SWE3032_41 Artificial Intelligence Project

Author
Jeonghoon Park
2016313844
hoonpk96@gmail.com
https://github.com/Team12-AIProject-Spring2023SKKU/Team12-AIProject-Spring2023SKKU/tree/main

ABSTRACT

We often don't realize it, but the price of food is important to our lives, not just for our wallets, but because it can cause wars between countries.

Worryingly, food production is increasingly volatile due to wars and climate change. We thought that if we could predict future food production to account for these variables, including war and climate change, we could be better prepared.

Our team created a forecasting model for the production of major grains. We created a model to predict rice, wheat, and corn production in the United States. We hope that the results of this model could be used in a meaningful way.

CCS Concepts

• Feature Selection, Prediction, Time-Series Data, EDA, Random Forest

Keywords

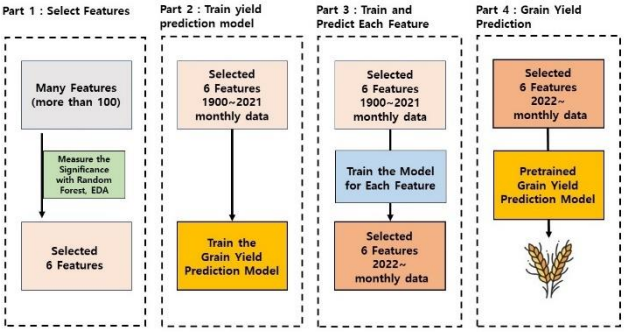
Grain Yield, Climate Indicators, Economic Indicators, Consumption Indicators

1. INTRODUCTION

While there are many AI models for predicting food production, we set out to differentiate our project with the following goals 1) We wanted to create a model that could reflect the complex and rapid changes in modern society by including climate change indicators, economic indicators, and consumer indicators. 2) We wanted to improve the accuracy of our predictive model by identifying which of the many features have a significant impact on food production. 3) We wanted to create a prediction model not only for food production, but also for each input features, so that we could predict the future food production.

Here are some of the challenges we faced during the project. 1) We collected data from various data providers. We had to preprocess the data to make it the same size, which was particularly challenging. 2) We built two models, one to predict features and one to predict grain yield. We created different models and tested them to see which one performed best, which was also challenging.

The overall outline of our project is shown in the figure below. 1) We collected as many features as possible and used Random Forest to select features that have a significant impact on grain yield. 2) We created a Grain Yield Prediction Model with the selected features. 3) We created a prediction model of the selected features and create predicted values. 4) Predicted future food production with the predicted values of the selected features.



2. Methodology

2.1 Data Collection

Our team collected over 100 features from a variety of data providers. We pulled data from the United States Department of Agriculture (USDA), NASA, National Weather Service, Food and Agriculture Organization of the United States (FAO), and the New York Stock Exchange (NYSE). The main features we collected include fertilizer prices, oil prices, land area under grain production, precipitation, sunshine, temperature, humidity, population, grain consumption, vegetable consumption, and meat consumption.

2.2 Data Preprocessing

We wanted to collect as much data as possible by month from 1900 to 2021, but some data was available by day, some by year, and some has null values. Here's how we handled that data.

2.2.1 Formatting data by month

Climate data such as sunshine, precipitation, temperature, and humidity were available on a daily basis, and since the United States is a large country, it was a matter of which region to base the data on. First, We found the maximum, minimum, and average of each climate data point to unify the format across months. Luckily, the grain production data was available by state

in the U.S., so We averaged the climate data back down to the state level.

2.2.2 Handling Null Data

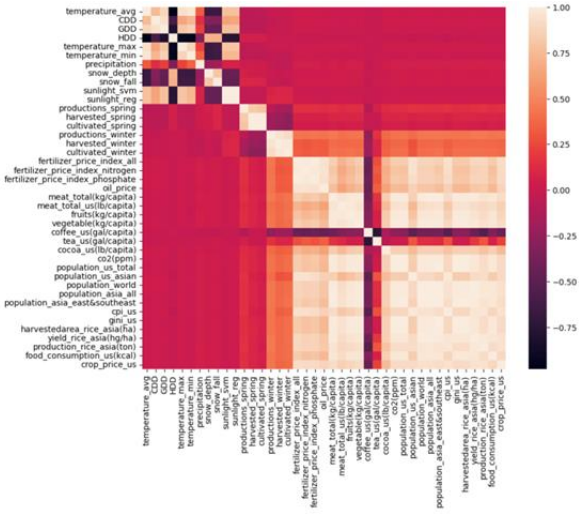
Not all data was available from 1900 to 2021. In many cases, we did not have data from 1900 to 1930, so we used SVM or Linear Regression to fill in the values. In the case of Linear Regression, we used it after verifying that the data appeared to be significantly linear when visually inspected.

2.3 Feature Selection using Random Forest and EDA

We wanted to select a few features for several reasons 1) If we simply train the model with all the features, there is a chance that the interactions between the features will reduce the performance of the model. 2) Since we will be making separate predictions for each feature to predict future grain yields, we can't build a model for every feature.

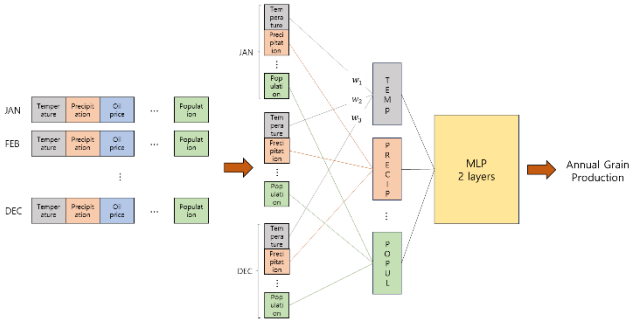
By training a Random Forest model to predict grain production, we can measure which features have a significant impact on the target value. Visual analysis was also performed using EDA.

In the case of Random Forest, the feature importance is calculated as the average of the feature importance of each ensembled decision tree. The MSE value was used as the criterion for each decision tree. In other words, the feature importance was judged by how much each feature reduced the MSE.



2.4 Grain Yield Prediction Model

The grain production prediction model was designed in the following way. 1) Sequential data of the same class were filtered and weighted by Weighted Sum to give weight to time-critical data. 2) The weighted data of each class were used as input to MLP to learn the nonlinear interactions on grain yield. K Fold Cross Validation was used for reliable performance evaluation. The structure of the model is shown in the image below. This process was done for rice, corn, and wheat.



2.5 Feature Prediction Model

When making predictions for features, we first split the data into climate-related data and other data. This is because climate-related data is time series data. Then, we tried several models for each feature and selected the most appropriate model.

For climate-related features, we created five models: RNN, LSTM, GRU, Ensemble (RNN, LSTM, GRU), and Ensemble (LSTM, GRU) to see which model performed the best.

For consumption and economic features, we applied LSTM and ARIMA, a statistical model, and checked which model performed well.

When creating the model, the input data was simply the year and month, and the target value was the value of the feature. This was not a high-dimensional prediction with many variables, but it was enough to identify trends over time and make meaningful predictions.

3. Results

3.1 Feature Selection with Random Forest

Random Forests were run for rice, corn, spring wheat, and winter wheat, respectively. The top six most influential Features for each crop were as follows. Among several features, we excluded plowed area and harvested area because it is self-evident that they have a direct impact on grain production.

1) Rice: population, Asian rice production, US consumer price index, US crop prices, oil prices, and US food consumption.

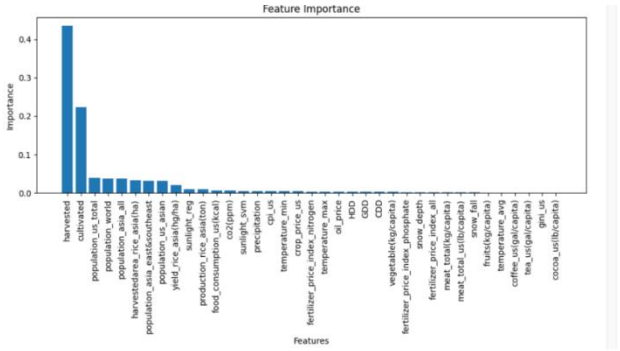
2) Corn: population, sunshine, Asian rice production, U.S. food consumption, carbon dioxide concentration, precipitation

3) Spring wheat: population, Asian rice production, crop price, oil price, US food consumption, sunshine

4) Winter wheat: population, Asian rice production, fertilizer prices, oil prices, sunshine, carbon dioxide concentration

Among these features, we selected 6 recurring features: population, US consumer price index, oil price, sunshine, US food consumption, and carbon dioxide concentration.

Below is one example of graph showing the importance of the features.



3.2 Feature Prediction Model

Our team selected six features: Population, US consumer price index, oil price, sunshine, US food consumption, and carbon dioxide concentration. We experimented with several models to make predictions for each feature.

As a result, the selected models are as follows.

1) Population: LSTM

2) US Consumer Price Index : LSTM

3) Oil Price : Ensemble(RNN, LSTM, GRU)

4) Sunlight : GRU

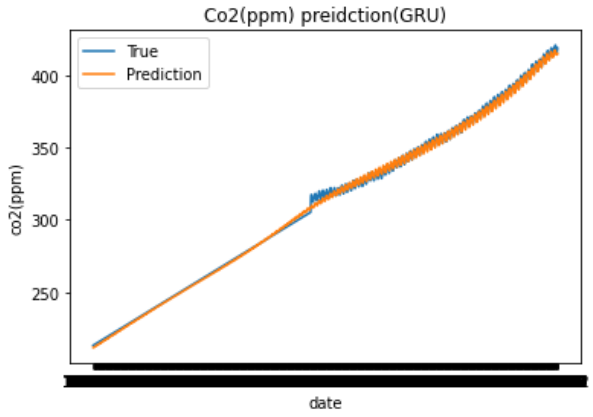
5) US Food Consumption : LSTM

6) CO2 Concentration : GRU

The overall performance results of the different models are shown in the following table. Test Loss was measured.

Model	Population	US CPI	Oil	Sun light	Food Consume	co2
LSTM	0.13	0.52	3145	797	3.3	2.24
GRU	-	-	3161	767	-	2.05
RNN	-	-	4717	4882	-	3.77
ARIMA	3.7	4.2	-	-	7.6	-
Ensemble(RNN,LSTM, GRU)	-	-	1531	794	-	2.20
Ensemble(LSTM, GRU)	-	-	2829	786	-	3.39

The image below shows a graph of the GRU model predictions for Co2 Concentration. In this way, we were able to create predictions for each feature.

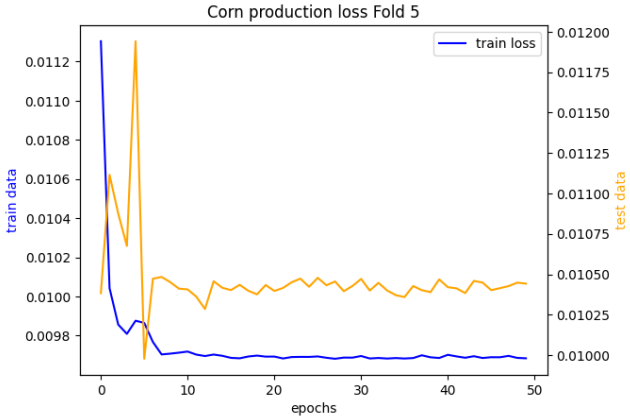


3.3 Grain Yield Prediction Model

We created models to predict yields for spring wheat, winter wheat, rice, and corn, respectively. The models were trained on 1900-2021 data for the six features we selected earlier. Below is the performance of each model. We used K Fold cross validation to measure the performance of the models, so the Test Loss is expressed as a range.

	Spring Wheat	Winter Wheat	Rice	Corn
Test Loss	0.0092 ~ 0.0115	0.0035 ~ 0.0043	0.006 ~ 0.0084	0.0105 ~ 0.0116

The image below shows the Training Loss values for the Corn Prediction Model over epochs.

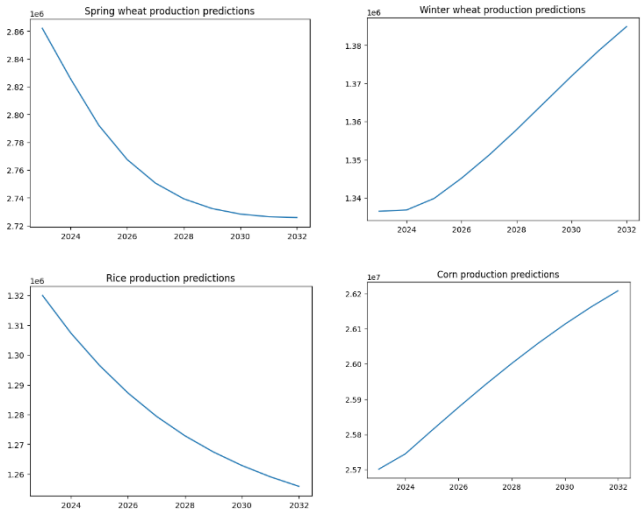


4. Result and Analysis

4.1 Make Predictions

With the models and data we have prepared so far, we actually predict how grain production will change over the next 10 years. Here are the projections for the next 10 years for each grain.

Predictions	Spring Wheat	Winter Wheat	Rice	Corn
2022	2861981	1336457	1319915	25701286
2027	2739279 (−4.2%)	1357848 (+1.6%)	1272858 (−3.5%)	26001388 (+1.1%)
2031	2725797 (−4.7%)	1384889 (+3.6%)	1255904 (−4.8%)	26207812 (+1.9%)



Here are just a few impressive metrics. Spring Wheat and Rice are forecast to decline rapidly in production over the next five years, followed by a slow decline after five years. Winter Wheat and Corn are forecast to rise steadily.

However, overall grain production is forecast to decline due to a large decline in Spring Wheat, which is the largest component of all grains.

This is not good news from the perspective of humanity's sustainability on the planet, as the world's population is forecast to continue to grow by 0.8% per year over the next decade. [1]

4.2 Feature Selection

Before we started this project, we thought that climate-related data would have the biggest impact on production, especially precipitation. But even more important features than precipitation were economic indicators like oil prices and consumer indicators like CPI and food consumption. An important piece of climate-related data was sunshine and co2 concentration.

Not surprisingly, oil prices have a significant impact on production. Oil is ubiquitous in modern agriculture: it powers tractors for harvesting and sowing, and light aircraft for spraying pesticides.

Surprisingly, CPI and food consumption were also very important data points. I suspect this is because producers increase or decrease production in response to demand over a longer time horizon. Obviously, the US government will subsidize producers based on demand.

The reason why sunshine is more important than precipitation is due to better infrastructure. Modern crop yields seem to be much less dependent on precipitation. However, there is no way to make up for the lack of sunlight. Therefore, the influence of sunlight seems to be higher.

4.3 Grain Yield Prediction Model

We wanted to see how accurate these projections we made were. Unfortunately, we could only find a few papers that provided longer-term projections of 10 years or so.

1) There was a paper from NASA that projected agricultural production under climate change. It predicted a 24% decrease in corn production and a 17% increase in wheat production in 2030. The main drivers of these changes were temperature and CO2 concentration. The increase in wheat production was attributed to the fact that warmer temperatures increase the area available for wheat production, which is not accounted for in our model. [2]

2) While it was difficult to find examples of using AI models to predict U.S. crop production, we were able to find examples of simply training and testing on historical data. In the case of training and testing Corn Yield using Ensemble CNN-DNN, the RMSE value was around 8.5~9%, but it is difficult to make a simple comparison. [3]

3) USDA's 2023 wheat production forecast is down 3% from 2022. Our projections show that combined spring and winter wheat production will be 0.8% lower in 2023 than in 2022. The trend is likely to be similar. [4]

5. References

[1] World Population Statistics :

https://www.macrotrends.net/countries/WLD/world/population-growth-rate

[2] https://climate.nasa.gov/news/3124/global-climate-change-impact-on-crops-expected-within-10-years-nasa-study-finds/

[3] Shahhosseini, Mohsen, et al. "Corn yield prediction with ensemble CNN-DNN." *Frontiers in plant science* 12 (2021): 709008.

[4] https://www.ers.usda.gov/topics/crops/wheat/market-outlook/