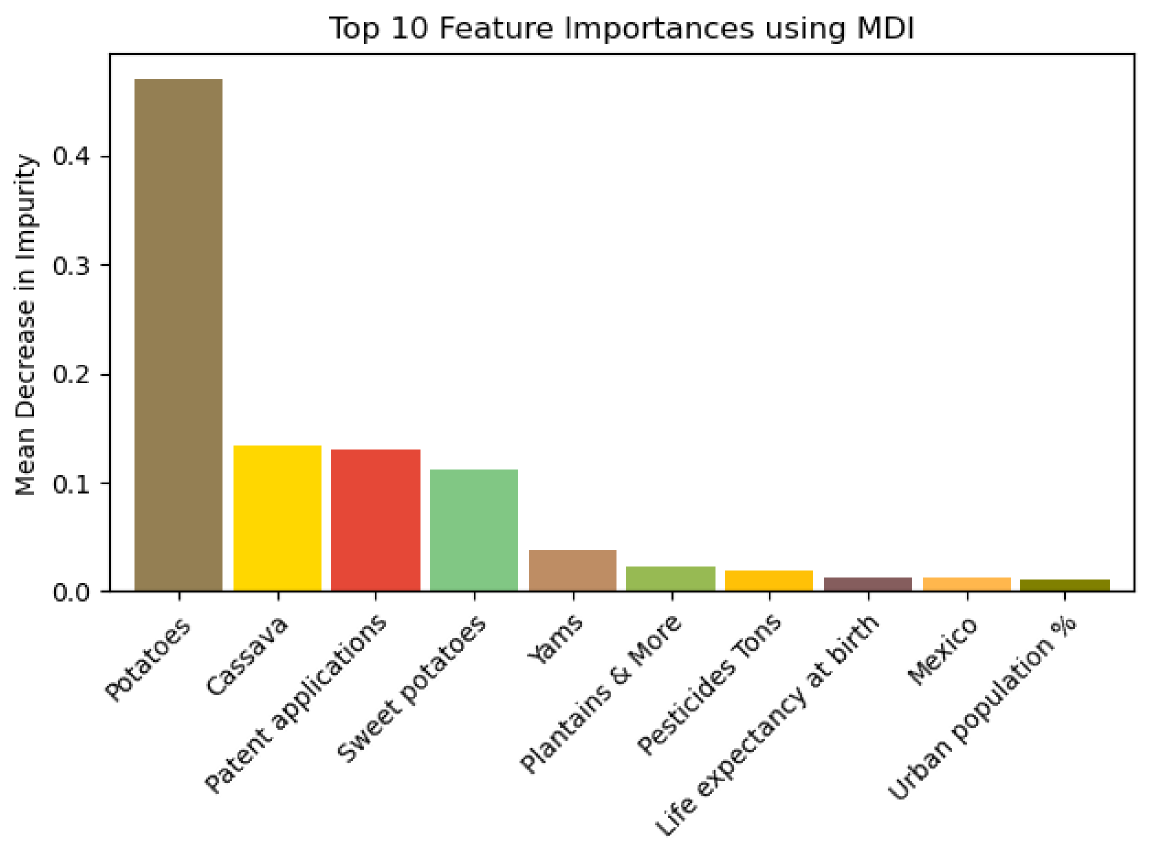




Model 4: Random Forest Regression

- Running a random forest regression model on all variables to conduct feature selection:

Model Performance on Validation Set	
R2	71.8%
RMSE	51026.08
MAPE	47.79%



Comparing Models

	Model 1: Base MLR	Model 2: Reduced MLR	Model 3: Regularized Regression	Model 4: Random Forest Regressor
R2	75.75%	81.3%	73.1%	71.8%
RMSE	50866.34	52720	49808.41	51026.08
MAPE	83.13%	44.83%	39.82%	47.79%

Conclusions

- The regularized regression model with tuned hyperparameters is selected as the most viable model owing to the fact the it minimizes the RMSE and MAPE.

- The metrics for the Random Forest model are comparable to that of regularized regression. However, we select the latter to the fact that it is simpler and easier to interpret.

- In terms of the socio-economic factors we initially set out to explore, there appears to be strong predictive power associated with these non-conventional features such as patent applications, life expectancy and urban population as evidenced by the MDI scores in the random forest model. This is a good starting point to further investigate these relationships using more advanced models.

Thank You

Team 8: Cassidy Gasteiger, Raymond Li, Suraj Shourie, Asadullah Syed, Angela Zheng