Impact of Drought on Agriculture

BU-NOAA Capstone Project Final Summary Report

Team 3:
Shamika Kalwe, Selma Sentissi, Jaya Nagesh, and Shuyi Zhu

Date:
16th July 2021

Table of Contents

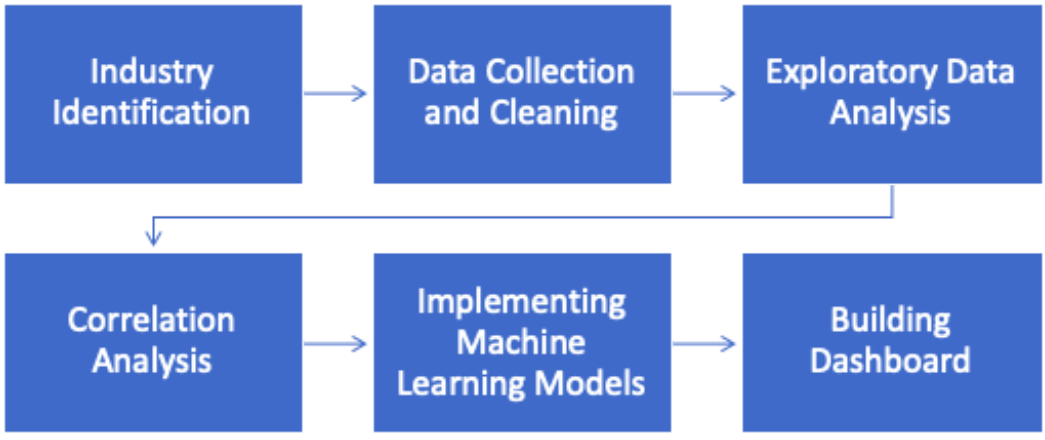
Introduction and Business Problem

In today's era, studying drought and its impact on the agricultural industry is more important than ever before. In simple terms, drought is defined as the absence of water for a prolonged period[1]. As a result, when drought is present, water availability and water quality both reduce, causing a decrease in crop production, an increase in pests/crop diseases, thus all contributing to a negative economic and socioeconomic impact. Within the US, 1,467 counties are experiencing drought currently which translates to 152.6 million acres of crops experiencing drought. In fact, the cost of drought events accounts for 9 billion dollars every year[2]. Hence it is vital for farmers to be notified about drought events so that they can prepare to combat the ill consequences of drought as far in advance as possible.

Hence, the purpose of this project is to increase dissemination of drought related information to farmers in a way that allows them to develop a multi-seasonal drought mitigation strategy. Even though there is vast amounts of data about drought monitoring already available, we believe a customized front-end may drive the users in the farming industry to use the data more effectively. In this project, we have collected relevant data, automated the data collection process, processed and analysed the data and created a front end that delivers relevant information to the farmers.

For this project, we focus our analysis on the New York state and the Corn crop. New York is known to be a leading agricultural state, producing an astounding $5.75 billion in revenue in 2017. Furthermore, it is critical to note that the crop production has decreased since the year 2012 to 6,866,171 acres in production, down from 7,183,576 in 2012, drought being a possible culprit for this decrease[3]. We decided to focus on Corn since it is one of New York's top 10 agricultural products and generated a revenue of $256 million 2017. We have explored the crop yield and drought related effects through the analysis that was conducted.

Methodology



[1] https://www.ncdc.noaa.gov/monitoring-references/dyk/drought-definition
[2] https://www.drought.gov/sectors/agriculture#key-issues
[3] https://www.nyfb.org/about/about-ny-ag

Data Source Identification and Data Collection

We collected the data for agricultural production and different indices for our analysis. We collected and worked with the following data: Temperature, Precipitation, U.S. Drought Monitor Index (USDM), U.S. Department of Agriculture (USDA), Evaporative Demand Drought Index (EDDI), Standardized Precipitation Index (SPI), Standardised Precipitation-Evapotranspiration Index (SPEI) and Palmer Drought Severity Index (PDSI).

Please refer to the Appendix for definition of these indices.

Machine Learning Models

We computed Machine Learning Models on different datasets to try and predict the Yield for future years using drought indicators. We also used feature importance models to figure out which indicators were the most important when predicting Yield. We sometimes needed to tune the models in order for them to be more accurate and thus, more useful.

(A) Different Models Used:

Linear Regression: The first model we tried was a simple linear regression. Since we wanted to keep this model as a baseline, we did not tune it. Generally, the model overfit as expected.

Kernel Ridge Regression: This model also offers a slightly more flexible approach to the linear regression model, hence we decided to try it.

Polynomial Regression: The polynomial regression model is more flexible than the linear regression model. First we tried running the simple polynomial regression model without any tuning (train and predicted on y train. The un-tuned model had a serious overfitting issue (sometimes the r^2 was even 1). Next, we conducted some hyper-parameter tuning using GridSearchCV, which allows us to vary different parameters and select the best ones. We varied things like the polynomial degree, and did a 5-fold cross validation. However, due to computational power limitations in Google Collaboratory, we were unable to vary the polynomial degrees as much as we intended to; we intended to test varying degrees from 2nd to 9th degree, however we could only test 2 to 4. Even with decreasing the degrees, the model still takes several minutes to run.

Random Forest Regressor: The random forest regressor is a flexible model compared to the linear regression model. At first when we did not tune, the model overfitted (not as bad as the polynomial regression however), which is expected from using flexible methods. Then we tuned the model using the GridSearchCV once again. We varied things like the number of estimators, max numbers of features, maximum depth, etc. and did a three-fold cross validation. The tuning did help the overfitting problem. We would have liked to vary even more parameters, however computational power was also a limitation.

Boosting: Boosting is an ensemble method, hence it is made up of multiple algorithms. First we did not tune the model, we used a base estimator of SVR, with the number of estimators fixed at 10. The model did not overfit as badly as the random forest regressor. When tuning the model, we tried different base estimators like Linear regression, KNN, and SVR, varied the number of estimators, and a few other parameters (bootstrap and bootstrap features).

Overall, the random forest regressor seemed to do best across all splitting variations and growing season/annual, indicating that the data is better fit with a more flexible method and there are non-linear patterns within the data.

(B) Results:

We computed the above models with two different datasets. The first dataset includes the different variables averaged for the months from May to September which is the Growing Season for Corn, and the second dataset includes the variables averaged yearly. The indices included in both these datasets are ADSCI yearly value, average precipitation mean as well as the 25th, 50th, and 75th percentile, the average temperature mean as well as the 25th, 50th, and 75th percentile, EDDI in the moving average of 5 and 6 months, the SPI 90, SPI 180, SPEI 90, SPEI 180, and PDSI. We tried three different ways of splitting the data for each one of them

- First, we split the data randomly with a train ratio of 75% of the data, a validation set of 15% and a testing set of 10%.
- Second, we split the data with the train set including the years 2000 to 20014, the testing set includes the years 2015 to 2017, and finally the validation set has the years 2018 and 2019.
- The third way we used to split the data has a test set going from 2000 to 2017, a test set with the year 2018 and a validation set with the year 2019.

This gives us a total of 12 different sets of models: Growing Season Unstandardized with the three splitting options, Growing Season Standardized with the three splitting options, Yearly Values Unstandardized with the three splitting options, and Yearly Values Unstandardized with the three splitting options.

We realized that the models were the most accurate for the yearly dataset, both standardized and unstandardized with the randomized splitting option. The yearly unstandardized dataset got a R^2 of .444 for the Linear Regression model, a R^2 of .420 for the Boosting model, and a .301 for the Random Forest Regressor model. The yearly standardized dataset got a R^2 of .434 for the Linear Regression model, a R^2 of .447 for the Kernel model, and a .431 for the Random Forest Regressor model. Thus, we decided not to go further with the second and third splitting options and focused on the randomized split option.

We also computed the absolute mean error for the models:

Models on Yearly Values Unstandardized	R2 train	R2 validation	R2 test	Error rate
Linear Regression	0.434	0.350	0.444	0.986
Polynomial Regression	1.000	-0.170	-0.064	0.185
Random Forest	0.913	0.443	0.301	0.185
Boosting	0.005	0.490	0.419	0.138
Kernel	0.375	0.263	0.350	0.140

Models on Yearly Values Standardized	R2 train	R2 validation	R2 test	Error rate
Linear Regression	0.434	0.350	0.444	0.881
Polynomial Regression	0.836	-0.168	0.100	2.240
Random Forest	0.917	0.441	0.431	1.975
Boosting	0.583	0.473	0.362	1.429
Kernel	0.431	0.341	0.447	0.881

Models on Growing Season Unstandardized	R2 train	R2 validation	R2 test	Error rate
Linear Regression	0.379	0.421	0.333	0.145
Polynomial Regression	0.990	0.120	0.207	0.149
Random Forest	0.903	0.450	0.439	0.134
Boosting	0.005	0.436	0.409	0.139
Kernel	0.365	0.386	0.310	0.148

Models on Growing Season Standardized	R2 train	R2 validation	R2 test	Error rate
Linear Regression	0.379	0.421	0.328	0.713
Polynomial Regression	1.000	0.011	0.001	2.734
Random Forest	0.908	0.459	0.445	2.383
Boosting	0.561	0.477	0.372	2.038
Kernel	0.375	0.404	0.330	0.795

(C) Feature Importance Methods:

Ordinary Least Squares (OLS):

Yearly:

In the Yearly, the statistically significant variables are ADSCI_yr, avg_precip_std, avg_precip_min, avg_precip_50%, avg_precip_75%, avg_temp_count, avg_temp_50%, max_temp_count, max_temp_min, max_temp_25%, max_temp_50%', EDDI_6, SPI_90, 'SPI_180', 'SPEI_90', SPEI_180, PDSI.

Within those, the avg_precip_min, the avg_precip_std, and the avg_precip_75% are the coefficients with the highest scores with respectively 1312.19, -342.69, and -182.25.

OLS Regression Results

Dep. Variable:	Yield	R-squared:	0.433
Model:	OLS	Adj. R-squared:	0.408
Method:	Least Squares	F-statistic:	17.70
Date:	Wed, 14 Jul 2021	Prob (F-statistic):	1.37e-64
Time:	18:03:21	Log-Likelihood:	-3093.1
No. Observations:	703	AIC:	6246.
Df Residuals:	673	BIC:	6383.
Df Model:	29		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
constant	814.4251	241.168	3.377	0.001	340.894	1287.957
ADSCI_yr	-0.0044	0.001	-5.597	0.000	-0.006	-0.003
avg_precip_std	-342.6886	123.113	-2.784	0.006	-584.421	-100.956
avg_precip_min	1312.1853	432.750	3.032	0.003	462.482	2161.889
avg_precip_25%	73.2988	94.360	0.777	0.438	-111.977	258.575
avg_precip_50%	179.3212	69.625	2.576	0.010	42.613	316.030
avg_precip_75%	-182.2452	52.921	-3.444	0.001	-286.155	-78.336
avg_precip_max	14.8443	14.062	1.056	0.292	-12.766	42.454
avg_temp_count	-7.4526	2.120	-3.515	0.000	-11.616	-3.289
avg_temp_mean	3.5826	2.823	1.269	0.205	-1.960	9.125
avg_temp_std	1.2643	3.755	0.337	0.736	-6.108	8.637
avg_temp_min	0.2160	0.726	0.298	0.766	-1.209	1.642
avg_temp_25%	1.6497	1.099	1.501	0.134	-0.509	3.808
avg_temp_50%	-2.5819	1.172	-2.203	0.028	-4.883	-0.280
avg_temp_75%	1.0117	1.396	0.725	0.469	-1.729	3.752
avg_temp_max	1.7904	1.072	1.670	0.095	-0.315	3.895
max_temp_count	-7.4526	2.120	-3.515	0.000	-11.616	-3.289
max_temp_mean	-0.5721	2.652	-0.216	0.829	-5.780	4.635
max_temp_std	-4.6386	3.582	-1.295	0.196	-11.672	2.395
max_temp_min	-2.0288	0.730	-2.778	0.006	-3.463	-0.595
max_temp_25%	-2.4522	0.987	-2.484	0.013	-4.391	-0.514
max_temp_50%	3.8638	0.854	4.522	0.000	2.186	5.542
max_temp_75%	-0.6436	1.193	-0.540	0.590	-2.986	1.698
max_temp_max	-1.3227	1.006	-1.315	0.189	-3.298	0.652
EDDI_5	4.0896	3.628	1.127	0.260	-3.035	11.214
EDDI_6	22.1277	4.075	5.430	0.000	14.126	30.130
SPI_90	-87.7821	29.821	-2.944	0.003	-146.335	-29.230
SPI_180	58.6655	22.168	2.646	0.008	15.138	102.193
SPEI_90	98.0340	27.598	3.552	0.000	43.845	152.223
SPEI_180	-64.4054	22.484	-2.864	0.004	-108.553	-20.258
PDSI	4.9504	0.955	5.184	0.000	3.075	6.825

Growing Season:

In the Growing Season, the statistically significant variables are ADSCI_yr, avg_temp_mean, avg_temp_25%, avg_temp_max, max_temp_mean, max_temp_min, max_temp_25%, EDDI_6, SPI_90, SPI_180, SPEI_90, SPEI_180, and PDSI.

Within those, the SPEI 90, SPI 90 are the coefficients with the highest scores with respectively 122.570, and -91.479.

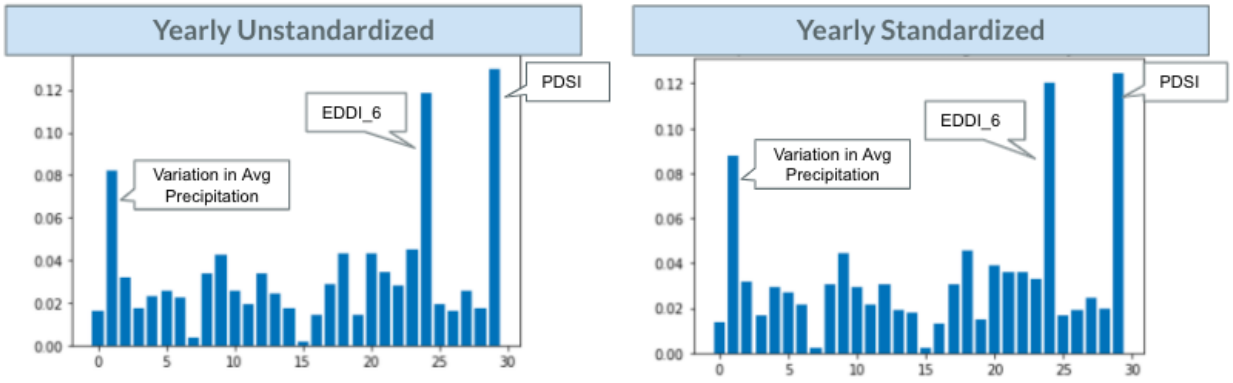
OLS Regression Results

Dep. Variable:	Yield	R-squared:	0.388
Model:	OLS	Adj. R-squared:	0.364
Method:	Least Squares	F-statistic:	16.46
Date:	Wed, 14 Jul 2021	Prob (F-statistic):	7.20e-56
Time:	17:55:28	Log-Likelihood:	-3120.0
No. Observations:	703	AIC:	6294.
Df Residuals:	676	BIC:	6417.
Df Model:	26		
Covariance Type:	nonrobust		

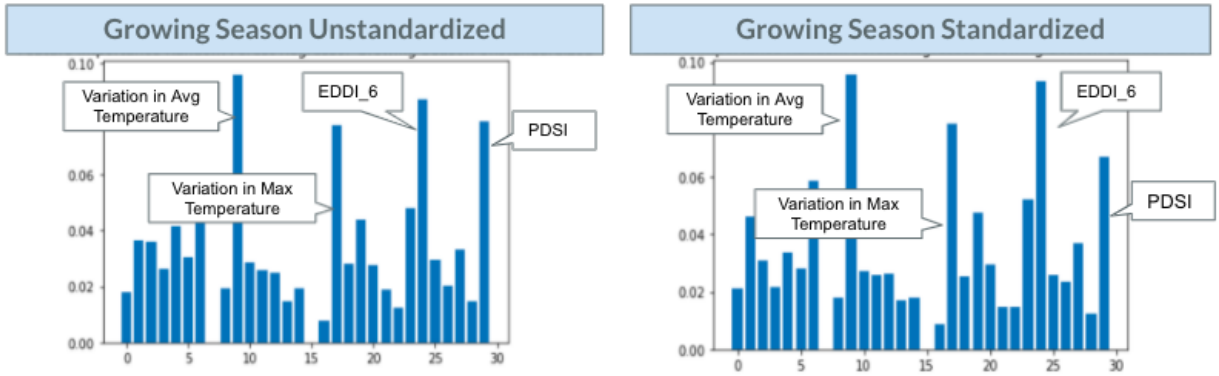
	coef	std err	t	P>|t|	[0.025	0.975]
constant	0.7882	0.868	0.908	0.364	-0.916	2.493
ADSCI_yr	-0.0042	0.001	-5.042	0.000	-0.006	-0.003
avg_precip_std	-285.1059	475.443	-0.600	0.549	-1218.627	648.416
avg_precip_min	-89.7150	157.227	-0.571	0.568	-398.427	218.997
avg_precip_25%	16.9970	85.380	0.199	0.842	-150.645	184.639
avg_precip_50%	-43.5005	49.284	-0.883	0.378	-140.268	53.267
avg_precip_75%	6.7053	68.295	0.098	0.922	-127.390	140.801
avg_precip_max	-7.5151	171.851	-0.044	0.965	-344.940	329.910
avg_temp_count	3.9408	4.340	0.908	0.364	-4.581	12.463
avg_temp_mean	0.8687	0.145	5.984	0.000	0.584	1.154
avg_temp_std	-25.9627	13.992	-1.856	0.064	-53.435	1.510
avg_temp_min	-4.6557	5.150	-0.904	0.366	-14.768	5.457
avg_temp_25%	-7.1817	2.320	-3.095	0.002	-11.738	-2.626
avg_temp_50%	2.3080	1.611	1.433	0.152	-0.855	5.471
avg_temp_75%	3.5479	2.971	1.194	0.233	-2.286	9.382
avg_temp_max	10.3249	4.231	2.440	0.015	2.017	18.633
max_temp_count	3.9408	4.340	0.908	0.364	-4.581	12.463
max_temp_mean	-0.4426	0.154	-2.877	0.004	-0.745	-0.141
max_temp_std	-0.9563	10.587	-0.090	0.928	-21.744	19.831
max_temp_min	-3.0958	3.850	-0.804	0.422	-10.655	4.464
max_temp_25%	5.1198	1.905	2.688	0.007	1.380	8.859
max_temp_50%	-0.4036	1.541	-0.262	0.793	-3.430	2.623
max_temp_75%	-2.3612	2.492	-0.947	0.344	-7.255	2.532
max_temp_max	-1.4720	3.194	-0.461	0.645	-7.743	4.799
EDDI_5	-4.0762	3.820	-1.067	0.286	-11.576	3.423
EDDI_6	24.7226	4.143	5.968	0.000	16.588	32.857
SPI_90	-91.4796	28.038	-3.263	0.001	-146.533	-36.427
SPI_180	42.7286	21.097	2.025	0.043	1.306	84.151
SPEI_90	122.5704	26.213	4.676	0.000	71.101	174.040
SPEI_180	-63.7877	22.639	-2.818	0.005	-108.239	-19.336
PDSI	3.7182	1.129	3.293	0.001	1.501	5.935

Random Forest Feature Importance:

Overall, using the Random Forest Feature Importance, the two most important features are PDSI and EDDI_6 (Appendix 1 (A, B, C, & D)). As far as the Yearly values are concerned, the unstandardized version has an EDDI_6 score of .119 and the PDSI a .130 *(Appendix 1A)*. Regarding the Standardized version, EDDI_6 is up to .121 and PDSI is .125 *(Appendix 1B)*.

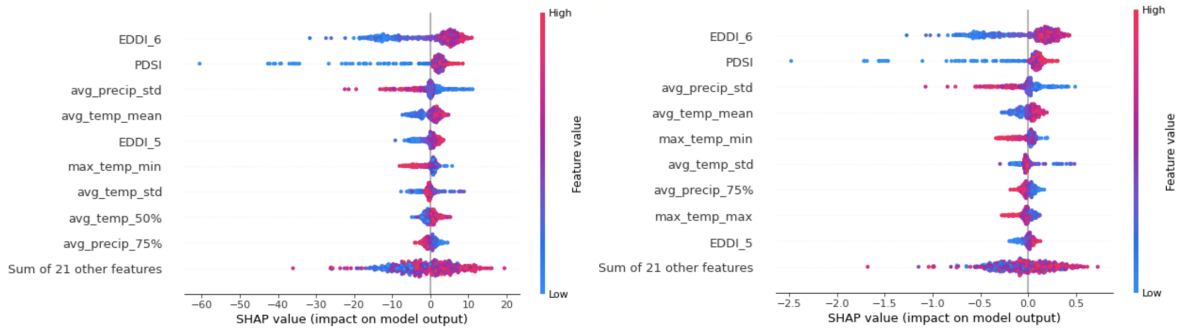


Concerning the Growing Season values, EDDI_6 and PDSI are also important but two other variables are very high: max_temp_std and avg_temp_std. The unstandardized version has an EDDI_6 score of .087, the PDSI a score of .079, an avg_temp_std score of 0.096, and a max_temp_std score of 0.078 *(Appendix 1C)*. For the Standardized version, EDDI_6 is up to .093, PDSI is .067, an avg_temp_std score of 0.058, and a max_temp_std score of 0.079*(Appendix 1D)*.



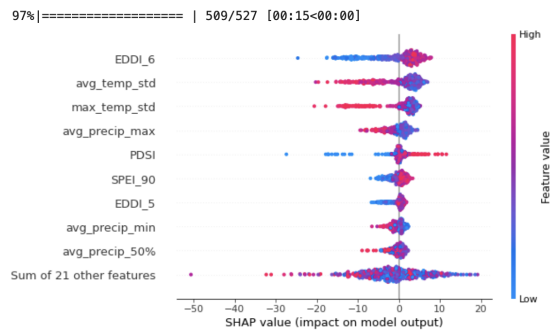
SHAP (SHapley Additive exPlanations):

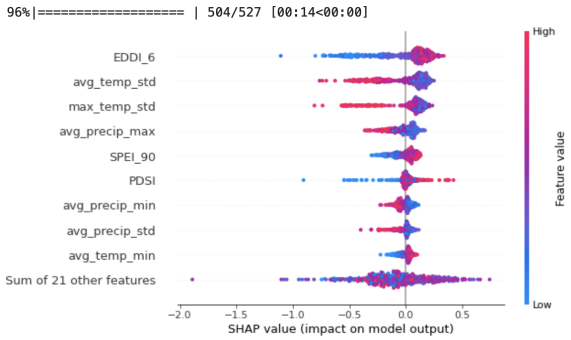
We also computed the feature importance using the SHAP method. Using the yearly dataset, EDDI_6 and PDSI were the most important features. Temperature and precipitation variables are also impactful variables as far as the yearly standardized and unstandardized Random Forest models.



(1) Unstandardized Yearly *(2) Standardized Yearly*

Regarding the Growing Season, EDDI_6 is still the most important feature followed by avg_temp_std. Max_temp_std is the third most important variable using the Growing Season dataset when it is both standardized and unstandardized.





(3) Unstandardized Growing Season *(4) Standardized Growing Season*

Dashboarding

Aim: As mentioned in the introduction, the purpose of this dashboard is to increase dissemination of drought related information to New York Corn growing farmers in a way that gives them information about the performance of Corn and drought metrics across the years and allows them to develop a multi-seasonal drought mitigation strategy and make data driven decisions.

Tools used: We used Tableau Public visualization tool, it is a free software to use. This made it publicly accessible without any charge.

Steps: We converted our excel data to google sheets to import data into Tableau from google drive, thus making the process cloud-based.

Interviews: We interviewed a few industry experts (from Northeast Regional Climate Center and College of Agriculture and Life Sciences, Cornell University) to get their opinion and feedback on the dashboards we built. Overall they had very positive feedback and suggested a few improvements which we incorporated into the dashboards.

Summary: The dashboard contains visualizations and information on the following topics:

Part 1 - Corn related data from USDA

1.1 Corn related information for the state of New York - This Dashboard shows county-wise values and trends of various Corn related metrics like: Acres planted, Sales, Yield etc.. This will give the farmers a complete overview of Corn production in New York across the years and across the counties.

1.2 NY's Top Corn Agricultural districts and Counties - This dashboard explores the Corn related data in terms of its distribution across the top Agricultural districts and Counties of New York. The farmers can see the top districts and the top counties here.

Part 2 - Data about Drought Indicators and Correlation Analysis Among Agricultural and Weather Indicators and Corn Yield

2.1 Temperature and Precipitation Trend across New York counties - This Dashboard shows annual trend of temperature (average of daily average temperature for the week,

maximum of the daily maximum temperature for the week, and minimum of the daily minimum temperature for the week) and precipitation (average of daily precipitation for the week) from 2000 to 2019. In the geographic heatmap, we focus on highlighting the DSCI (Drought Severity and Coverage Index) distributed county-wise along with the exact values of temperature and precipitation.

2.2 New York County-wise Temperature Trend Analysis - The dashboard narrows down the scope from New York State to New York counties. Farmers can select on the specific year county name filters to see the exact past temperature values. They can also gain a quick look on the future forecasted temperature values for each county specifically.

2.3 New York County-wise Precipitation Trend Analysis - Similar to the 2.2 dashboard, this dashboard focuses on the county-wise precipitation. Farmers can select the year and county name to see the exact past precipitation values. They can also gain a quick look on the future forecasted precipitation values for each county specifically.

2.4 DSCI and Agricultural Indicators Analysis for New York counties - This dashboard focuses on agricultural indices (EDDI, SPI, PDSI, SPEI) and drought level (ADSCI) analysis. This Dashboard shows county-wise values of these agricultural inputs and the overall drought severity for each county.

2.5 Modelling Analysis based dashboard - Here we showcase what indices explain the variation in Corn yield the most. This is to give farmers an indication as to which index they should be more aware of as compared to others while trying to plan their next steps.

Dashboard Link:

Part 1: https://public.tableau.com/app/profile/shamika.kalwe/viz/NOAA-USDA/Story1

Part 2: https://public.tableau.com/app/profile/shuyi.zhu/viz/NOAA-Part2_final/Story1

Conclusion

We were able to identify, download, clean, explore and model the agricultural and drought related data to derive actionable insights. In the models that we computed, Random Forest Regressor gave the best results overall. EDDI_6, PDSI, and Temperature variations turned out to be the most important features across the feature importance methods.

We successfully built a Tableau dashboard to facilitate effective information dissemination. As we used Tableau public and used publicly available data, our dashboards are accessible to anyone with the link. This dashboard will equip farmers with all the relevant information at one place and facilitate data driven decision making.

Challenges Faced

The data sources were spread out. Thus, for the Data identification and collection part, having to identify relevant contacts, contacting them and waiting for them to get back to us slightly slowed down the process. In addition, each dataset had a different process to download (different websites and API processes) and had a different format. Thus the data took a lot of effort in terms of downloading, transforming, cleaning, and finally merging to make it ready to use.

One of the issues we ran into was that some of the models took a while to compute and would

sometimes crash after a long period, so we sometimes had to keep checking on our models for them not to crash. Another issue we faced was that whenever we would add another index (eg. EDDI_5, PDSI…), the dataset would not be complete so we would lose some of the values when merging with the final dataset.

We used the Tableau Public as the data visualization tool. Although Tableau Public is a free software to create data visualization charts, it only allows one author per workbook. It is not feasible to merge multiple workbooks done by different authors; thus, Tableau Public may not be an ideal platform for collaborative data visualization tools.

Future Work

It is also worth mentioning that our analysis and model applied with a limited number of weather and agricultural factors. There are other variables such as soil moisture, snowfall, rainfall that can be added to the features selection. Also, some of the agricultural indices we have in the model analysis, such as Palmer Drought Severity Index (PDSI), are the combination of several independent variables. For the further analysis, if we could also include those independent variables in the input variables, it would also make the model more precise. As an extension of our work, it would be a good idea to try to try more tree based models (eg. XGBoost).

Another possible extension would be to extend our dashboards and models to other kinds of crops as well as other states. Due to the limited time scope of this project, we did not reach out directly to farmers and ask them feedback about the dashboards we have. It would be beneficial to collect their feedback and revise the dashboard accordingly since new york corn farmers are the primary business target for this project. It would be important to know if the dashboard contains informative data that they will likely use in their corn growing business.

Accessible Code Repository

https://github.com/BU-NOAA-Capstone/Capstone

Appendix

1 -

(A, B, C, D)

Yearly Unstandardized	
Features	Score
ADSCI_yr	0.01642
avg_precip_std	0.08194
avg_precip_min	0.03193
avg_precip_25%	0.01759
avg_precip_50%	0.02341
avg_precip_75%	0.02584
avg_precip_max	0.02257
avg_temp_count	0.0035
avg_temp_mean	0.0339
avg_temp_std	0.04258
avg_temp_min	0.0254
avg_temp_25%	0.01912
avg_temp_50%	0.03348
avg_temp_75%	0.02447
avg_temp_max	0.01726
max_temp_count	0.00182
max_temp_mean	0.01441
max_temp_std	0.02903
max_temp_min	0.0429
max_temp_25%	0.01434
max_temp_50%	0.04317
max_temp_75%	0.03434
max_temp_max	0.02795
EDDI_5	0.04506
EDDI_6	0.11882
SPI_90	0.01939
SPI_180	0.01649
SPEI_90	0.02568
SPEI_180	0.01719
PDSI	0.13001

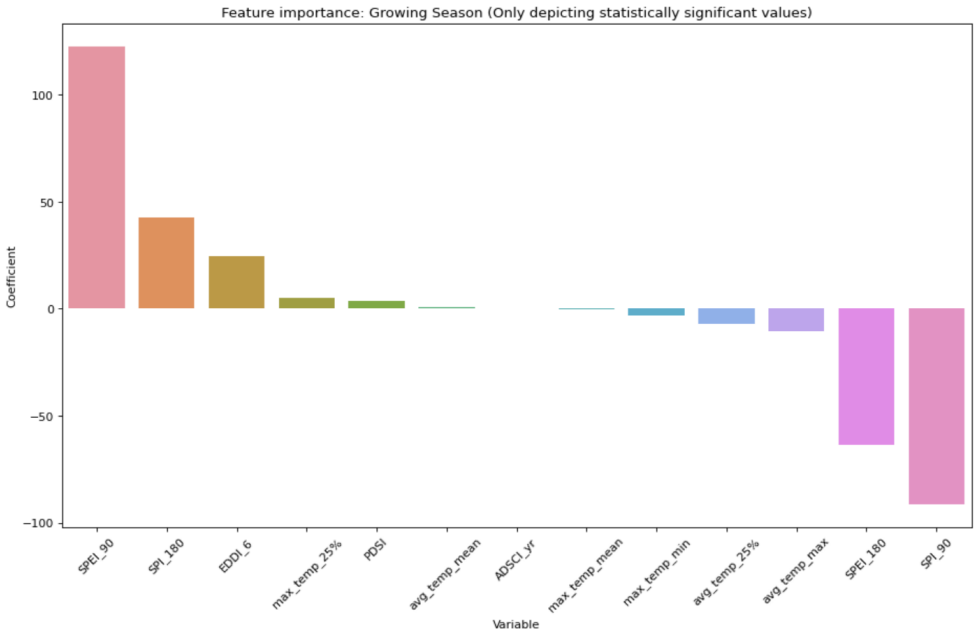
Yearly Standardized	
Features	Score
ADSCI_yr	0.01357
avg_precip_std	0.08758
avg_precip_min	0.032
avg_precip_25%	0.01683
avg_precip_50%	0.02937
avg_precip_75%	0.02716
avg_precip_max	0.02181
avg_temp_count	0.0025
avg_temp_mean	0.03074
avg_temp_std	0.04424
avg_temp_min	0.02935
avg_temp_25%	0.02129
avg_temp_50%	0.03061
avg_temp_75%	0.01941
avg_temp_max	0.01812
max_temp_count	0.00206
max_temp_mean	0.01338
max_temp_std	0.03037
max_temp_min	0.04535
max_temp_25%	0.01507
max_temp_50%	0.03891
max_temp_75%	0.0361
max_temp_max	0.03623
EDDI_5	0.03315
EDDI_6	0.12051
SPI_90	0.01648
SPI_180	0.01889
SPEI_90	0.02436
SPEI_180	0.01989
PDSI	0.12467

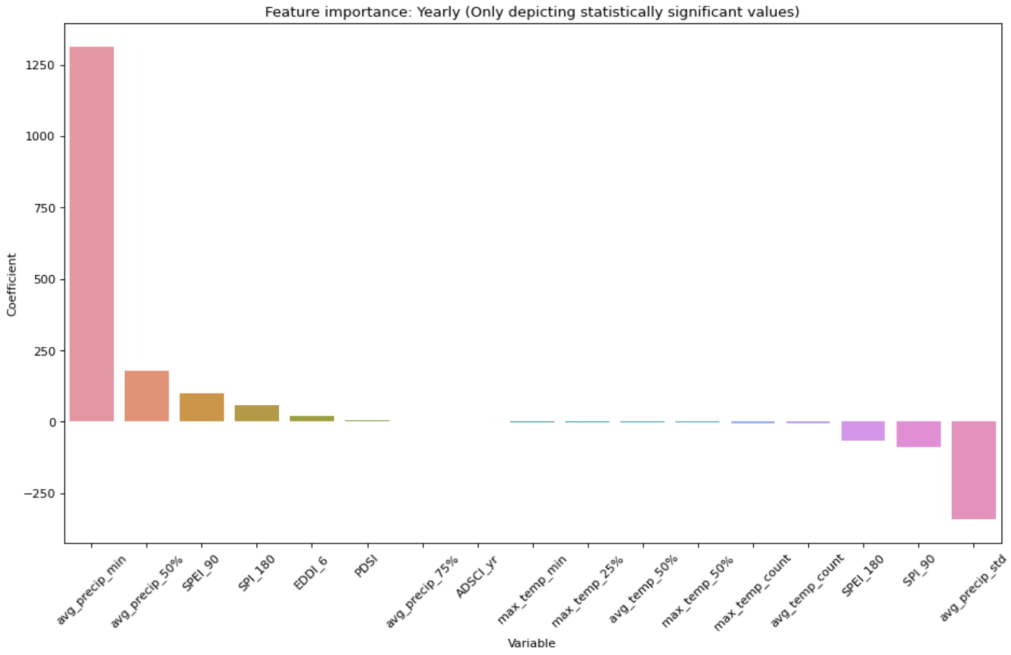
Growing Season Unstandardized	
Features	Score
ADSCI_yr	0.01784
avg_precip_std	0.03668
avg_precip_min	0.03595
avg_precip_25%	0.02624
avg_precip_50%	0.0416
avg_precip_75%	0.03052
avg_precip_max	0.05263
avg_temp_count	0
avg_temp_mean	0.0194
avg_temp_std	0.09593
avg_temp_min	0.02844
avg_temp_25%	0.02579
avg_temp_50%	0.02515
avg_temp_75%	0.01491
avg_temp_max	0.0193
max_temp_count	0
max_temp_mean	0.00777
max_temp_std	0.07774
max_temp_min	0.02802
max_temp_25%	0.0441
max_temp_50%	0.02782
max_temp_75%	0.01906
max_temp_max	0.01237
EDDI_5	0.04796
EDDI_6	0.0872
SPI_90	0.0297
SPI_180	0.0204
SPEI_90	0.03347
SPEI_180	0.01467
PDSI	0.07934

Growing Season Standardized	
Features	Score
ADSCI_yr	0.02123
avg_precip_std	0.04633
avg_precip_min	0.03104
avg_precip_25%	0.02182
avg_precip_50%	0.03371
avg_precip_75%	0.02807
avg_precip_max	0.05877
avg_temp_count	0
avg_temp_mean	0.01797
avg_temp_std	0.09579
avg_temp_min	0.02723
avg_temp_25%	0.02594
avg_temp_50%	0.02613
avg_temp_75%	0.01708
avg_temp_max	0.01781
max_temp_count	0
max_temp_mean	0.00887
max_temp_std	0.07873
max_temp_min	0.02526
max_temp_25%	0.04768
max_temp_50%	0.02952
max_temp_75%	0.01486
max_temp_max	0.01489
EDDI_5	0.05219
EDDI_6	0.09337
SPI_90	0.02588
SPI_180	0.02351
SPEI_90	0.03691
SPEI_180	0.01251
PDSI	0.06688

2 -

Feature Importance:





3 -

Definition of all the indices used:

Temperature and Precipitation:

- About:
  - *Temperature:* We computed the average and maximum daily temperature for every New York county for the years 2000 to 2020 in Fahrenheit.
  - *Precipitation:* We computed the average daily precipitation for every New York county for the years 2002 to 2020 in inches.
- Data Source: We used the Applied Climate Information System API services to download and store the data via a Python script as a JSON file. We collected this data for New York state for the period of 2000-2020.

http://www.rcc-acis.org/docs_webservices.html#griddata

U.S. Drought Monitor (USDM):

- About: It is a composite index, generated weekly, which gives a measure of drought for a particular region. It uses five classifications: None: No Drought, D0: Abnormally Dry, D1: Moderate Drought, D2: Severe Drought, D3: Extreme Drought, D4: Exceptional Drought. A higher value indicates higher drought.
- Data Source: We used the USDM REST API services to download and store the data via a Python script. We collected this data for New York state for the period of 2000-2020

https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx

Accumulated Drought Severity and Coverage Index (ADSCI):

The US Drought Monitor data mentioned above was manipulated to compute DSCI and

eventually ADSCI using following steps:

1. The values of D0 to D4 in the data represents percent of area of that county under a particular drought level or worse
2. Meanings of the Drought levels - None: No Drought, D0: Abnormally Dry, D1: Moderate Drought, D2: Severe Drought, D3: Extreme Drought, D4: Exceptional Drought
3. We combine the above values to get DSCI i.e. Drought Severity And Coverage Index using formula: DSCI = 1(D0) + 2(D1) + 3(D2) + 4(D3) + 5(D4). This helped us get one value for each county for each week.
4. Next, to aggregate USDM over time we compute Accumulated DSCI (ADSCI) for each year using formula: ADSCI = Summation of DSCI from i=1 to n weeks

https://droughtmonitor.unl.edu/About/AbouttheData/DSCI.aspx

Evaporative Demand Drought Index (EDDI):

EDDI is the measure of how anomalous the atmospheric evaporative demand is ("the thirst of the atmosphere"). It is for a given location and across a time period of interest.

https://psl.noaa.gov/eddi/

Standardized Precipitation Index (SPI):

SPI is a multiscalar drought index based on precipitation. A negative value is indicative of drought, and positive value is indicative of wet conditions. It can be computed for a given location and across a time period of interest.

https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/indices/spi/div#select-form

Standardised Precipitation-Evapotranspiration Index (SPEI):

SPEI is a multiscalar drought index based on climatic data that includes both Precipitation and temperature. Similar to SPI, a negative value is indicative of drought, and positive of wet conditions. It also can be computed for a given location and across a time period of interest.

https://spei.csic.es/

Palmer Drought Severity Index (PDSI):

It is a drought index calculated using temperature and precipitation data along with information on the water-holding capacity of soils.

https://www.droughtmanagement.info/palmer-drought-severity-index-pdsi/

Note:

Wherever we have a digit following an index (EDDI_6, SPI_180) it refers to the time period over which this index was computed. Thus, EDDI_6 is the value of the EDDI index computed for 6 months, similarly SPI_180 is the SPI index computed over 180 days.

4 -

References:

1. Vose, E. (2021). Did You Know? | Monitoring References | National Centers for Environmental Information (NCEI) https://www.ncdc.noaa.gov/monitoring-references/dyk/drought-definition
2. Agriculture. (2021). Retrieved 15 July 2021, https://www.drought.gov/sectors/agriculture#key-issues
3. New York Agriculture :: New York Farm Bureau. (2021) https://www.nyfb.org/about/about-ny-ag
4. Drought Severity and Coverage Index | U.S. Drought Monitor. (2021) https://droughtmonitor.unl.edu/About/AbouttheData/DSCI.aspx
5. RCC-ACIS. (2021) http://www.rcc-acis.org/docs_webservices.html#griddata
6. What is the USDM | U.S. Drought Monitor. (2021) https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx
7. Team, P. (2021). Evaporative Demand Drought Index (EDDI): NOAA Physical Sciences Laboratory https://psl.noaa.gov/eddi/
8. Santiago Beguería, F. (2021). Index: SPEI, The Standardised Precipitation-Evapotranspiration Index. Retrieved 15 July 2021, from https://spei.csic.es/
9. Palmer Drought Severity Index (PDSI) | Integrated Drought Management Programme. (2021). Retrieved 15 July 2021, from https://www.droughtmanagement.info/palmer-drought-severity-index-pdsi/
10. Enloe, H. (2021). North American Drought Monitor | Temperature, Precipitation, and Drought | National Centers for Environmental Information (NCEI). Retrieved 15 July 2021, from https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/indices/spi/div#select-form