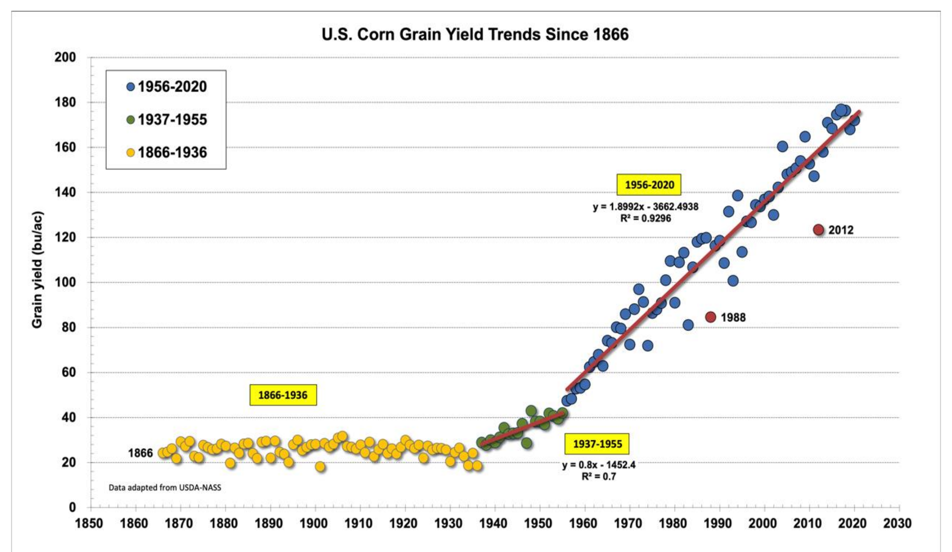
The Effect of Climate Change on the Yield of Maize

PURDUE UNIVERSITY — The Data Mine

Malak Amireh, Archis Dhar, Kosei Dohi, Maggie Garg, Abhilash Gogineni, Lexie Hvostal, Julianne Iaccarino, Ameya Jumde, Josh Knull, Wilson Lee, Yash Mehta, Saravana Polisetti, Rohan Rajesh, Matt Schnell, Aditya Tripathy, Sara Xiao, Bing Yuan

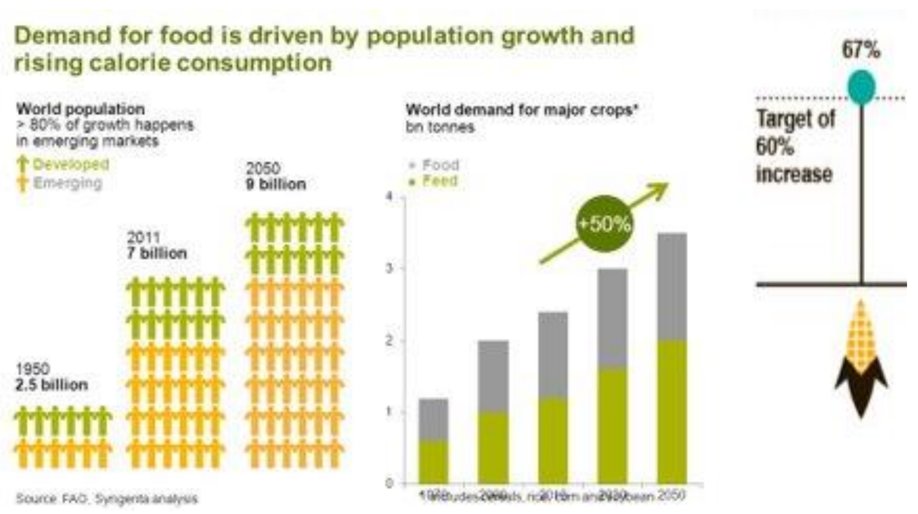
BAYER

Problem Scoping

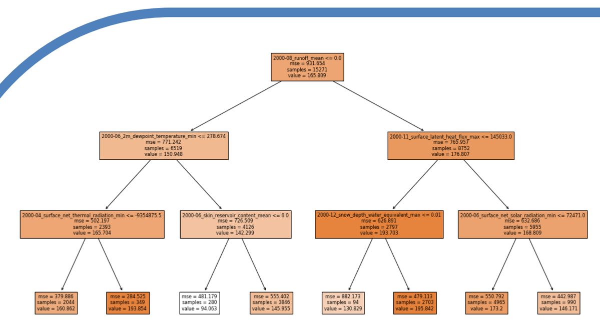


Scope
The plot on the left demonstrates the trend of US Corn Yield over the last century, noting a drastic increase in yield since pivotal advancements in hybrid breeding in the 1930s. Corn accounts for 95% of total feed grain production and use in the US; and it is estimated that global corn production will need to double to meet the demand of 2050.

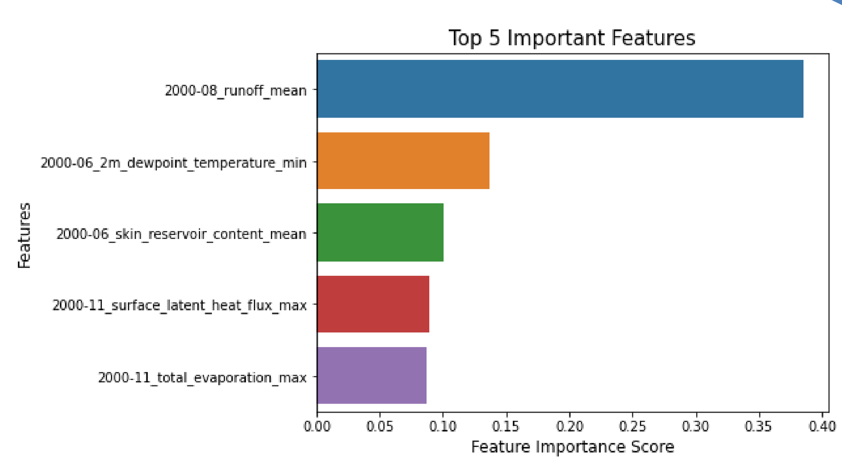
Project Goal
Construct machine learning models and conduct causal inferencing to provide seed product development and testing operations with insights on climate impact on predicted crop growth.



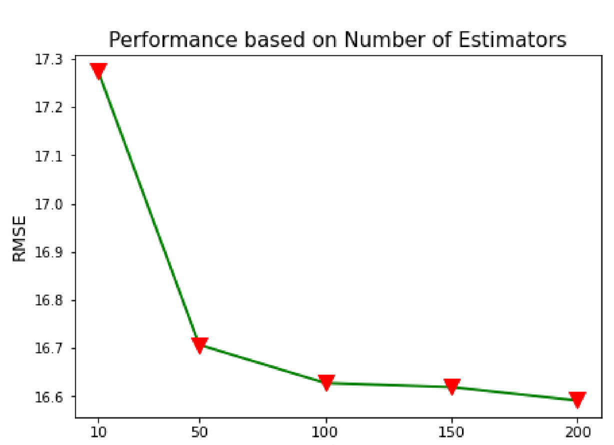
Yield Prediction



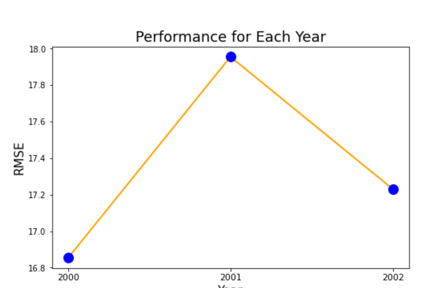
1. Algorithm choice - **Random Forest Regressor.** Takes ensemble of multiple boosted decision trees (estimators).

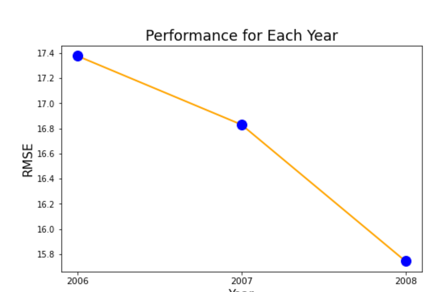


3. Feature importance- Ranked the relative feature importance and graphed the 5 most important ones. The most important feature is runoff.



2. Hyperparameter tuning- RMSE score based on the number of estimators (decision trees). The optimal number is around 100 based on speed and performance.





4. Training for multiple years - Trained a random forest regressor for different years and plotted performance for each year. Future work would consist of creating an ensemble of each regressor.

Data Methodology

Genotypic data of Maize varieties were paired with several sources of abiotic stress. These abiotic stressors consisted of weather, air quality, and soil information.

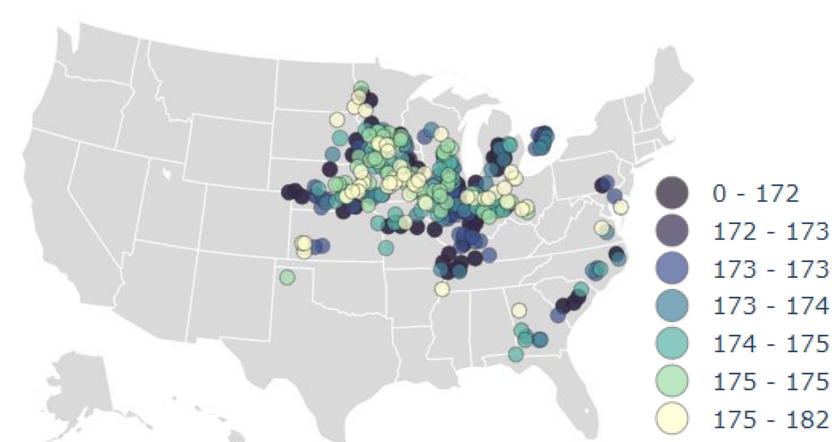
- Incomplete genomic data
- Beagle imputation for many of the missing values

- ERA5 hourly weather estimates used to replace historical NOAA data
- Resolution up to 0.1° x 0.1° grid

- Downloaded from EPA database
- Provided by the prior student Bayer team

- Downloaded from EPA database
- Also provided by the prior student Bayer team

Maps

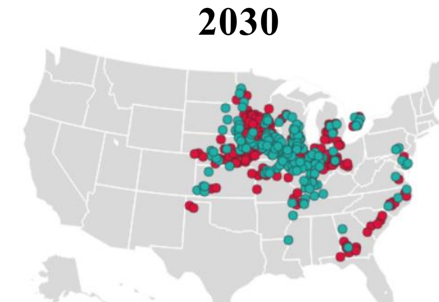


2022 Predicted Yield Distribution (bushels)

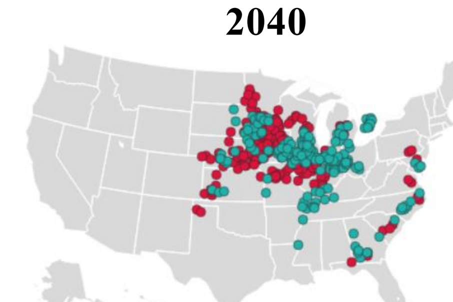
- One strain randomly selected for comparison
- Predicted yield using data
- Size denotes size of yield
- Overlapping circles denote multiple plantings at the location
- 7 colors for 7 quantiles; brighter color means higher yield
- Overall mean is 174.5

- All maps use 12 CMIP6 Scenario 5 weather-related variables predicted by CAS (8), MRI (1), NOAA-GFDL (3)
- Soil and air quality information are assumed to be the same as in 2008
- Green values represent those higher than the baseline mean (174.5) of the 2022 prediction
- Red values represent those lower than the baseline mean (174.5) of the 2022 prediction

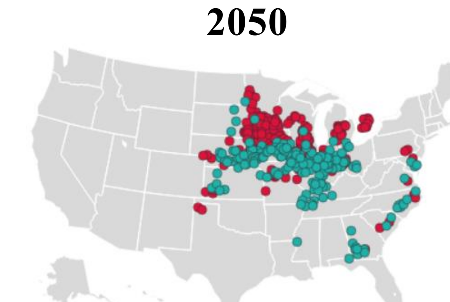
2030



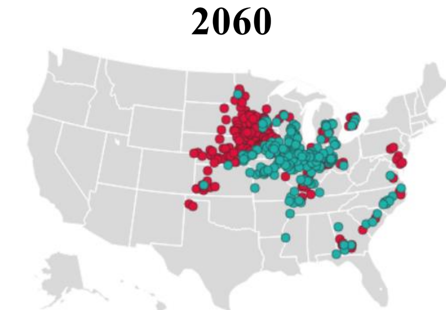
2040



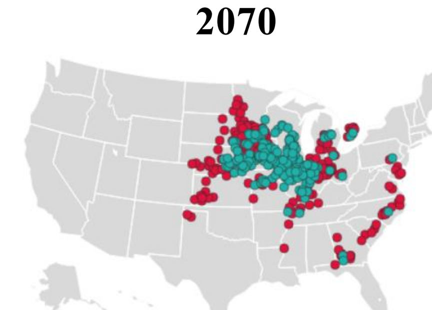
2050



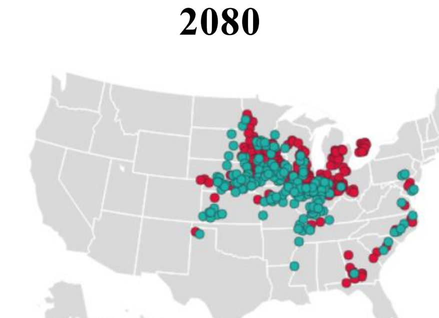
2060



2070

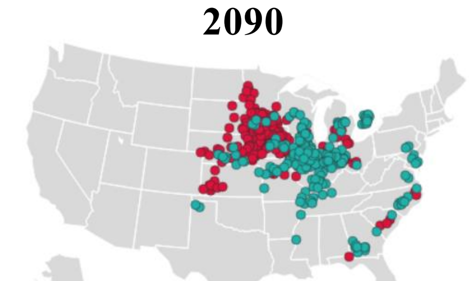


2080



- Performance is expected to decrease in the northwestern states over time for this strain

2090



0.0 - 174.5
174.5 - 184.3

Causal Inference

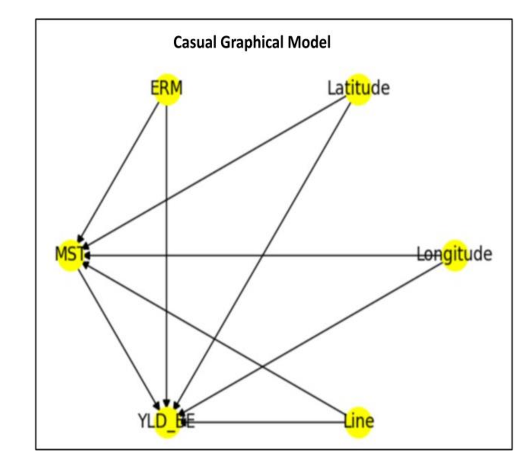
Multivariate Linear Regression Causal Model

Hypothesis	After accounting for the effect of precipitation, low temperature in the summer will reduce yield.
Methodology	• Defined the treatment of low temperature in the summer as having an average temperature for June, July, and August that was lower than the median • Split the dataset into control and treatment group • Implemented a dummy binary categorical variable to assign a treatment value for each observation • Verified normal distribution of target variable (yield) • Predicted yield based off precipitation and average temperature for June, July, and August
Results & Interpretation	After controlling for the effects of precipitation in the summer, a crop in a location with lower average temperature in the summer will yield 0.0741 bushels/acre less on average than a crop in a location with ideal average temperature in the summer.

Causal Inference using Dowhy Library

Machine learning that will model causal assumption and validate through 4 steps:

- **Model:** Dowhy creates causal graphical model from dataset
- **Identify:** Identify desired causal effect criteria based on graphical model
- **Estimate**: Estimate causal effect based on identified criterion
- **Refute** : Refute the obtained estimate through various refutation method



Refute: Bootstrap Sample Dataset:

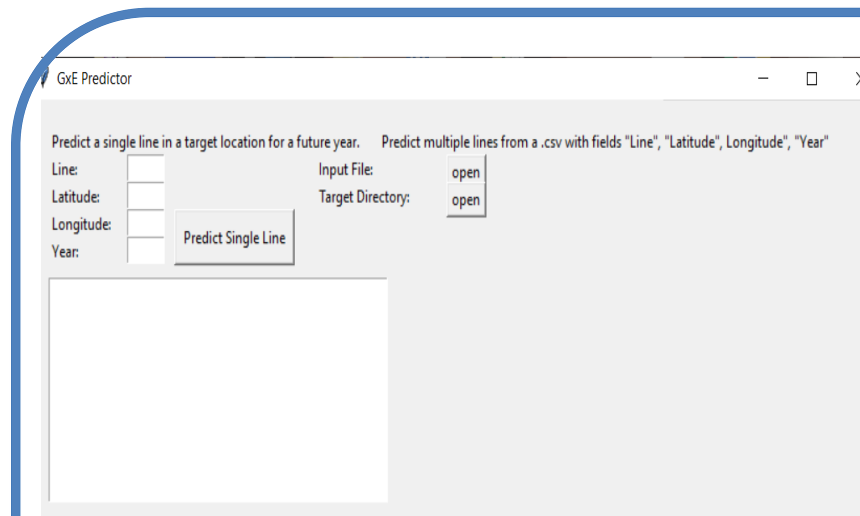
P-value: 0.41

Refute : Use a subset of data

P-value: 0.49

Using Dataset from 2000 with Moisture variable used as treatment

GUI & Future Work



More complete genotype imputation should be explored which would allow the prediction of a specific strain planted in a specified year and location.

Acknowledgments

- We would like to thank the Purdue Data Mine and Bayer Crop Science for the opportunity to explore and learn throughout this year

- Thank you to Alfi Hasan, Richard Sun, and Nima Hamidi for their help & guidance throughout this project