Dashboarding Solution on US Seed Biotechnology and Climate Impact

Denver R. Headings Cantu

D195 Capstone

Western Governors University

Table of Contents

Contents

A. Project Overview

A1. Research Question or Organizational Need

My organization ABC Seeds is seeking to create a climate impact dashboard with public data on genetically modified seed usage in the United States, specifically the impact from severe climate conditions on crop yields to understand the impacts of severe climate conditions on crops to facilitate development of new seed technologies. The company currently only produces genetically engineered seeds for one state in the US and is now wanting to expand to other regions and states. The current region we manufacture seeds for is not frequently affected by severe climate conditions, but the regions we are looking to expand into likely do experience these conditions. The organization needs to understand how much crops are impacted by severe climate conditions, what variables contribute to impact, and how different regions or states might be impacted differently than others. Further, the organization is seeking to understand how the contribution of genetically engineered seed technology has helped mitigate crop yield losses during severe climate conditions, and how that could translate into environmental benefits through reduced water usage and potential water savings.

A2. Scope of Project

The scope of the project is to create a comprehensive Tableau dashboard utilizing various datasets from publicly available sources that are used to create a regression model. The project will focus on agricultural production, genetically engineered seed utilization for the major row crops in the US, and climate data through the years for each state. The scope is limited to the United States and the three major row crops: corn, soy, and cotton. Included in this project is a report of the cleaning, validation and transformation process used to create the datasets in the dashboard. A summary report will also be included to provide an in-depth overview of the project, its results, and outcomes.

The goals of this project are to create visuals that depict the relationships between the variables, provide statistical information and interactive visuals of the data and a predictive model that can be used for next step initiatives. Within these goals, we will deliver a random forest regression model, and a comprehensive dashboard solution in Tableau. Requirements include filter capabilities for commodity and state name.

A3. Summary of Data Analytics Solution

ABC Seeds has asked for a pilot analysis for the climate impact research team initiative. The initial analysis will pave the way for future solutions in climate impact resistant seed technologies, provide information for teams within the organization, and further the goals of expansion into other regions in the United States. The dashboard solution will be created off the cleaned and analyzed datasets we will work though in python. With our Tableau dashboard solution, we will be able to create interactive visuals that users will be able to dive into and receive insightful information about the data. The goal is to create a user-friendly dashboarding solution that provides a clear look into the question at hand, while providing increasing levels of granularity into the data. The dashboard will focus on visuals that depict the relationship between genetically engineered seed usage and crop performance, climate impact and the contribution genetically engineered seed technology has made. The dashboard will also be useful in creating the buy in needed for expanding research into climate tolerant seeds, and the model can be used for identifying the necessary conditions for testing and development.

The main dataset will be accessed through the NASS (National Agricultural Statistics Service) API through python where it will be loaded through a function, cleaned, and transformed into a csv file that will be used for the dashboards. The Palmer Drought Severity Index and climate data will be downloaded via text files and transformed and then combined with the data from NASS. This solution will provide clarity into how the data was captured and processed. Leadership has requested a report of the wrangling and cleaning process to be provided for transparency and verification.

B. Project Plan

B1. Goals, Objectives, and Deliverables

The goal of the project is to conduct an analysis of genetically engineered seeds from three major US row crops (Corn, Soybeans and Cotton) and the climate impact associated with reduction in crop yields.

Goal	Objective	Deliverables
Data cleaning and transformation	Load the data, analyze the cleanliness, clean and transform into a csv format	Jupyter Notebook, CSV files
Tableau Visuals	Create visuals in Tableau sheets	Scatter Plots, Line Charts, Cluster Analysis, Interactive Maps, Bar Charts, informational tables
Tableau Dashboard	Combine the visuals into meaningful dashboards	Tableau workbook dashboards created from individual visuals
Summary Analysis	Create a summary report of the analysis conducted through Jupyter Notebook	PDF document

B3. Standard Methodology

In this project I planned to work in the waterfall method. The original assumption was that I would need to complete each phase before moving on to the next phase. Due to circumstances experienced when working with the data in Tableau, I ended up with more of an agile approach. Discoveries in the data during the visualization process lead me back to the cleaning and wrangling process. Also, during the dashboard visualization process, I found I needed to go back to the original visualizations created to edit and reconfigure to best fit in the dashboard.

For project planning I used the Smartsheet software to outline my phases and tasks. The Gantt chart in fig [1] illustrates the methodology I used to complete the project.

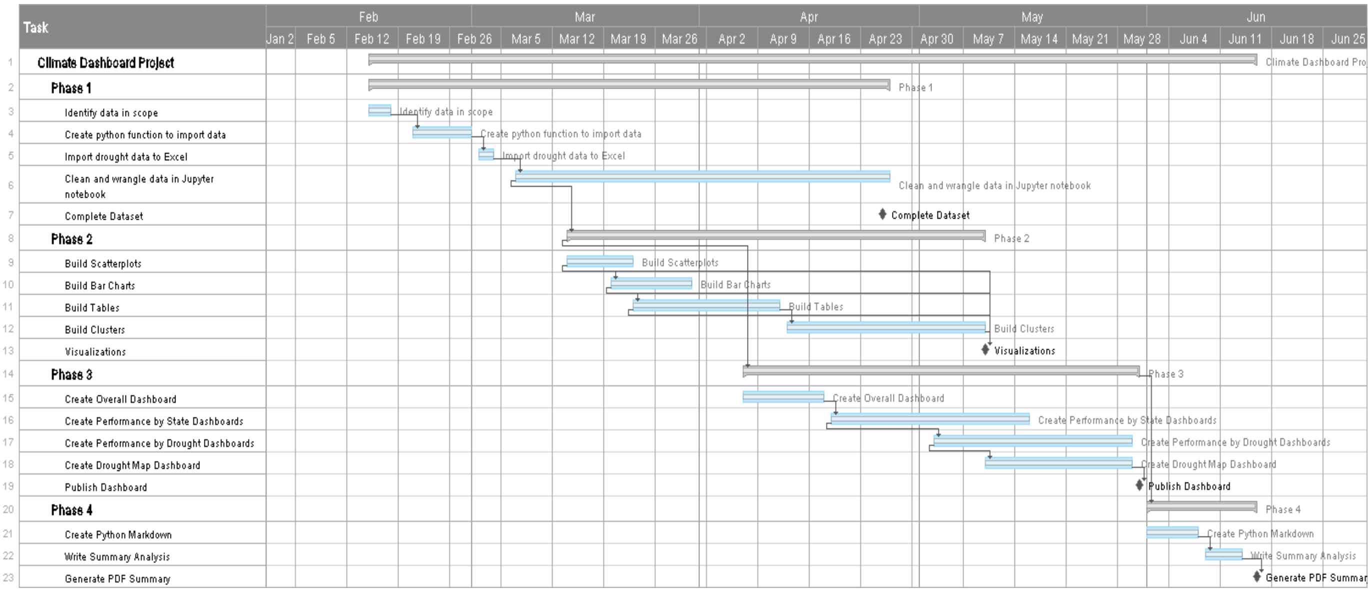


Figure 1: Gantt chart showing project plan

B3. Timeline and Milestones

The below table outlines the projects planned milestones and completion dates along with the executed actual milestone delivery states. The project milestones experienced changed from the original plan due to the methodology changing and unforeseen blockers in working with Tableau.

Table 1: Timelines and Milestones

Milestone	Projected Start Date	Projected End Date	Duration	Actual Start Date	Actual End Date	Actual Duration
Data Cleansing	2/15/2023	2/28/2023	13 days	2/15/2023	4/26/2023	51d
Data Visualization	3/1/2023	3/20/2023	19 days	3/14/2023	5/09/2023	41d
Dashboard Creation	3/21/2023	4/1/2023	12 Days	4/07/2023	5/29/2023	38d
Analysis Summary Report	4/11/2023	4/17/2023	6 days	6/01/2023	6/15/2023	10d

C. Data Methodology

C1. Data Collection Methods

My data selection methodology did not change from my original plan. I was able to connect to the NASS databank via API, download the csv and text sets from the other sources with no issues.

I encountered some issues with syntax when sending the NASS base URL with the required parameters. The data item parameter used specific spacing, commas and where the placement was not easily identifiable. To overcome this obstacle, I tested different parameter syntax and was able to configure the correct syntax for the data item parameter.

There were challenges in formatting the climate data from the NclimDiv archives as the data was in a very raw state and the format of the geocodes were different from other standards. The documentation was available in a readme.txt, but had to be manually extracted and re-created in excel. To overcome this obstacle I created a master table from the documentation table and used Microsoft excel to join the raw dataset and create the state names, extract the dates, and create the value column header. I was able to easily save the climate data for the PDSI(Palmer Drought Index), average temperature and average rainfall for all states and years needed for the analysis.

For this project I did not encounter any unplanned data governance issues.

The advantages of the data set I used are that the data quality was very reliable and no observable outliers existed in the data. A few limitations exist in some states not having complete data, or missing a few years of data.

D. Data Extraction

D1. Data Extraction Technique

The main data extraction technique I used was a python API wrapper for the NASS databank API using the nasspython library. I requested a developer key from the NASS quickstats website and used this key as a parameter in the function to authenticate the request. For the function, it was designed to take an input of the parameter variables, piece together the request URL with the authentication key then return the results in json format. The final piece of the function transforms the json data into a pandas data frame. The function was then ran for each datapoint needed, ACRES PLANTED and YIELD BU/ACRE(LB ACRE). This extraction method was chosen as it utilized the functions within python to read json and parse to a data frame, allowing the data to be viewed in a tabular format.

To prepare the data I used the pandas and missingo libraries. Pandas was used to perform transformations on the shape of the data sets as well as adding formatting and semantic naming conventions. Missingo was used to visualize any missing data that might be present.

To prepare this data, several cleaning operations were needed. The dataset contained many columns that were not necessary, and the data item needed to be pivoted to have its own column with the value for that item in the row. To accomplish this, I pivoted the data frames using pandas and kept only the necessary fields while pivoting the data item description and value to their own column. The data frame then needed to be flattened and the columns renamed to create the appropriate semantics.

Each crop required two sets of data that were extracted and then pivoted in the same method. These two sets of data were then joined together creating a data frame, had the ACRES PLANTED and YIELD BU/ACRE (LB/ACRE) with their own column and respective values.

Next, the data would be checked for missing items and outliers. From here I could see which states did not have a complete row of data for any said year using the missingo library. These states were identified as not being necessary to the results and were dropped.

The same process was repeated for each crop and then at the end, the data frames were unioned into one dataset containing all of the states, years, acres planted and yield for each crop. The dataset was then checked for cleanliness and consistency by looking at the unique entries for state and commodity (crop type). The data contained an entry for “Other States” which were dropped.

Next, the biotech data was imported from a csv file to a dataframe and merged with the completed NASS dataset. This merge would create the columns for percentage of genetically engineered seeds planted for each state, by year and commodity.

I performed the same step for the climate data csv and merged it with the completed dataframe from the step before.

The data was then checked for missing data, and at this step missing data was dropped to produce a completed data frame ready to export to a csv.

Python in Jupyter was the appropriate tool for data extraction and preparation in this project as it works in chunks that allows me to view the results with each step and easily troubleshoot along the way. The nasspython, pandas and missingo libraries were the only needed additional tools to complete this process, they provided the necessary functions for the extraction and preparation process.

E. Data Analysis Process

E1. Analysis Methods

Originally for this project the plan was to analyze the data through multi-linear regression with interpretations of estimated coefficients and k-means clustering. I chose to complete the multi-linear regression models in python and then create k-means clusters in Tableau for added visual and statistical story telling. Though, during the modeling process I encountered issues within my data that lead me to test other models. I determined that the model that best fit this data and environment would be a random forest model. I encountered issues with multi-collinearity and overall fit, and with the limited data was unable to satisfy the classic assumptions of an ordinary least squares regression model. I attempted many methods to improve the model but decided to test a random forest model which produced much better results. I compared the results between the two models and it was decided that the random forest model would be the direction for the project.

Further, when testing the data in Tableau, limitations were realized in the ability to display the statistical results of k-means clusters in Tableau. Visually, they have some appeal, but the text limitations would lead to the creation of static visuals with no dynamic filters, which would violate the requirements to include filters. It was decided to drop the k-means clusters in Tableau, as they did not add much value to the project or visual capabilities.

E2. Advantages and Limitations of Tools

Python is a very strong analytic tool and provides the necessary capabilities to create multi-linear regression models. Using python in Jupyter Notebook provides its own advantages with being able to complete the code in chunks, allowing me to test each module of code and change or alter where needed. Jupyter Notebook also provides useful markdown chunks, that facilitate the creation of documentation and storytelling of the process and approaches used through the data cleaning, wrangling and analysis process. The main tool for visualization used was Tableau. The advantage of using Tableau is that it allows a drag and drop feature of variables to create a variety of visualizations with the ability to add trendlines and provides statistical calculations. Further, filters can be added to group specific visualizations and view the outcomes with high granularity and doesn't require extensive code to accomplish this. Calculated fields allow you to create additional metrics for analysis, providing additional depth and discovery into your models. Tableau has limitations with the creation of advanced regression models, and, it does not have many functions that allow dynamic results from models to be displayed by viewers. For instance, regression trendlines added to plots provide r-squared and p values, but the viewer can only see these values if they hover over the line. Copy/pasting these values to a text field would not work either, as the visual is intended to be dynamic with filters, and the text remains static. For the general audience of this project, this does not pose a huge issue, as the main function is to provide a general indication through easy to interpret visuals.

E3. Step by Step Process of Analytical Methods

Step 1: Initial Look Into The Data

After cleaning and wrangling the main dataset that included all the data in one frame, the data was exported to a csv and imported into Tableau for the base analysis. In this step, simple plots were created to visualize linear relationships between the dependent variable, yield per acre, and year. The data was also checked for outliers and anomalies in this step, using Tableau for this function accelerated the process it reduced the amount of code needed in the Jupyter notebook. In this step it was identified that California was an outlier in the cotton commodity (fig.2), it has much higher yields per acre than any other state in that group. In this plot, I was able to quickly

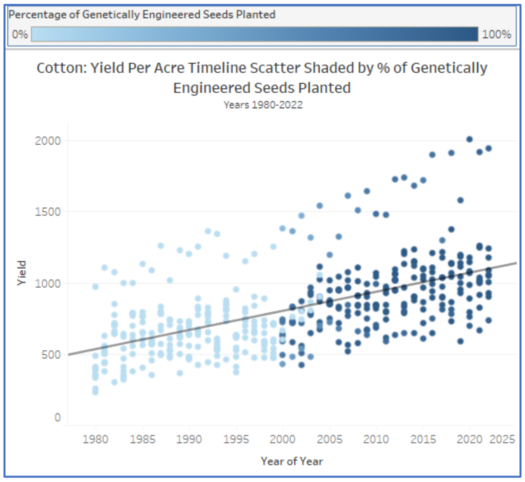


Figure 2: Scatter plot of cotton yields through the years 1986-2020

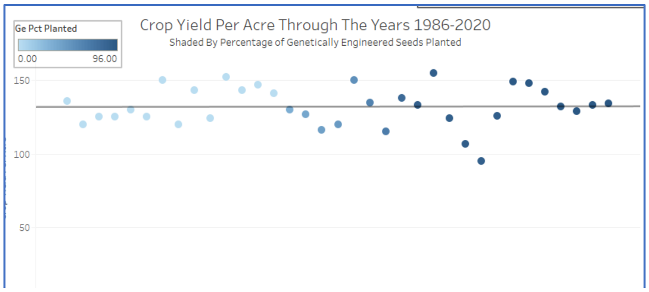


Figure 3: Scatter plot of corn yields through years 1986-2020 for the state of Kansas

isolate California as the source for the data points that were much higher than all of the others. According to an article from UCDAVIS (2016) *"The soils and climate in California are favorable for cotton production resulting in yields that were at least twice those of the U.S. average between the mid-1920s and the 1970s. Even today, California's yields are close to twice as high as the U.S. average."* At this step it was decided to remove California from the cotton analysis.

Also identified as an outlier is the state of Kansas (fig.3) in the corn commodity. This state had no linear relationship between the yield per acre and year with a completely straight trendline. Because Kansas shows no linear relationship, it would be removed from the analysis as an outlier.

Step 2: Investigate Climate Features

With initial analysis on the dependent variable yield per acre and percentage of genetically engineered seeds completed, I moved on the looking at yield per acre and the PDSI(Palmer Drought Severity Index) fig(4,5). With this simple plot, I was able to identify that the relationship between the two variables was not exactly linear. A curve appeared in the plot, which indicated that there is a quadratic relationship, and plotting a polynomial trendline with 2 degrees quickly proved this to be the case.

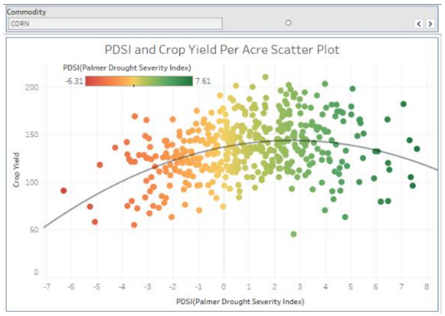


Figure 2: PDSI and Crop Yield Scatter Plot for Corn

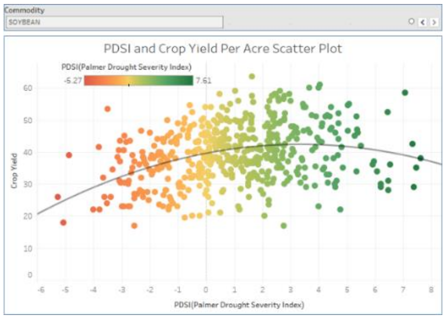


Figure 3: PDSI and Crop Yield Scatter Plot for Soybeans

Though, I found that the significance of the relationship between the cotton commodity and PDSI was not as significant at p value = 0.04 (fig. 6). Upon further research I found that as a glycophyte, cotton is more tolerant to abiotic stress than other staple crop species. With this information and the observations from the plot, it was decided that cotton would be dropped for this analysis. Also, even with the removal of California from the data, the yield distribution was negatively skewed. These additional circumstances with the cotton commodity data do not make it a good fit for the multi linear regression models and would require its own separate analysis.

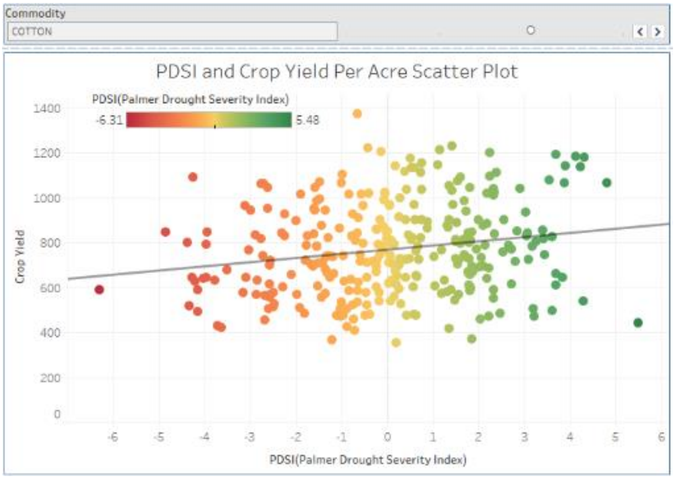


Figure 4: PDSI(Palmer Drought Severity Index) and Yield Per Acre Scatter Plot for Cotton

Step 2a: Plotting the additional climate independent variables

Next, I plotted the average temperature and average rainfall in inches against the crop yield. Both variables also indicated a parabolic shape in the plots. Within this process I did find evidence of an outlier event, which would be a natural disaster that occurred in the year 1993 (fig. 7). "The Great Flood of 1993 (or Great Mississippi and Missouri Rivers Flood of 1993) was a flood that occurred in the Midwestern United States, along the Mississippi and Missouri rivers and their tributaries, from April to October 1993." (USGS,2023). This disaster is known as one of the worst flooding events to occur in the United States and resulted in billions of dollars in losses. Crop devastation was major, and heavily impacted the corn commodity and some of the soybean commodity. Also, another event was noted for 2012, where a large drought event occurred in the United States. According to The National Weather Service "The 2012–2013 North American drought, an expansion of the 2010–2013 Southern United States drought, originated in the midst of a record-breaking heat wave. Low snowfall amounts in winter, coupled with the intense summer heat from La Niña, caused drought-like conditions to migrate northward from the southern United States, wreaking havoc on crops and water supply". While these events can introduce outliers into our model I decided to keep them as extreme weather events are an important part of climate analysis on crops.

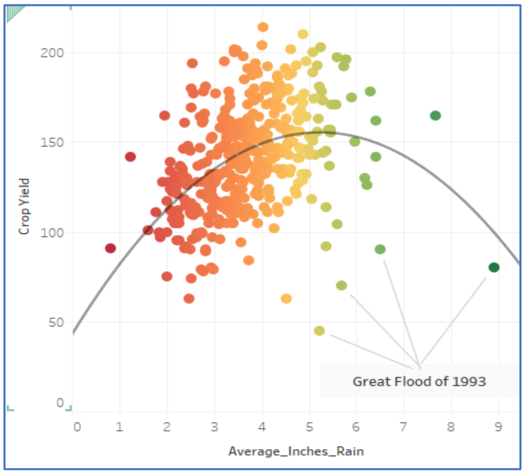


Figure 5: Average Inches Rain Scatter Plot highlighting a significant climate event

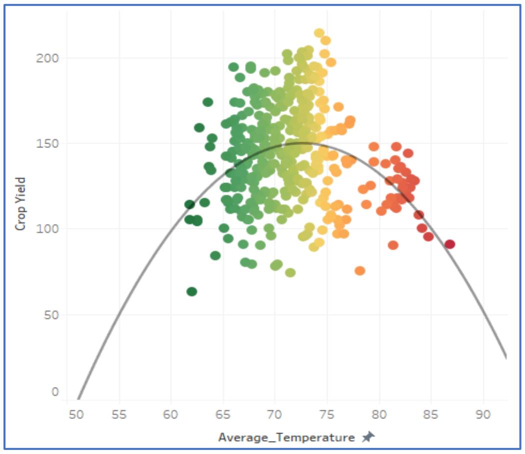


Figure 6: Average Temperature Scatter Plot

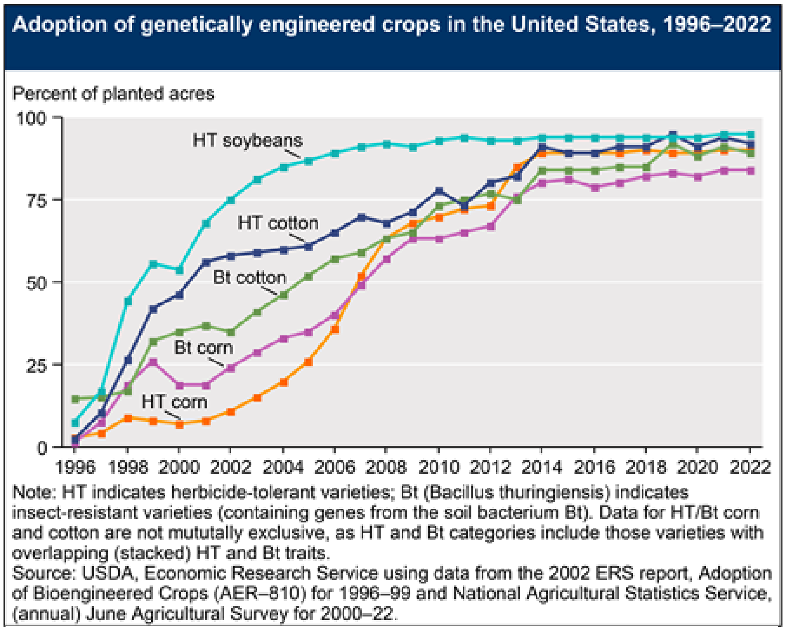
Looking at the plot (fig. 8) for crop yield and average temperature, the same shape is apparent, with lower yields in lower temperatures and lower yields in higher temperatures. Due to the parabolic shape, the model will be tested with centered and squared terms of the climate features. While random forest is a robust model, testing these terms will be necessary to determine the best fit for the model.

Step 3: Model Preparation

With the information from the initial look into the data, the next step was to look at the approach in testing a random forest model in python and check for some basic assumptions and fit of the model. Being that random forest is a robust model and does not require that all data be normalized, or be linear, there is more freedom within the assumptions. Though, I did investigate the distributions and linear relationships between the features.

The first step was to look at the distributions of the data, which the dependent variable yield per acre, and independent variables GE %, PDSI, average temperature, average inches rainfall. All of the climate features were relatively normally distributed.

The % of GE planted variable though was not distributed evenly. Being that half of the years in the analysis are considered “pre GE” (1986-1999), and the other half “post GE” (2000-2020), all of the pre GE years had zeros for the % GE planted values. There is a set of missing years 1996-1999 where genetically engineered seeds were introduced, but specific data is not available. Data does not become available until 2000. To account for this, I decided to perform a simple interpolation on the data to create values for years 1996-1999. An infographic from USDA (fig.9) is the only available information on what implementation looked like for those years, and it is clear there was a strong increase during that time. This provided more validity to the interpolation route, and I de-incremented the first value available in the year 2000, all the way to 0 in year 1995. This provided values for the years 1996-1999 which would provide a better dataset for the model. Though, the question remained on how many zeros should I include in the model, and I decided that it would not be wise to include 20 years of zeros, and instead opted for 5 years of zeros and the model was created on years 1990-2022. This would give the model a more even distribution for the genetically engineered percentage feature.


Adoption of genetically engineered crops in the United States, 1996–2022

Note: HT indicates herbicide-tolerant varieties; Bt (Bacillus thuringiensis) indicates insect-resistant varieties (containing genes from the soil bacterium Bt). Data for HT/Bt corn and cotton are not mututally exclusive, as HT and Bt categories include those varieties with overlapping (stacked) HT and Bt traits.
Source: USDA, Economic Research Service using data from the 2002 ERS report, Adoption of Bioengineered Crops (AER–810) for 1996–99 and National Agricultural Statistics Service, (annual) June Agricultural Survey for 2000–22.

Figure 7 (USDA): Line chart that shows the adoption of genetically engineered crops in the United states

Step 4: Assessing Model Features

The next step was to create additional features to test in the model. Knowing that the climate data follows a parabolic shape, I created centered and squared terms of those features to test in the model. Also, a category was created to bin the percentage of genetically engineered seeds. The state feature was one-hot encoded to be included in the model.

Next, I would test these features in the model and assess the fit. It was determined that the squared and centered terms were not necessary for the model, and the features fit best in their original form. When running the model with all the squared term features, then running without them and seeing improvement, this suggests that multicollinearity might have been affecting the model's ability to discern the importance of individual features. The climate features appear to have high correlation to the dependent variable, and among themselves. I decided to keep only the PDSI, and average temperature, state, and percentage of genetically engineered seeds planted as features to run against the yield dependent variable. Average rainfall appears to be moderately correlated with the PDSI, so that was excluded.

Step 5: Validating The Model

After running the model, I checked the rest of the assumptions for the model to assess the fit. I first referenced the MSE and R-Squared values from the model to determine how well the model performed. The Mean Squared Error (MSE) is a measure of how well the model's predictions match the true values. It calculates the average squared differences between the predicted and actual values. A smaller MSE indicates a model that predicts more accurately, whereas a larger MSE indicates a model that predicts less accurately.

R-Squared, also known as the coefficient of determination, is a statistical measure that shows the proportion of variance in the dependent variable (in this model, crop_yield) that is predictable from the independent variables.

Next, I evaluated the feature importance to give insight into which features were most influential in predicting crop yields. Looking at feature importance would also provide domain-specific insights that could guide further analysis or intervention. Understanding which variables most influence crop yield is likely just as important as the predictive capability of the model.

Last step was to test the residuals and assess for the null hypothesis for homoscedasticity. To accomplish, I first plotted the residuals against the fitted values to observe the plots. Then, I ran a Breusch-pagan test to statistically assess the presence of heteroscedasticity.

F. Results

F1. Evaluation of Statistical Significance

F1.a Corn Model

MSE and R-Squared

The first step in interpreting the results from the random forest model was to access the MSE and R-Squared values. For the corn commodity model the values were:

```
MSE:  180.58852750000005
R-Squared:  0.7997604304564911
```

The MSE (Mean Squared Error) indicates that on average the model's predictions are about 13.46 units away from the true crop yield values, on average. Given the average crop yield is about 140 bushels per acre, this error represents about 9.6% of the average yield. With the variability in yields, this is acceptable for our use case.

For the R-Squared value, expressed as a percentage, the model now explains approximately 79.93% of the variability in crop yield, which suggests a good fit to the data.

OOB (Out-Of-Box) Score

I ran an OOB test to assess the validity of the model when run against the untrained samples using the RandomForestRegressor. It is calculated using the samples that are not used in the training of the model, which are referred to as out-of-bag samples. These samples are used to provide an unbiased estimate of the model's performance and provides us with an additional validation metric.

```
OOB Score: 0.6483027551904614
```

The OOB score, in the context of a regression model like RandomForestRegressor, is the R-squared (coefficient of determination) calculated using out-of-bag samples. The R-squared value is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with 1 indicating perfect prediction and 0 indicating that the model does not account for any of the variability in the target variable.

The OOB R-squared score of 0.6483 suggests that approximately 64.83% of the variation in the crop yield (the dependent variable) can be explained by the model's predictors when tested on out-of-bag samples. This is a relatively good score, indicating that the model captures a substantial portion of the variability in the crop yields. However, around 33.14% of the variability is still unaccounted for, which gives us a measure to look for additional variables that contribute to crop yield.

One agricultural topic that would likely have a significant impact is disease and pest infestations. Currently, there is no comprehensive data set that captures this information and could be another subject for a project within the company. While climate conditions typically exacerbate disease and pest devastation, disease and pest infestations can and often occur in normal climate conditions. This is likely an important feature that would improve the predictive ability of the model.

For the scope of our project, this provides enough validation for the model to be used for research and generalized assumptions.

Feature Importance

Feature importance was analyzed as a part of the process to validate the models results and fit. Knowing which variables were the most important contributors to the model would provide crucial information for expansion and improvement of the model. Also, this would provide investigative topics to expand upon. In a domain like agriculture, this could lead to actionable insights.

- *Feature: Ge_Pct_Planted, Importance: 0.39974737522499076*
- *Feature: PDSI, Importance: 0.1613940270054687*
- *Feature: Average_Temperature, Importance: 0.18706558572554915*
- *Feature: state_name_ILLINOIS, Importance: 0.01285590373479315*
- *Feature: state_name_INDIANA, Importance: 0.004773523495417663*
- *Feature: state_name_IOWA, Importance: 0.01450798451936228*
- *Feature: state_name_MICHIGAN, Importance: 0.0095819360714204*
- *Feature: state_name_MINNESOTA, Importance: 0.003459306697172654‍8*
- *Feature: state_name_MISSOURI, Importance: 0.009528317462351155*
- *Feature: state_name_NEBRASKA, Importance: 0.00897179734515709*
- *Feature: state_name_NORTH DAKOTA, Importance: 0.09328171529947404*
- *Feature: state_name_OHIO, Importance: 0.005019249621446916*
- *Feature: state_name_SOUTH DAKOTA, Importance: 0.08103473412033452*
- *Feature: state_name_TEXAS, Importance: 0.003559787984973342*
- *Feature: state_name_WISCONSIN, Importance: 0.005218755692088177*

1. Top Features by Importance
Ge_Pct_Planted (37.97%): This is by far the most significant feature. It suggests that the proportion of genetically engineered seeds planted has the most considerable influence on crop yields. This could imply that genetically engineered seeds have a distinct performance profile when it comes to yield outcomes, maybe due to resistance to certain pests, diseases, or better adaptability to various soil types and weather conditions.

Average_Temperature (18.70%): Temperature plays a pivotal role in crop yield, impacting growth phases, moisture availability, pest activity, etc. Its high importance emphasizes the role of temperature in determining yields.

PDSI (16.13%): The Palmer Drought Severity Index measures prolonged and abnormally dry or wet periods. It's essential because drought or overly wet conditions can severely impact crop growth and development.

2. State Specifics
NORTH DAKOTA (9.32%) and SOUTH DAKOTA (8.10%): These states have considerably higher importance than other states, suggesting they may have distinct patterns or conditions that are especially relevant in determining crop yields. Perhaps there are unique weather conditions, soil types, or farming practices in these states that set them apart from others. For these two states, they were most heavily impacted by the Great Flood of 1993 with off the charts low yields for that year. Though, the climate data does not capture the severity of the flood, since none of those features capture flood

conditions, rather, just moisture and rainfall. Rainfall during that time was rather high, but the devastation from flowing water is a different category of data that would be outside of the scope of the data. It is important to note that the model was able to capture *reduced* yields during these events, just not the *severity*.

Most other states have feature importance values less than 1%, with a few exceptions like ILLINOIS (1.28%) and IOWA (1.45%). This indicates that, after accounting for other factors like temperature, PDSI, and genetically engineered seed usage, the specific state becomes less crucial for the overall model. However, it doesn't mean the state isn't important; instead, it means the unique variance attributed solely to the state, after considering other features, is relatively low.

Residuals Analysis

To check the model for the presence of heteroscedasticity, first a plot (fig. [10]) was created against the residuals and predicted values. While the initial view shows a large degree of randomness we can see that it is not completely random.

Next, a Breusch Pagan hypothesis test was run using the statsmodels package from python. The presence of heteroscedasticity would indicate variance of the errors is not consistent across all levels of the independent variables. This method will access the presence of heteroscedasticity with the following assumptions:

Null Hypothesis (H0): Homoskedasticity is present (variance of the errors is constant).

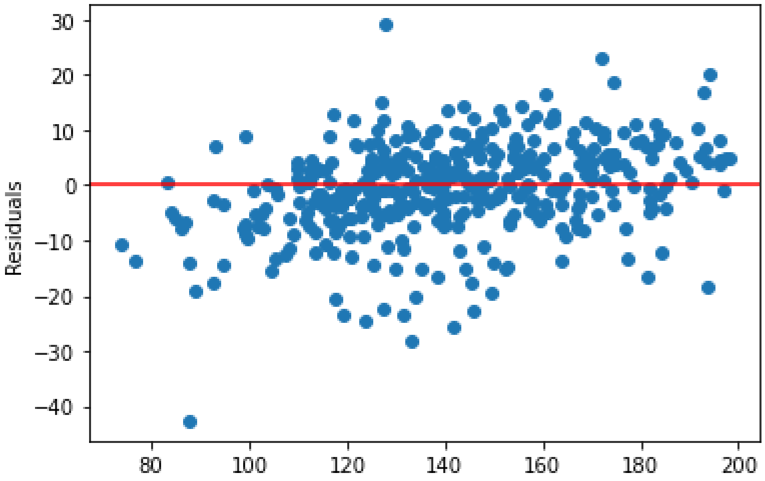


Figure 8

Alternative Hypothesis (H1): Heteroskedasticity is present.

The significance value is to measure the results will be (p > 0.05).

The results from the Breusch Pagan hypothesis test came out as follows:

```
'LM Statistic': 19.48567629880924, 'LM-Test p-value': 0.14720849483770648,
'F-Statistic': 1.4084159054536463, 'F-Test p-value': 0.14584314411989935
```

These results indicate that we are unable to reject the null hypothesis (H0) that homoscedasticity is present as our LM-Test p-value and F-Test p-values are greater than the indicated significance (p > 0.05). This provides evidence that there is not a large degree of heteroscedasticity in the model and the variance of the errors is constant. This gives further weight to the reliability of the model and its predictions.

Predictions vs Actuals

The actuals and predictions were plotted in Tableau (fig. 11) to include in the dashboards as evidence of the relationships and accuracy of the model. The overall plot indicated a linear consistency in the actuals vs fitted (predicted) values. Near the bottom we can observe a specific outlier and some scatter. These values represent the crop disaster likely from the Great Flood of 1993. While the model was mostly accurate in its predictions, it was unable to account for the severity of impact from an infrequent natural disaster. Given the spread and cohesiveness of the data points, I am satisfied with the results from the model.

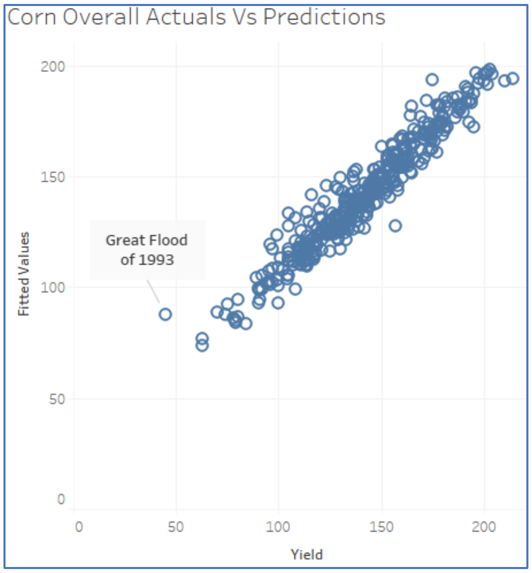


Figure 9: Corn Actuals vs Predicted Values Scatter Plot

In the next chart (fig. 12) we can observe the actual crop yields plotted compared to the fitted values for each GE seed category and year. The data points are shaded by drought severity – orange/dark red capturing severe drought and dark blues capturing extreme moisture. We can observe the predicted values decreasing where there are severe climate conditions indicated by the PDSI. This aligns with the actuals yields being lower and corresponding with the PDSI severity. The model was able to predict lower yields per acres in periods of extreme or severe dryness that correspond to the actuals. The main time period of interest in 2012, where widespread drought affected many of the states in this analysis. The Great Flood of 1993 is also observable in the visualization, where the model was able to capture the reduced yields due to the disaster. This gives us clear evidence of the impact that climate severity has on crop yields.

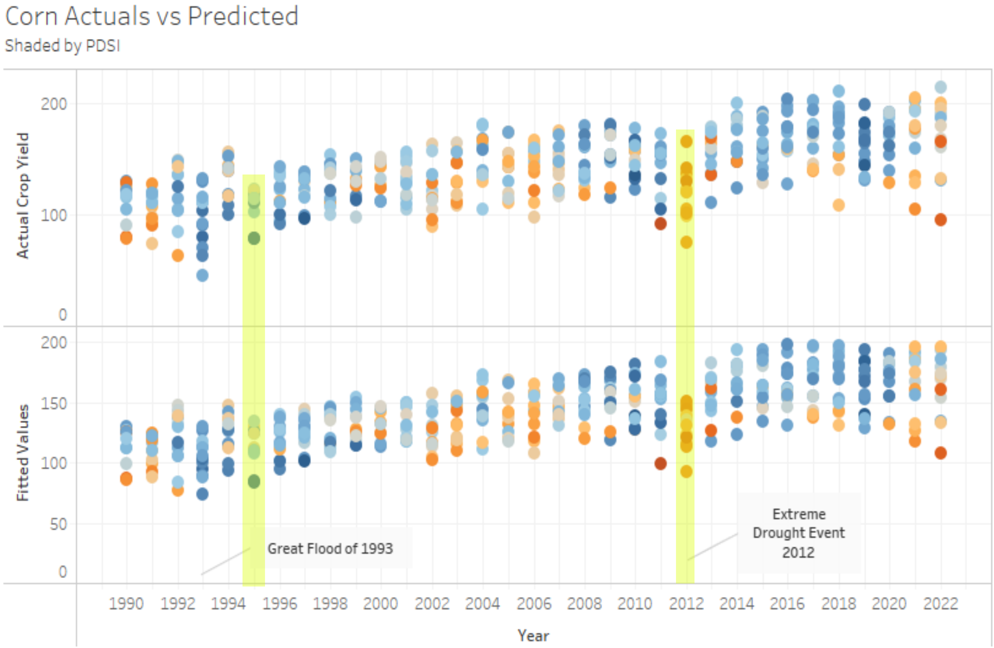
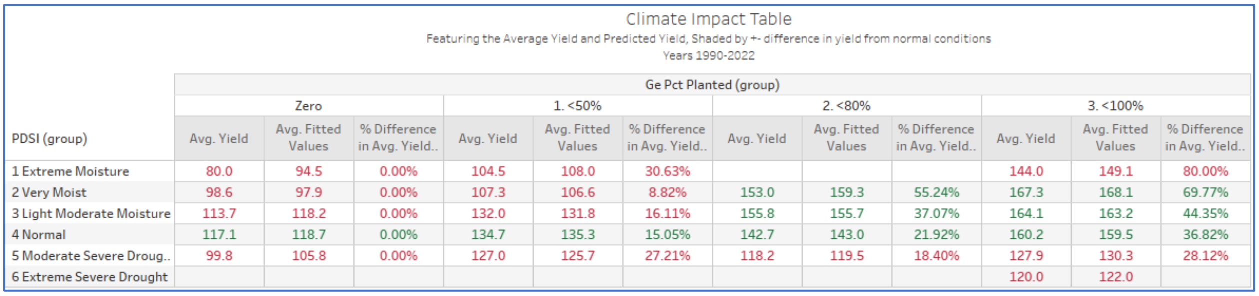


Figure 10: Corn Actuals vs Fitted Values chart shaded by the PDSI(Palmer Drought Severity Index). Highlights on two significant climate events

Climate Impact Assessment

A crucial feature of the dashboard that was created is the Climate Impact Assessment. The data was grouped into bins for the percentage of genetically engineered seeds planted to give an indication on the performance of implementation. This includes a table that captures the different degrees of climate conditions, the actual average yields, the average predicted yields, and the difference in the actual yields from the prior genetically engineered planted percentage category.



Climate Impact Table

Featuring the Average Yield and Predicted Yield, Shaded by +- difference in yield from normal conditions

Years 1990-2022

	Ge Pct Planted (group)											
	Zero			1. <50%			2. <80%			3. <100%		
PDSI (group)	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..
1 Extreme Moisture	80.0	94.5	0.00%	104.5	108.0	30.63%				144.0	149.1	80.00%
2 Very Moist	98.6	97.9	0.00%	107.3	106.6	8.82%	153.0	159.3	55.24%	167.3	168.1	69.77%
3 Light Moderate Moisture	113.7	118.2	0.00%	132.0	131.8	16.11%	155.8	155.7	37.07%	164.1	163.2	44.35%
4 Normal	117.1	118.7	0.00%	134.7	135.3	15.05%	142.7	143.0	21.92%	160.2	159.5	36.82%
5 Moderate Severe Droug..	99.8	105.8	0.00%	127.0	125.7	27.21%	118.2	119.5	18.40%	127.9	130.3	28.12%
6 Extreme Severe Drought										120.0	122.0	

Figure 11: Climate Impact Table features the average yield per acre against the predicted yield per acre for each climate condition (PDSI) group. A comparison is made on the difference in actual yields for each category compared to zero.

We can make clear observations about the data from this table, including the consistent increase in crop yields from 'Zero' to '<100%'. We can observe that in the zero and <50% genetically engineered seeds planted that performance was reduced in all conditions except normal. This indicates that crops were more sensitive to dry and moisture conditions. In the <80% to <100% categories, we can see improvement in the light-moderate moisture category, but reduction in the extreme moisture category. This could be an indication that as seed technology has improved and implementation increased, resistance to moisture conditions improved.

For moderate-severe drought we can observe in the <80% category that performance was actually less than in the <50% category. This is likely due to the severe drought event of 2012, where most of the states affected were implemented in the <80% range. Though, we can observe in the <100% category that while the moderate-severe category performance is improved from the prior category, it isn't much better than the <50% category. This indicates there needs to be more research and investment in developing more robust seeds to withstand drought conditions.

A field capturing the percentage difference in the actual yield in each category compared to the first category of 'Zero', gives us insight into how seeds perform compared to the years analyzed when no genetically engineered seed technology existed. We can see a steady and significant increase in each category among all climate conditions. Where the <100% category does not have significant performance compared to the <50% or <80% category in moderate-severe drought climate conditions, it does have a 28% increase overall when compared to the 'Zero' category. This indicates that farmers are able to realize a nearly 30% increase in yields during adverse climate conditions compared to when seed technology did not exist.

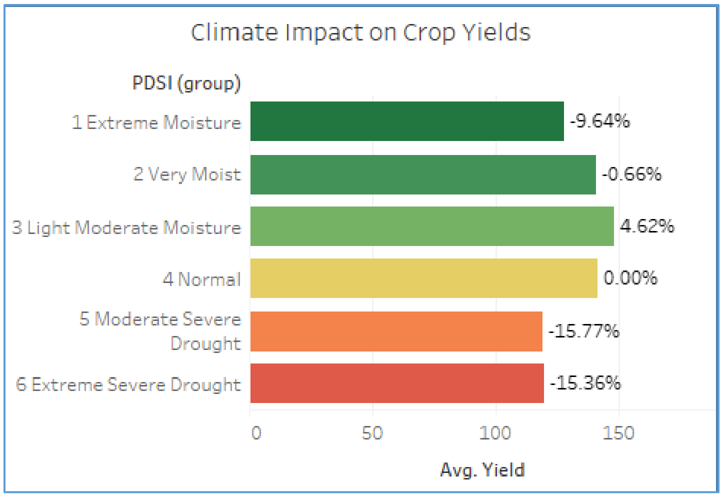


Figure 12: Bar chart of Climate Impact on Crop Yields

A chart (fig. 14) was created to assess the impact of potential loss from severe climate conditions. Overall we can observe that it can be expected about a 15% decrease if conditions are in the moderate-severe to extreme drought conditions. For the extreme moisture conditions, we could expect an average of about 10% loss in crop yields. This information can provide us crucial insight into the anticipated loss farmers can expect as severe climate conditions impact. Also, this provides a measurement for us to research the development of new seed technologies to become more resistant to these conditions.

The dashboard provides us the ability to look into each individual states results. This provides very useful information into the individual impacts and how they can dramatically differ from the overall assessment. For example (fig. 15), with the state of Illinois, we can observe a much more severe impact within the moderate-severe drought climate conditions. An average of 34% loss when those conditions exist. We will be able to investigate and look for other features that might be a contributing factor, like soil composition, or unique disease and pests that affect only that region and are exacerbated by the high stress climate condition. Further, we can test and develop new seed technologies in created environments that mimic specific conditions like what might exist in Illinois. We can then compare our results in this model to determine outcomes and effectiveness of the developments and technology.

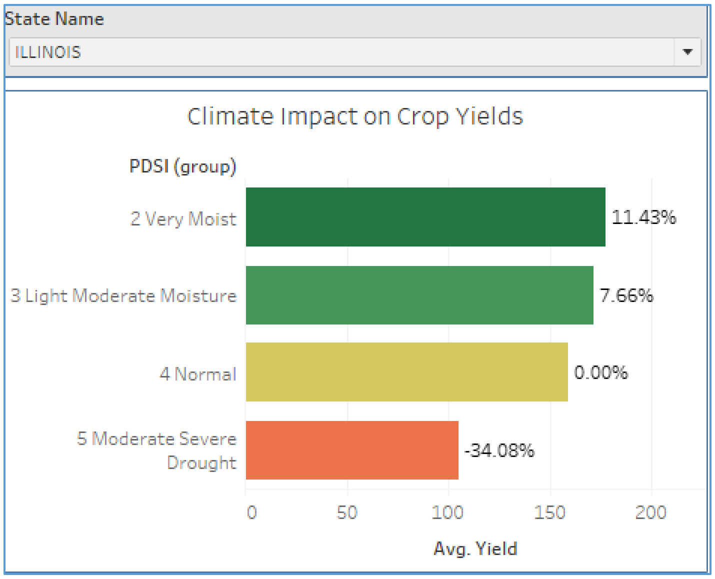


Figure 13: Bar chart of Climate Impact on Yields for the state of Illinois

F1.b Soybean Model

MSE and R-Squared

The first step in interpreting the results from the random forest model was to access the MSE and R-Squared values. For the corn commodity model the values were:

```
MSE:  19.259744892473115
```

```
R-Squared:  0.7422583387423574
```

The MSE (Mean Squared Error) indicates that on average the model's predictions are about 4.38 units away from the true crop yield values, on average. Given the average crop yield is about 41 bushels per acre, this error represents about 10.6% of the average yield. With the variability in yields, this is acceptable for our use case.

For the R-Squared value, expressed as a percentage, the model now explains approximately 74.22% of the variability in crop yield, though lower than our corn model, it still suggests a good fit to the data.

OOB (Out-Of-Box) Score

I ran an OOB test to assess the validity of the model when run against the untrained samples using the RandomForestRegressor. It is calculated using the samples that are not used in the training of the model, which are referred to as out-of-bag samples. These samples are used to provide an unbiased estimate of the model’s performance and provides us with an additional validation metric.

```
OOB Score: 0.6686160609340628
```

The OOB score, in the context of a regression model like RandomForestRegressor, is the R-squared (coefficient of determination) calculated using out-of-bag samples. The R-squared value is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with 1 indicating perfect prediction and 0 indicating that the model does not account for any of the variability in the target variable.

The OOB R-squared score of 0.6686 suggests that approximately 66.86% of the variation in the crop yield (the dependent variable) can be explained by the model's predictors when tested on out-of-bag samples. This is a relatively good score, and higher than our corn model, indicating that the model captures a substantial portion of the variability in the crop yields. However, around 33.14% of the variability is still unaccounted for, which gives us a measure to look for additional variables that contribute to crop yield.

One agricultural topic that would likely have a significant impact is disease and pest infestations. Currently, there is no comprehensive data set that captures this information and could be another subject for a project within the company. While climate conditions typically exacerbate disease and pest devastation, disease and pest infestations can and often occur in normal climate conditions. This is likely an important feature that would improve the predictive ability of the model.

For the scope of our project, this provides enough validation for the model to be used for research and generalized assumptions.

Feature Importance

Feature importance was analyzed as a part of the process to validate the models results and fit. Knowing which variables were the most important contributors to the model would provide crucial information for expansion and improvement of the model. Also, this would provide investigative topics to expand upon. In a domain like agriculture, this could lead to actionable insights.

```
•  Feature: Ge_Pct_Planted, Importance: 0.317698967641399
•  Feature: PDSI, Importance: 0.1697004806545178
•  Feature: Average_Temperature, Importance: 0.25713017830800516
•  Feature: state_name_ARKANSAS, Importance: 0.003031145436720671
•  Feature: state_name_ILLINOIS, Importance: 0.016237956544813702
•  Feature: state_name_INDIANA, Importance: 0.00567274012071869
•  Feature: state_name_IOWA, Importance: 0.010779776912229713
•  Feature: state_name_KANSAS, Importance: 0.038119183629038406
•  Feature: state_name_MICHIGAN, Importance: 0.0029379781598586987
•  Feature: state_name_MINNESOTA, Importance: 0.00375839824109468
•  Feature: state_name_MISSISSIPPI, Importance: 0.0022661816993791314
•  Feature: state_name_MISSOURI, Importance: 0.011441333306111997
•  Feature: state_name_NEBRASKA, Importance: 0.009646445831574948
•  Feature: state_name_NORTH DAKOTA, Importance: 0.10756952980294593
•  Feature: state_name_OHIO, Importance: 0.0027517837494373475
•  Feature: state_name_SOUTH DAKOTA, Importance: 0.03861193707499506
•  Feature: state_name_WISCONSIN, Importance: 0.00264598288715902
```

1. Top Features by Importance
Ge_Pct_Planted (31.76%): This is by far the most significant feature. It suggests that the proportion of genetically engineered seeds planted has the most considerable influence on crop yields. This could imply that genetically engineered seeds have a distinct performance profile when it comes to yield outcomes, maybe due to resistance to certain pests, diseases, or better adaptability to various soil types and weather conditions.

Average_Temperature (25.71%): Temperature plays a pivotal role in crop yield, impacting growth phases, moisture availability, pest activity, etc. Its high importance emphasizes the role of temperature in determining yields. To note, this is a more important feature for the soybean crop than it is for the corn crop.

PDSI (16.97%): The Palmer Drought Severity Index measures prolonged and abnormally dry or wet periods. It's essential because drought or overly wet conditions can severely impact crop growth and development.

2. State Specifics
NORTH DAKOTA (10.75%): This state has considerably higher importance than other states, suggesting it may have distinct patterns or conditions that are especially relevant in determining crop yields. Perhaps

there are unique weather conditions, soil types, or farming practices in these states that set them apart from others. For North Dakota, it was heavily impacted by the Great Flood of 1993 with off the charts low yields for that year. Though, the climate data does not capture the severity of the flood, since none of those features capture flood conditions, rather, just moisture and rainfall. Rainfall during that time was rather high, but the devastation from flowing water is a different category of data that would be outside of the scope of the data. It is important to note that the model was able to capture *reduced* yields during these events, just not the *severity*.

Most other states have feature importance values less than 1%, with a few exceptions like SOUTH DAKOTA (3.86%) and KANSAS (3.81%). This indicates that, after accounting for other factors like temperature, PDSI, and genetically engineered seed usage, the specific state becomes less crucial for the overall model. However, it doesn't mean the state isn't important; instead, it means the unique variance attributed solely to the state, after considering other features, is relatively low. For these two states, more investigation is likely needed into the particulars that affect their crop yield.

Residuals Analysis

To check the model for the presence of heteroscedasticity, first a plot (fig.16) was created against the residuals and predicted values. While the initial view shows a large degree of randomness we can see that it is not completely random.

Next, a Breusch Pagan hypothesis test was run using the statsmodels package from python. The presence of heteroscedasticity would indicate variance of the errors is not consistent across all levels of the independent variables. This method will access the presence of heteroscedasticity with the following assumptions:

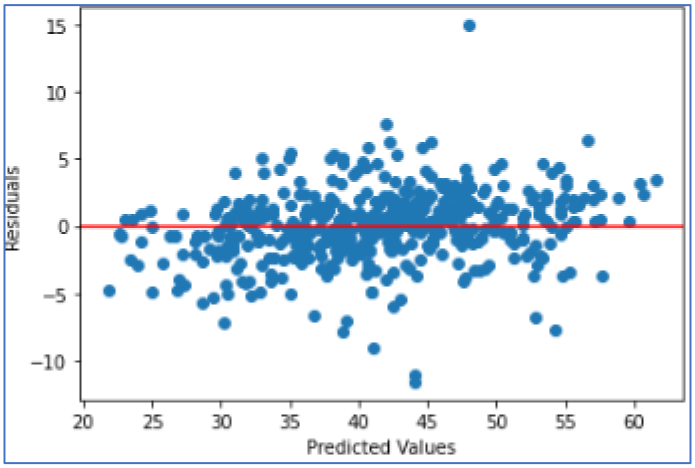


Figure 14: Residuals and Predicted Values Scatter Plot

Null Hypothesis (H0): Homoskedasticity is present (variance of the errors is constant).

Alternative Hypothesis (H1): Heteroskedasticity is present.

The significance value is to measure the results will be (p > 0.05).

The results from the Breusch Pagan hypothesis test came out as follows:

```
'LM Statistic': 25.330831341365997, 'LM-Test p-value': 0.06420153340825306, '
F-Statistic': 1.6133810153024453, 'F-Test p-value': 0.06172424768818806
```

These results indicate that we are unable to reject the null hypothesis (H0) that homoscedasticity is present as our LM-Test p-value and F-Test p-values are greater than the indicated significance (p > 0.05). This provides evidence that there is not a large degree of heteroscedasticity in the model and the variance of the errors is constant. This gives further weight to the reliability of the model and its predictions.

Predictions vs Actuals

The actuals and predictions were plotted in Tableau (fig. 17) to include in the dashboards as evidence of the relationships and accuracy of the model. The overall plot indicated a linear consistency in the actuals vs fitted (predicted) values. Near the bottom we can observe a specific outlier and some scatter. These values represent the crop disaster likely from the Great Flood of 1993. For the soybean model we notice more outliers within the predictions. It is important to note that while the Great Flood of 1993 did impact some states in the analysis, that a majority of the states in this analysis were not within the path of destruction. The outliers within these predictions are likely due to some other accounted factor, which could be pest and disease loss. The soybean commodity is known to be susceptible to many diseases and pests which can results in severe reduction in crop yields and loss. In 2018, more than 556 million bushels of soy were lost due to pest and disease devastation. (Journal of Integrated Pest Management, 2020) A comprehensive study on disease and pest loss will be the next step to continue this initiative.

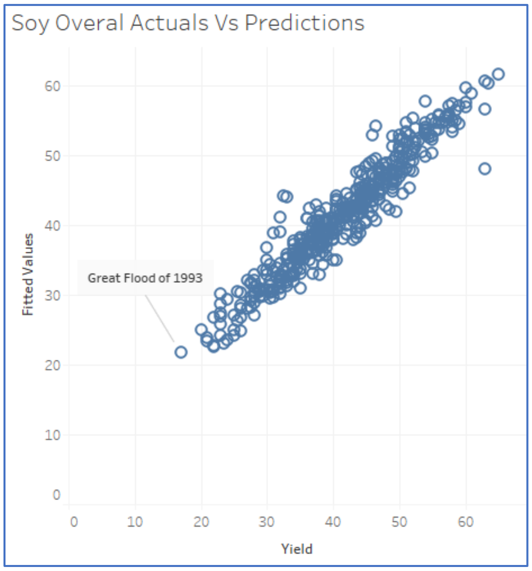


Figure 15: Actuals vs Predictions Scatter Plot

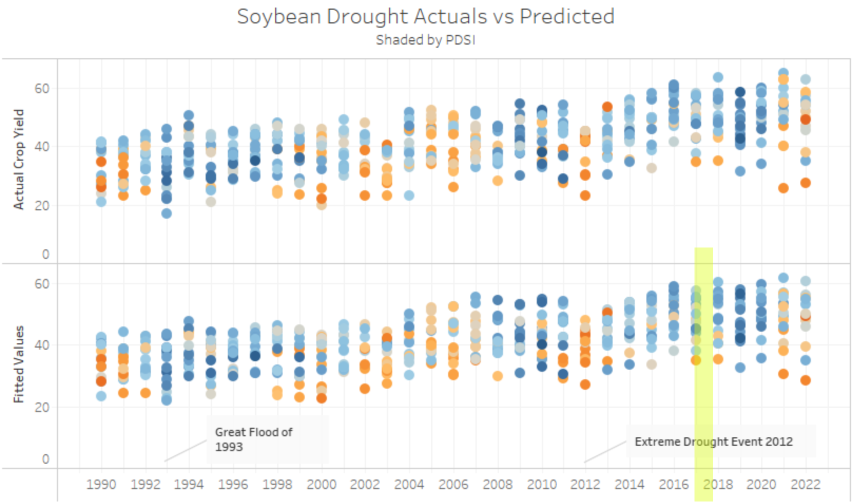
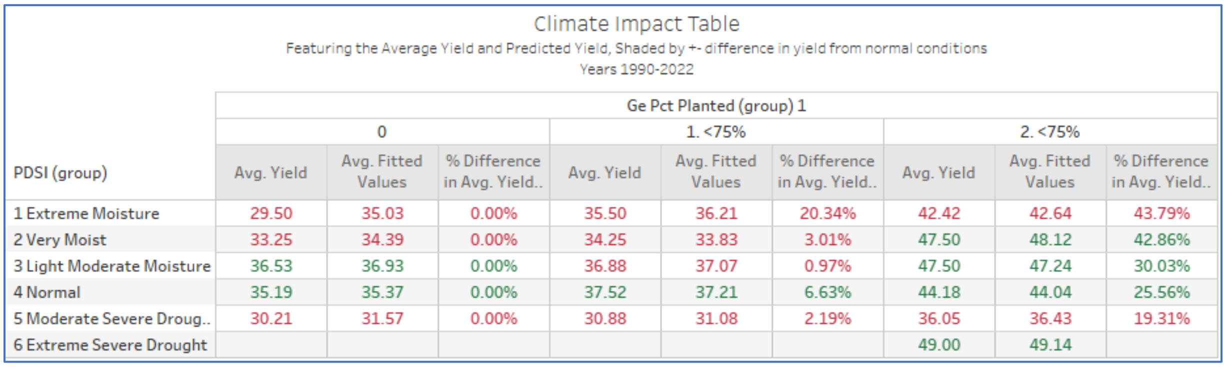


Figure 16: Actuals vs Predicted values chart shaded by the PDSI(Palmer Drought Severity Index). Highlights for severe climate events.

In the next figure (fig. 18)we can observe the actual crop yields plotted compared to the fitted values for each GE seed category and year. The data points are shaded by drought severity – orange/dark red capturing severe drought and dark blues capturing extreme moisture. We can observe the predicted values decreasing where there are severe climate conditions indicated by the PDSI. This aligns with the actuals yields being lower and corresponding with the PDSI severity. The model was able to predict lower yields per acres in periods of extreme or severe dryness that correspond to the actuals. The main time period of interest in 2012, where widespread drought affected many of the states in this analysis. The Great Flood of 1993 is also observable in the visualization, though it is important to note that fewer states were affected by the disaster within the soybean commodity. There is much less reduction in yield than compared to the corn commodity for this time period.

Climate Impact Assessment

A crucial feature of the dashboard that was created is the Climate Impact Assessment. The data was grouped into bins for the percentage of genetically engineered seeds planted to give an indication on the performance of implementation. This includes a table that captures the different degrees of climate conditions, the actual average yields, the average predicted yields, and the difference in the actual yields from the prior genetically engineered planted percentage category.



Climate Impact Table
Featuring the Average Yield and Predicted Yield, Shaded by +- difference in yield from normal conditions
Years 1990-2022

	Ge Pct Planted (group) 1								
	0			1. <75%			2. <75%		
PDSI (group)	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..	Avg. Yield	Avg. Fitted Values	% Difference in Avg. Yield..
1 Extreme Moisture	29.50	35.03	0.00%	35.50	36.21	20.34%	42.42	42.64	43.79%
2 Very Moist	33.25	34.39	0.00%	34.25	33.83	3.01%	47.50	48.12	42.86%
3 Light Moderate Moisture	36.53	36.93	0.00%	36.88	37.07	0.97%	47.50	47.24	30.03%
4 Normal	35.19	35.37	0.00%	37.52	37.21	6.63%	44.18	44.04	25.56%
5 Moderate Severe Droug..	30.21	31.57	0.00%	30.88	31.08	2.19%	36.05	36.43	19.31%
6 Extreme Severe Drought							49.00	49.14	

Figure 17: Climate Impact Table features the average yield per acre against the predicted yield per acre for each climate condition (PDSI) group. A comparison is made on the difference in actual yields for each category compared to zero

We can make clear observations about the data from this table, including the consistent increase in crop yields from ‘Zero’ to ‘<100%’. We can observe that in the zero and <75% genetically engineered seeds planted that performance was reduced in the moderate-severe drought range and the extreme moisture range.

For soy, implementation was extremely fast, with adoption rates soaring past 75% by 2002. The percentage groups were reduced for the soybean category to account for this. We can observe between the zero category to the <75% not much increase in the yield per acre. Though, once adoptions reaches greater than 75%, we can observe a significant increase in crop yields.

In the <75% category, we can also observe a data point in the extreme drought condition, which represents as something of an outlier. This represents one data point for the state of Nebraska, which on average has higher yields than the majority of the other states. Being that there was only one occurrence of extreme drought conditions, it appears that the yield is higher than the moderate-severe drought. Though, we can observe in the moderate-severe drought a reduction in yield compared to the normal conditions and a reduction in the extreme moisture conditions. Overall, the soybean commodity appears to be less impacted by drought conditions than the corn commodity.

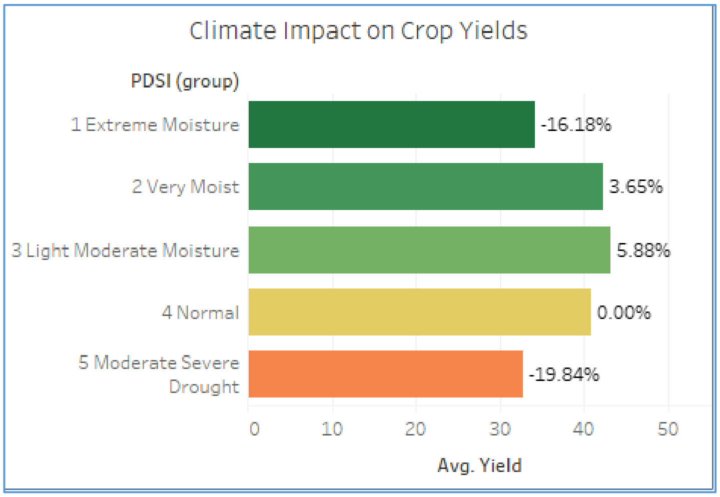


Figure 18: Bar chart of Climate Impact on Yields

A chart was created to assess the impact of potential loss from severe climate conditions. Overall we can observe that it can be expected about a 20% decrease if conditions are in the moderate-severe to extreme drought conditions. For the extreme moisture conditions, we could expect an average of about 16% loss in crop yields. This information can provide us crucial insight into the anticipated loss farmers can expect as severe climate conditions impact. Also, this provides a measurement for us to research the development of new seed technologies to become more resistant to these conditions.

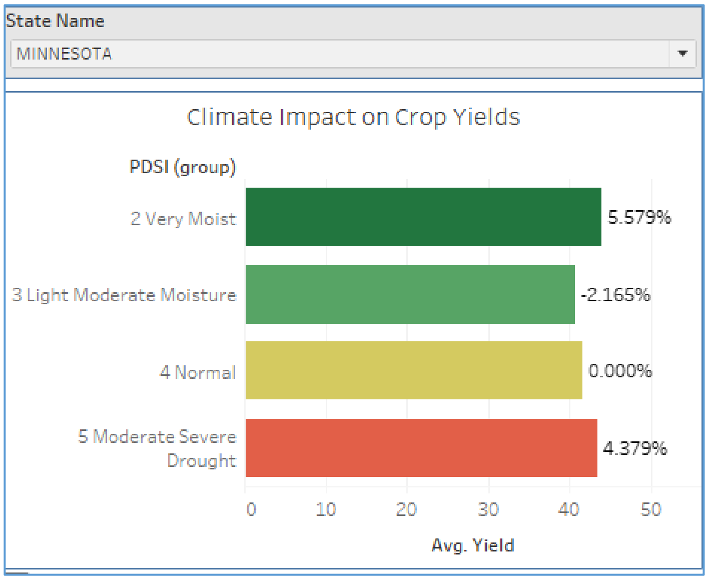
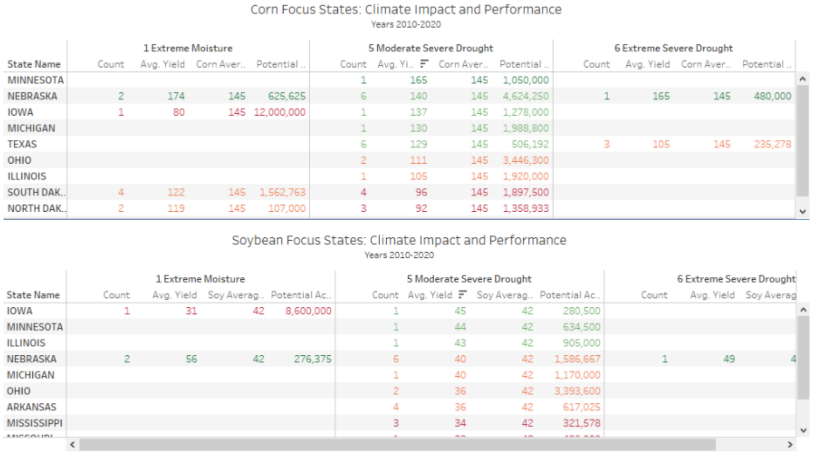

State Name
MINNESOTA
Climate Impact on Crop Yields
PDSI (group)
2 Very Moist
3 Light Moderate Moisture
4 Normal
5 Moderate Severe Drought
5.579%
-2.165%
0.000%
4.379%
0
10
20
30
40
50
Avg. Yield

Figure 19: Bar chart on Climate Impact on Yields for state of Minnesota

The dashboard provides us the ability to look into each individual states results. This provides very useful information into the individual impacts and how they can dramatically differ from the overall assessment. For example, with the state of Minnesota, we can observe a much less severe impact within the moderate-severe drought climate conditions. An average of 4% loss when those conditions exist. We will be able to investigate and look for other features that might be a contributing factor, like mitigation techniques or unique features like soil composition that might increase performance in those conditions. Further, we can test and develop new seed technologies in created environments that mimic specific conditions like what might exist in Minnesota. We can then compare our results in this model to determine outcomes and effectiveness of the developments and technology.

Recommended Focus States: Combined Commodities

As part of the project, creating recommendations for states identified as being significantly impacted more than others for outreach and additional research were visualized in a dashboard. These tables house the states that experienced adverse climate conditions within the past 10 years, with the average yield per acre, the overall yield per acre for the entire commodity and the potential selling acres. The values are color coded based on the yield performance within each climate condition compared to the overall commodity average. This provides information into which states perform better in adverse climate conditions than other states. This will give teams a focus area of where to start on research initiatives generated from this project, they will be able to reach out to the better performing states to gather information on specific circumstances, mitigation techniques and technologies that are employed and use this information to compare to the lower performing states. Mitigation techniques, and technology adoption recommendations could then be created for farmers who plant in the lower performing states to increase the yields during adverse climate conditions. The dashboard also includes potential selling acres, measuring the percentage of acres that are potentially not using seed technology. Our sales teams will be able to use this information to reach out to states to create additional buy in and sell our seed technology to farms that have not implemented seed technology in their crops.



Corn Focus States: Climate Impact and Performance
Years 2010-2020

	1 Extreme Moisture				5 Moderate Severe Drought				6 Extreme Severe Drought			
State Name	Count	Avg. Yield	Corn Aver..	Potential ..	Count	Avg. Yi..	Corn Aver..	Potential ..	Count	Avg. Yield	Corn Aver..	Potential ..
MINNESOTA					1	165	145	1,050,000				
NEBRASKA	2	174	145	625,625	6	140	145	4,624,250	1	165	145	480,000
IOWA	1	80	145	12,000,000	1	137	145	1,278,000				
MICHIGAN					1	130	145	1,988,800				
TEXAS					6	129	145	506,192	3	105	145	235,278
OHIO					2	111	145	3,446,300				
ILLINOIS					1	105	145	1,920,000				
SOUTH DAK..	4	122	145	1,562,763	4	96	145	1,897,500				
NORTH DAK..	2	119	145	107,000	3	92	145	1,358,933				

Soybean Focus States: Climate Impact and Performance
Years 2010-2020

	1 Extreme Moisture				5 Moderate Severe Drought				6 Extreme Severe Drought		
State Name	Count	Avg. Yield	Soy Averag..	Potential Ac..	Count	Avg. Yield	Soy Averag..	Potential Ac..	Count	Avg. Yield	Soy Averag
IOWA	1	31	42	8,600,000	1	45	42	280,500			
MINNESOTA					1	44	42	634,500			
ILLINOIS					1	43	42	905,000			
NEBRASKA	2	56	42	276,375	6	40	42	1,586,667	1	49	4
MICHIGAN					1	40	42	1,170,000			
OHIO					2	36	42	3,393,600			
ARKANSAS					4	36	42	617,025			
MISSISSIPPI					3	34	42	321,578			
[illegible]					[illegible]	[illegible]	[illegible]	[illegible]			

Figure 20: Dashboard snapshot that captures the states that were impacted by adverse climate conditions and shaded by their performance measured in yield per acre

F2. Evaluating Practical Significance

This data analytics solution provides in-depth insight into the performance of the major row crops corn and soybean in the United States. Where our company data is limited to internal statistics and a limited amount of farm data, creating this dashboard as a data analytics solution provides a greater look into agricultural data on a national basis with the utilization of all available agricultural data.

Our model can facilitate the growth of research in expanding our seed technologies by giving us the conditions and investigative paths to access growing conditions. Also, as a climate impact monitoring solution, we will be able to expand and add new data to the model as climate events occur. We will be able to use the model to predict a range of loss based on a forecasted or occurring severe climate event. This will be useful to many teams within the organization to provide information and outreach to our farmers and customers, as well as developing plans for mitigation techniques extracted from investigations prompted by this project. Teams will be able to investigate the unique conditions presented by states, compile information and data on what other conditions and mitigation strategies are used, and also gather information on what other data points would be relevant to the model.

The project results also provide outreach recommendations that can be used by our sales team to focus on states that have low performance and also experience significant impact from climate conditions. We may be able to offer solutions, mitigation techniques and additional buy-in for utilization of our seeds. This has the potential to generate additional revenue for the company.

This solution will also be useful to our public relations and marketing teams for use with public outreach and education. Bringing awareness to the benefits of using seed technology, and the positive impacts it can have on the environment and our food supply is an important mission of the company. Seed technologies allow us to create solutions that can benefit the environment, the economy, and the global food supply, and projects like these help to educate the public on those specifics.

G. Summary

In summary, the conclusions drawn from this project are that the introduction of genetically engineered seeds has had significant impact on crop yields, and this impact has allowed farmers to realize higher yields and less loss during severe climate conditions. The increase in crop yields measured per acre indicate that farmers are able to harvest more ears of corn within one acre of land. Genetically engineered seed technology contributes to this through many different factors like allowing a higher plant population per acre, resistance to pests and disease, and resistance to stresses from climate or other factors. Looking at the impact on a nationwide scale from all states who utilize genetically engineered seed technology provides a different view into the subject, where we are able to capture the differences that exist between states and regions that we do not have data on within our company.

The research and development of new seed technologies through the creation of new traits is an expensive and lengthy process, and the business will require significant evidence that the pursuit of new specific traits are warranted. This Climate Impact Analysis dashboard will provide the evidence of the impact of adverse climate conditions, within specific states and conditions, and that there is significant area for improvement. Further, we will be able to utilize this model in the process of development of the traits to understand what conditions we will need to test, and be able to measure our results. We will be able to investigate conditions that exist in states that are most severly impacted and replicate those conditions to expose our developing plants to, and also have the ability to create different traits for different conditions based on other inputs.

This model can also be used for marketing and outreach, and provide useful insight to our farmers about what kinds of losses they can expect from forecasted or ongoing climate events. We will be able to develop additional tools to provide to our customers and farmers to help them mitigate losses during periods of severe climate conditions. This will also help strengthen our relationship with our customers, creating buy in for when our new technologies release.

The company has a mission to educate the public about the benefits and contributions that genetically engineered seed technology contributes to the world. This dashboard provides information and material that be used as education material and info graphics to distribute the benefits and positive impacts. Where food supply and water supply are both critical economic and environment concerns, seed technologies can provide positive solutions to these issues. Crops that typically employ irrigation techniques to mititate drought will be able to use less water and reduce the impact on local watersheds. Farmer's have to pay for the water they use, so the less water, the less cost it takes to produce a successful harvest. This equals more food, and less impact on the environment. Creating more robust seeds that are tolerant to high stress conditions means increased immunity, meaning they are also less affected by pests and diseases. Robust traits in seeds can then contribute to a reduction in the need for fungicides, herbicides and pesticides that are used as mitigation techniques to combat those problems.

In conclusion, the project take-away is that we have created a tool that allows us to expand and continue to grow in the agricultural industry, while progressing the mission values of the company. We have opportunity to create new and robust technologies and increase our share in the market, as well as creating positive buy in from the public on this subject.

Recommended Course of Actions

The first recommended course of action would be to gather a team to investigate and gather a dataset that captures the pest and disease impact on the corn and soy crops in the United States. This data would provide a crucial contribution to the model, and would facilitate improvements in the predictions. There is no comprehensive dataset that exists with this information, but there are various sources for each state and generalized numbers that could be put together to create a dataset to add to the model. Within this action, project submissions could be made to the USDA and other agencies to propose collection of this type of agricultural data to be added to the NASS databank.

The second recommended course of action would be to gather a team to investigate the recommended focus states identified in the dashboard. Gathering data and information from these states will provide critical information needed for the next steps in the development phase of new seed technologies that are severe climate tolerant. Also, this will provide information for other teams to develop mitigation strategies based on the findings in each state, and provide us more detail into the unique circumstances that might impact a specific state or region. This data could also be used to improve the model and add additional features that may or may not impact a specific state or set of conditions.

H. Panopto Presentation

A Panopto presentation was created and can be viewed at the following link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7bc320a0-8393-4eb2-b17b-b059001f23d8

Tableau Dashboard

A link to Tableau dashboard published in Tableau Public

https://public.tableau.com/shared/Y2T276S5B?:display_count=n&:origin=viz_share_link

I. Sources

Central Midwest Water Science. (2018, August). *Great Flood 1993*. Retrieved from USGS: United States Geological Service: https://www.usgs.gov/centers/cm-water/science/great-flood-1993

Kumar, A. (2023, May). *Pearson Correlation Coefficient & Statistical Significance*. Retrieved from Data Analytics: https://vitalflux.com/pearson-correlation-coefficient-statistical-significance/#:~:text=If%20p%2Dvalue%20is%20less,favor%20of%20the%20alternate%20hypothesis.

National Agricultural Statistics Service. (2023). *Adoption of Genetically Engineered Crops in the US.* Retrieved from USDA: https://www.ers.usda.gov/data-products/adoption-of-genetically-engineered-crops-in-the-u-s

Obergfell, C. (n.d.). *2012 Drought*. Retrieved from National Weather Service: https://www.weather.gov/iwx/2012_drought

Roth, M. G. (2020). *Integrated Management of Important Soybean Pathogens of the United States in Changing Climate*. Retrieved from Journal of Integrated Pest Management, Volume 11, Issue 1: https://doi.org/10.1093/jipm/pmaa013

USDA. (2022, September). *Recent Trends in GE Adoption*. Retrieved from Economic Research Service: https://www.ers.usda.gov/data-products/adoption-of-genetically-engineered-crops-in-the-u-s/recent-trends-in-ge-adoption/#:~:text=Genetically%20Engineered%20(GE)%20seeds%20were,are%20produced%20using%20GE%20varieties.

USDA. (2023). *National Agricultural Statistics Service (NASS)*. Retrieved from Quick Stats: https://quickstats.nass.usda.gov/

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., . . . Arndt, D. (2014). *NOAA Monthly U.S. Climate Divisional Database (NClimDiv)*. Retrieved from NOAA National Climatic Data Center. doi:10.7289/V5M32STR