Forecasting Corn & Soybean Yield/Analog Years

PURDUE UNIVERSITY
The Data Mine

Team Member Names: Ayush S. Arora, Matthew J. Tolla, Sathvik Hegde, Vikash Gunalan, Hashim Khan, Prisha Cocasse, Chung-Yuan Liang

BASF
We create chemistry

1 Introduction

BASF creates chemistry for a sustainable future by combining economic success with environmental protection and social responsibility. With publicly available data they tasked us with digitizing a model to help farmers forecast analog years for corn and soybeans for 2024.

Objectives:

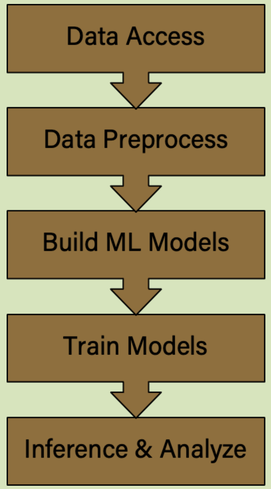
1. Create a 2024 forecast of corn & soybean yield in Indiana, Iowa, and Illinois and determine likeliest analog years based on bioclimatic factors.
   - Analog years are the most similar years in the past to the predicted year based on weather factors.
2. Create a digital means for interacting with the platform to help farmers

Criteria and Constraints:

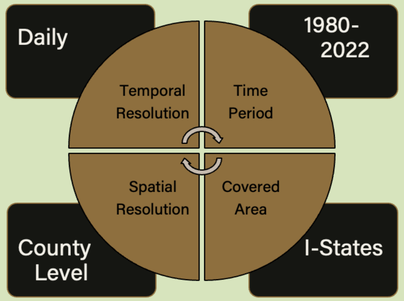
1. Use open-source climate and soil data
2. Accuracy: Weightage of each input factor has on farming
3. Variability: Changing environmental conditions

2 Methodology

Working Flowchart:



Data Access:



Attained data for various weather factors from open-source data based on the criteria above:

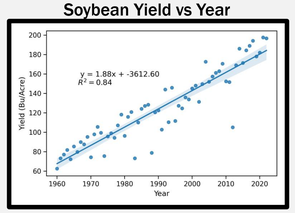
- Temperature (avg, min, max)
- Precipitation
- Solar radiation
- Vapor pressure
- Growing Degree Days (GDD)

Preprocessing:

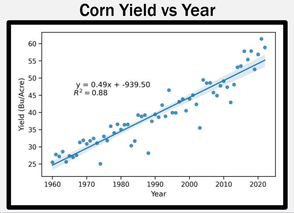
Linear Regression Models

We used Linear Regression to determine the trend of the yield throughout history.

Soybean Yield vs Year



Corn Yield vs Year



Random Forest (RF) Models

- A RF model itself does not account for the temporal relationships.
- The input data of the RF models are monthly and weekly features in the target year and annual features in the previous 3 years (including the target year).
- The number of overall features is 276.
- Hyperparameters:
  - Number of estimators = 400
  - Maximum depth = 12

Merged features in Anvil: R & Python

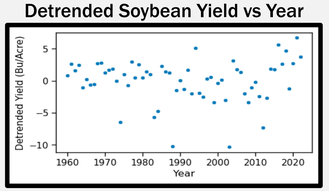
```
df = read.csv('./MergedDataIN.csv')
head(df)
```

	year	Temp_Avg_high_Jan	Temp_Avg_high_Feb	Temp_Avg_high_Mar	Temp_Avg_high_Apr
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	1980	36.000	32.931	44.903	60.067
2	1981	32.645	40.643	50.032	67.500
3	1982	28.581	33.679	49.419	57.533
4	1983	36.258	42.857	51.419	56.200
5	1984	29.677	45.069	39.871	58.633
6	1985	28.323	33.821	54.032	67.767

3 Results

Analog Year Classification

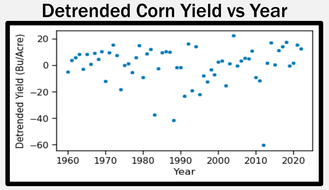
Detrended Soybean Yield vs Year



Removes the effect that increasing technology capabilities has on yield

- Focus solely on weather factors

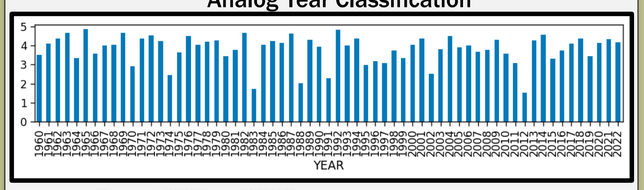
Detrended Corn Yield vs Year



To remove trend of the yield throughout history, we detrended data

- Model Accuracy: R^2= 0.88

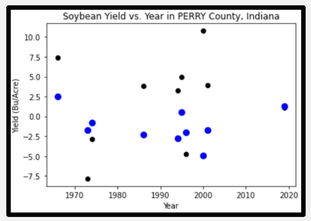
Analog Year Classification



From 1960 to 2022:

- The best years: 2020, 2016, 2004, 1994, 1985
- The worst years: 2012, 1988, 1983, 1974, 1967

Accounting For Error



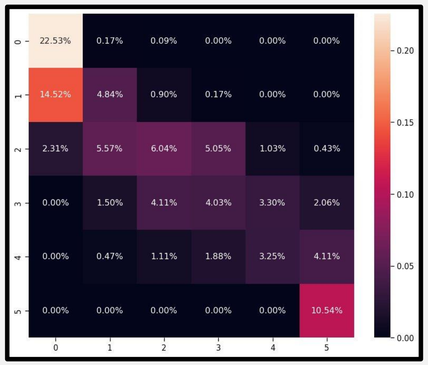
Random forest model for predicting soybean yield

- Very low accuracy/R^2 value
- Black dots = test data
- Blue dots = prediction

Error caused by using a bad model

How can we improve this?

The confusion matrix of predicting analog year by the RF model



- The confusion matrix shows that the predictions from the RF model do not largely deviate from the expectation (the diagonal).
- Evaluation
  - Training
    - R^2 = 0.95
    - RMSE = 0.06
  - Testing
    - R^2 = 0.55
    - RMSE = 0.96
- The RF classifier is used to predict the analog year, but the analog year is categorized by levels. Thus, RMSE that evaluate the continuous difference is more important.

4 Conclusions

- This project develops a framework to load and process data in high resolutions (county-level and daily) automatically. With this framework, the models can be updated easily in the future.
- The linear regression models are applied indicate and remove the trends of yields along year.
- The random forest classifier performs well in predicting the analog year of yields based on the climate features in the target year. Through this model, farmers can get expected yields within 10% errors.

References

5 Future Goals

The model cannot predict the analog yield in the target year, but we must obtain the climate features in the present year. Coupling with models for climate feature predictions is necessary for this framework. An alternative is to train another model that can predict the analog year only depending on the historical data. A user-friendly interactive web interface will also be important to share this useful information with farmers.

Acknowledgements

Thank you to following people for their expertise and guidance throughout our project.

- Dr. Dan Quinn, Agronomy Department, Purdue
- Gregory Ury, Seed Agronomist, BASF
- Emily L Hoeing, Corporate Partners Advisor, The Data Mine
- Cai Shun Chen, Corporate Partners Technical Specialist, The Data Mine
- Justin J Moritz, Product Manager, BASF
- Sai Shashank Mukkera, Teaching Assistant/Project Lead, The Data Mine

The Data Mine Corporate Partners Symposium 2024