SWE3032\_41 Artificial Intelligence Project

Author

Jeonghoon Park  
2016313844

hoonpk96@gmail.com

**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 coulb be used in a meaningful way.

인지하지 못할 때가 많지만 식량 가격은 우리들 삶에 중요하다. 왜냐하면 단순히 우리 지갑 사정이 팍팍해지는 것을 넘어 국가간 전쟁의 원인이 될 수도 있기 때문이다.

우려스럽게도 식량 생산량의 변동성은 전쟁과 기후 변화로 날로 커지고 있다. 전쟁과 기후변화를 포함한 변인들을 반영하여 미래 식량 생산량을 예측할 수 있다면 미리 대비할 수 있으리라는 생각했다.

우리 팀은 주요 곡물 생산량에 대한 예측 모델을 만들어 보았다. 미국의 쌀, 밀, 옥수수 생산량 예측 모델을 만들었다. 이 모델의 결과값들이 의미 있게 쓰여 지기를 바란다.

**CCS Concepts**

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

**Keywords**

Grain Yield, Climate Indicators, Economic Indicators, Consumption Indicators

# 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.

텍스트, 스크린샷, 폰트, 직사각형이(가) 표시된 사진

자동 생성된 설명

우리 프로젝트의 전체적인 개괄은 아래 그림과 같다. 1) 가능한 많은 Feature들을 모아서 곡물 생산량에 큰 영향을 미치는 Feature를 Random Forest를 사용하여 선택했다. 2) 선택한 Feature를 가지고 Grain Yield Prediction Model을 만들었다. 3) 선택한 Feature들의 예측 모델을 만들고 예측 값들을 만들었다. 4) 선택한 Feature들의 예측값을 가지고 미래의 식량 생산량을 예측했다.

# Methodology

## 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.

우리 팀은 다양한 데이터 제공자로부터 100가지가 넘는 Feature들을 수집했다. USDA(United States Department of Agriculture), NASA, National Weather Service, FAO(Food and Agriculture Organization of the United States), NYSE(New York Stock Exchange)로부터 데이터를 가져왔다. 수집한 주요 Feature로는 비료 가격, 유가, 곡물 생산 토지 면적, 강수량, 일조량, 온도, 습도, 인구수, 곡물 소비량, 채소 소비량, 육류 소비량 등이 있다.

## 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.

우리는 가능한 많은 데이터를 1900년부터 2021년까지 월별로 모으고자 했다. 그렇지만 일별, 연별로 제공되는 데이터도 있었으며 계측치가 없는 데이터들도 있었다. 해당 데이터를 핸들링 한 방법을 소개하겠다.

### **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.

일조량, 강수량, 온도, 습도 같은 기후 데이터들은 일별로 데이터가 있었다. 그리고 미국이 넓기 때문에 어느 지역을 기준으로 할 지도 문제였다. 우선 월별로 형식을 통일하기 위해서 각 기후 데이터들의 최댓값, 최솟값, 평균을 구했다. 우리가 구한 곡물 생산량 데이터가 미국의 주별로 곡물 생산량을 나타낸 데이터였기 때문에 해당 데이터들을 주별로 정리했다.

### **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.

모든 데이터가 1900년부터 2021년까지 갖추어져있지는 않았다. 1900~1930년까지의 데이터가 없을 때가 많았는데 이런 데이터들은 SVM이나 Linear Regression을 사용하여 값을 채웠다. Linear Regression의 경우 해당 데이터를 시각적으로 확인했을 때 유의미하게 선형으로 나타나는 지 확인한 뒤 사용하였다.

## 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.

텍스트, 스크린샷, 라인, 직사각형이(가) 표시된 사진

자동 생성된 설명

많은 Feature들 중 몇 가지를 선택하고자 한 이유가 있었음. 1) 단순히 모든 Feature를 가지고 모델을 학습시키면 Feature들 끼리의 상호작용으로 모델의 성능을 떨어트릴 가능성이 있다. 2) Feature들에 대한 예측을 따로 진행하여 미래 곡물 생산량을 예측할 것이기 때문에 모든 Feature들에 대해서 모델을 만들 수는 없다.

곡물 생산량을 예측하기 위한 Random Forest 모델을 학습 시키면 어떤 Feature가 Target Value에 중요한 영향을 미치는 지 측정할 수가 있음. EDA를 진행하여 시각적인 분석도 함께 진행하였음.

Random Forest의 경우 Ensemble한 Decision Tree 각각의 Feature Importance의 평균으로 Feature Importance를 계산함. 각각의 Decision Tree의 Criterion으로는 MSE 값을 사용하였다. 즉 Feature Importance는 각 Feature 가 MSE를 얼마나 감소 시켰는 지로 판단하였다.

## 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.

곡물 생산량 Prediction 모델은 같은 class의 Sequential 한 데이터를 필터를 거쳐 Weighted Sum 함으로써 시간상 중요한 데이터에 가중치를 부여하였다. 가중치를 부여한 각 class의 데이터들을 MLP의 input으로 사용하여 쌀 생산량, 옥수수 생산량, 밀 생산량에 미치는 비선형적 상호작용을 각각 학습하였다. K Fold Cross Validation을 사용하여 신뢰성 있는 성능 평가를 진행하였다. 이 과정을 쌀, 옥수수, 밀에 대해 각각 진행하였다. 모델의 구조는 아래 이미지처럼 나타낼 수 있다.

텍스트, 스크린샷, 사각형, 직사각형이(가) 표시된 사진

자동 생성된 설명

## 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.

기후 관련 Feature들에 대해서는 RNN, LSTM, GRU, Ensemble(RNN, LSTM, GRU), Ensemble(LSTM, GRU) 총 5 가지의 모델들을 만들어 어떤 모델이 가장 좋은 성능을 보이는지 확인하였다.

소비나 경제 관련 Feature 들에 대해서는 LSTM과 ARIMA라는 통계적인 모델을 적용해 보았고 어떤 모델이 좋은 성능을 보이는 지 확인하였다.

모델을 만들 때의 input 데이터는 단순히 연-월을 넣었고 target value는 해당 Feature들에 대한 값을 넣었다. 이렇게 했을 때 여러가지 변수들을 이용하여 가치 있는 예측을 할 수는 없었지만 시간에 대한 경향성을 파악할 수는 있었다.

# Results

## 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.

텍스트, 스크린샷, 라인, 평행이(가) 표시된 사진

자동 생성된 설명

## 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.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

## 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.

우리는 봄밀, 겨울밀, 쌀, 옥수수에 대해 각각 생산량을 예측하는 모델을 만들었다. 이 모델은 앞에서 선택한 6가지 Feature들의 1900~2021 데이터를 학습하여 만들었다. 다음은 각 모델들의 성능들이다. K Fold cross validation을 사용하여 모델의 성능을 측정하였기 때문에 Test Los를 범위로 표현하였다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

# Result and Analysis

## 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%) |

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

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]

## 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.

이 프로젝트를 시작하기 전에는 기후 관련 데이터들이 생산량에 가장 큰 영향을 미칠 줄 알았다. 특히 강수량이 매우 중요할 것으로 생각했다. 그렇지만 강수량보다도 더 중요한 feature는 유가와 같은 경제 지표와 CPI, 음식 소비량 같은 소비자 지표들이었다. 기후 관련 데이터에서 중요한 데이터는 일조량과 CO2 농도였다.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명유가가 생산량에 중요한 영향을 미치는 것은 당연해 보인다. 수확이나 파종을 할 때 사용되는 트랙터, 농약을 살포할 때 사용되는 경비행기 운용 등 현대 농업에서 석유가 사용되지 않는 곳은 거의 없다.

의외로 CPI와 음식 소비량 또한 매우 중요한 데이터였다. 시간 범위를 길게 했을 때 생산자들도 수요에 따라 생산량을 늘리거나 줄이기 때문인 것으로 추측된다. 분명 미 정부에서도 수요량에 따라 생산자들을 보조할 것이다.

강수량 보다도 일조량이 중요한 이유로는 관계 시설이 잘 돼 있기 때문으로 추측된다. 현대 작물의 생산량은 강수량 의존도를 많이 낮춘 것으로 보인다. 그렇지만 햇빛이 부족한 것은 보강할 방법이 없다. 그래서 햇빛의 영향력이 더 높은 것으로 보인다.

## 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. [1]

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. [2]

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.

우리가 만든 이 예측 값들이 얼마나 정확한 지 확인해보고 싶었다. 아쉽게도 10년 정도 되는 장기적인 예측 값을 제시하는 논문은 몇 가지 밖에 찾을 수 없었다.

1) NASA에서 기후 변화에 따른 농업 생산량을 예측한 논문이 있었다. 이 논문에서는 2030년에 Corn의 생산량이 24% 감소할 것으로 예측했으며 밀 생산량은 17% 증가할 것으로 예측했다. 이 변화량의 주요 원인은 온도 변화와 Co2 농도가 지목됐다. 밀 생산량 증가는 온도가 높아지면서 밀을 생산할 수 있는 면적이 넓어지기 때문으로 설명했는데 이는 우리 모델에서 고려하지 못한 부분이다. [1]

2) 인공지능 모델을 사용하여 미국 작물 생산량을 예측한 사례를 찾기는 어려웠지만 단순히 과거 데이터를 학습시키고 테스트해본 사례를 찾아볼 수는 있었다. Ensemble CNN-DNN을 사용하여 Corn Yield를 학습시키고 테스트한 사례에서 RMSE 값이 8.5~9% 정도로 나왔다. 그렇지만 단순 비교하기는 힘들 것으로 보인다. [2]

3) USDA에서 2023년 밀 생산 예측량을 2022년에 비해 3% 감소할 것으로 봤다. 우리의 예측치로 계산했을 때 봄 밀과 겨울밀을 합쳐 2023년은 2022년보다 생산량이 0.8% 감소할 것으로 나왔다. 추세는 비슷하다고 할 수 있을 것 같다.

# REFERENCES

1. world population statistics:

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

1. https://climate.nasa.gov/news/3124/global-climate-change-impact-on-crops-expected-within-10-years-nasa-study-finds/
2. Shahhosseini, Mohsen, et al. "Corn yield prediction with ensemble CNN-DNN." *Frontiers in plant science* 12 (2021): 709008.
3. Winter Wheat 예측 <https://www.ers.usda.gov/topics/crops/wheat/market-outlook/>