A Statistical Analysis on the Economic Impact of Droughts

Citadel Datathon
Team 19

Patrick Li Kevin Wang Alex Xiao Richard Zhang

September 29, 2018

1 Executive Summary

Out of all natural resources, water is perhaps the most commonly associated with life itself. Thus, it is no surprise that we treat droughts as one of the most socially and economically threatening natural events. However, not every drought causes the same magnitude of damage; different industries respond differently to the absence of water. Understanding the structure of economic damages, and recovery of specific industries affected by drought allows governments and company managers to better allocate resources to mitigate the threat of droughts.

Our investigation revealed that droughts are in fact a very common phenomenon across the US, with severe negative impacts on the output levels of various industries. More specifically, we found that the American agriculture and manufacturing industries suffer the most severe losses in the presence of long term droughts. However, not every drought results in the same magnitude of economic loss for our industries of interest. Depending on variables such as drought severity, fall weather variance, percentage of agriculture/manufacturing jobs in a region, public water supply, and domestic water use can mitigate or accentuate drought driven economic losses.

Intuitively, as water is a core element in agriculture, a lack of water would lead to a decrease in crop output. Additionally, as the availability of water decreases, the output of labor intensive industries such as manufacturing suffers as employees are unable to match their usual output while worrying about the health implications of water scarcity. As two massive drivers of the US economy, accounting for a total of around $3.5 trillion in annual GDP or 13% of the American GDP (1)(2), the American economy suffers as the American people suffer. However, with proper economic planning and initiatives, municipal and state governments can minimize future drought induced suffering and lift up local economic output.

Contents

2 Introduction and Topic Question

Droughts in the US have riddled the availability of water over the past couple of years. As record breaking heat waves become the norm, droughts have become a common occurrence, leading to heavy damages to counties' and states' economic settings(5). In fact, as of August 2018, 23% of the US is undergoing severe to extreme droughts(6).

The negative impact on one's health and safety is a domino effect. First, agriculture and farming yield is impacted from the lack of water, leading to reduced food production and supply. As a result of basic supply and demand, food and private water usage costs are driven up. Conflicts over the availability of water, dehydration, starvation and even death are all possible consequences. As individuals are struggling with these basic essentials, their ability to work is impacted: leading to negative GDP output on their respective industries.

To gain a understanding of the impact of these events, we will be focusing on the economic impact as a result of droughts. Although other natural disasters such as hurricanes and earthquakes have gained media and Hollywood fame for their destructive power, droughts are in fact the second most disastrous weather disaster, leading to annual losses of $9 billion in the US(7).

This report will seek to answer: **What is the measurable change economic measures over the duration of a drought and its recovery period?**

3 Exploratory Data Analysis

Before performing inference, we analyzed the distribution of drought duration.

We notice an almost tri-modal distribution in drought lengths; the plot somewhat resembles a mixture of three Gaussian distributions with widely varying means. Motivated by the quasi-mixture normality of the drought length distribution, we attempted to cluster counties by predictors of drought induced economic loss. Unfortunately, this avenue did not reveal much information beyond geographical relationships between counties, so we decided to pursue other methods of analysis.

3.1 Drought Severity Distributions

United States Drought Monitor classifies a short term droughts as events where $DSCI$ remains high for 6 months or lower.

Plotting the distribution of short term droughts severity between 2010 and 2016, we notice the distribution is heavily right skewed. The distribution of short term drought duration was incredibly similar. Due to the heavy right tails of these distributions, most short term droughts in the United States appear to be of relatively low severity, persisting for two weeks or less. We do not anticipate such low duration events will have significant influence

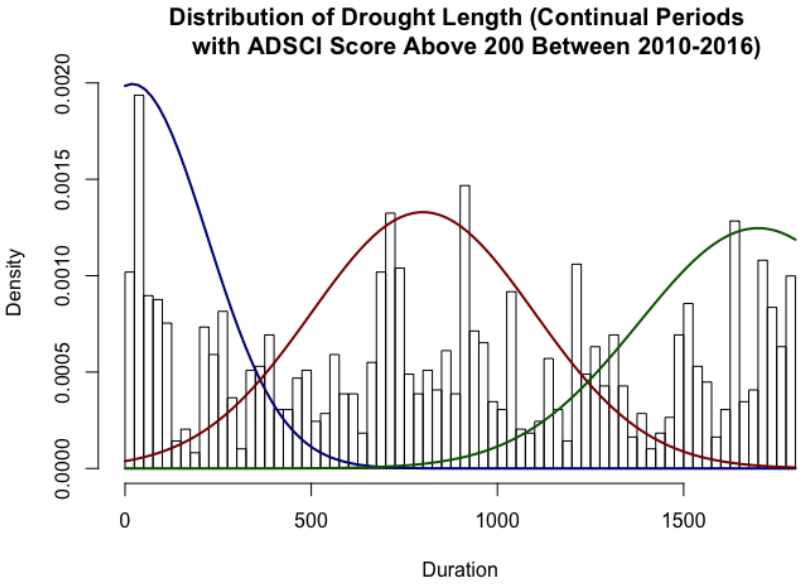
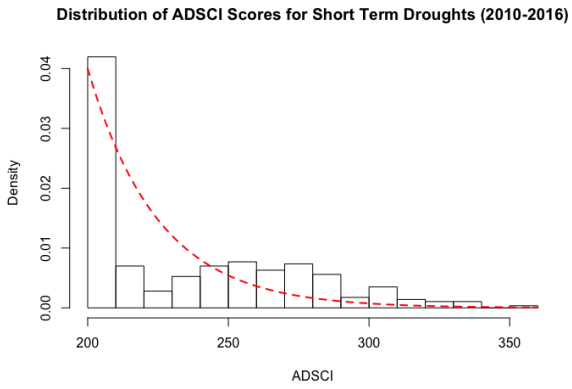
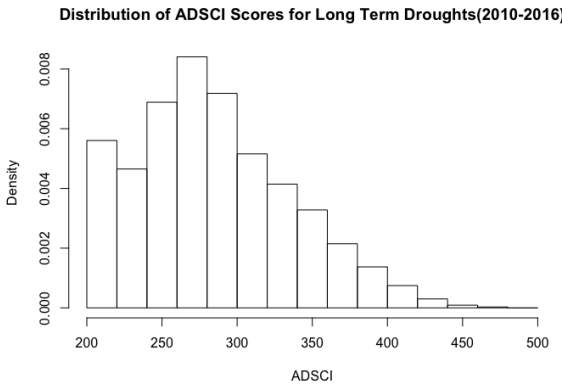


Figure 1: Distribution of Drought Length in Days Between 2010 and 2016 Juxtaposed with Three Normal Distribution Curves, $X_1 \sim \mathcal{N}(20, 200), X_2 \sim \mathcal{N}(800, 300), X_3 \sim \mathcal{N}(1700, 320)$



(a) Distribution of Short Term Droughts vs. Exponential Curve $X \sim Exp(\frac{1}{25})$



(b) Distribution of Long Term Droughts

Figure 2: Short and Long Term Comparisons

on agricultural and manufacturing jobs, and thus focus on analysis on long term droughts, which persist for 6 months or more.

Though the distribution of long term drought severity is also right skewed, the right tail is far less pronounced.

3.2 Geographical Visualizations

In Figure 3, through a visualization of drought severity across the US, we notice that certain regions, particularly certain Southern and Southwestern states such as Nevada and Texas experience more severe droughts by average DSCI (ADSCI) score when compared to other states. This is in line with our expectations as southern states are known to be drier and warmer.

Next, as the severity of droughts increases, we would expect greater negative impacts on industry growth. To confirm this hypothesis, we plotted the average industry growth by state normalized by the average national growth in Figure 4, where we can see that there is an approximate inverse relationship between industry growth and drought severity. The industry growth for states affected by severe droughts such as California fell below the national average growth rate, while other states such as North Dakota exceeded the national average.

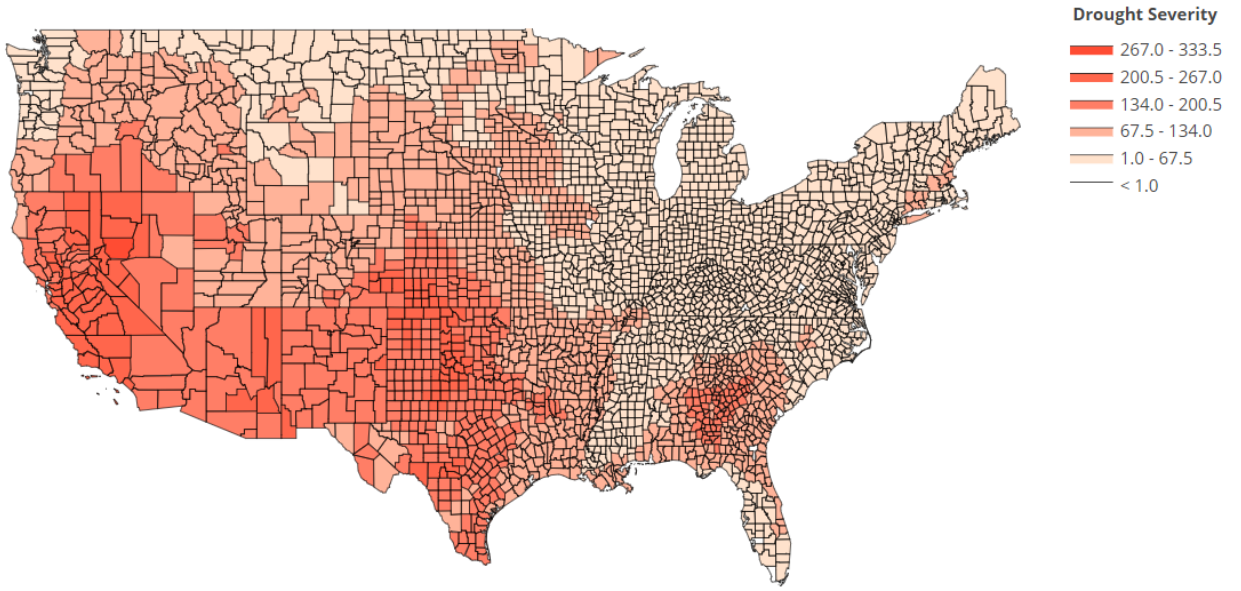


Figure 3: ADSCI scores across the U.S.

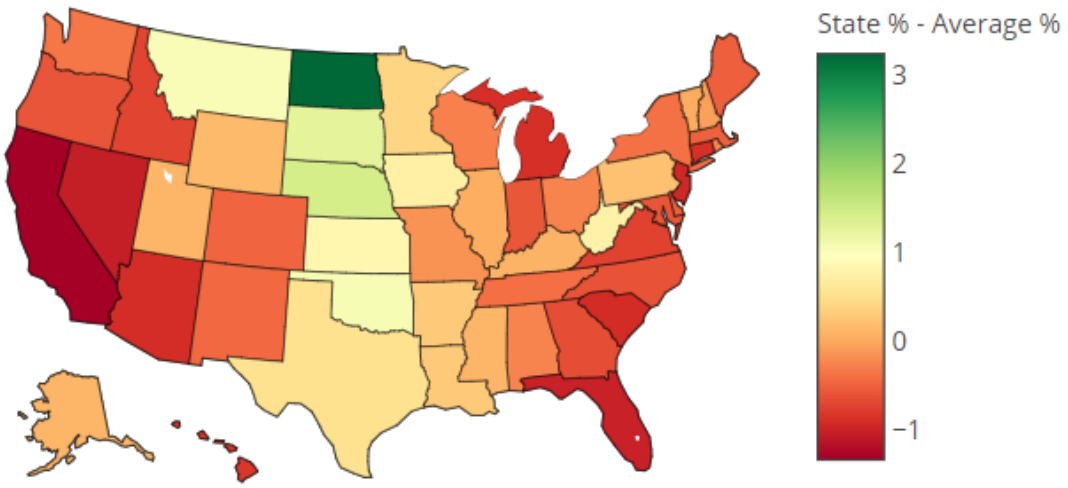
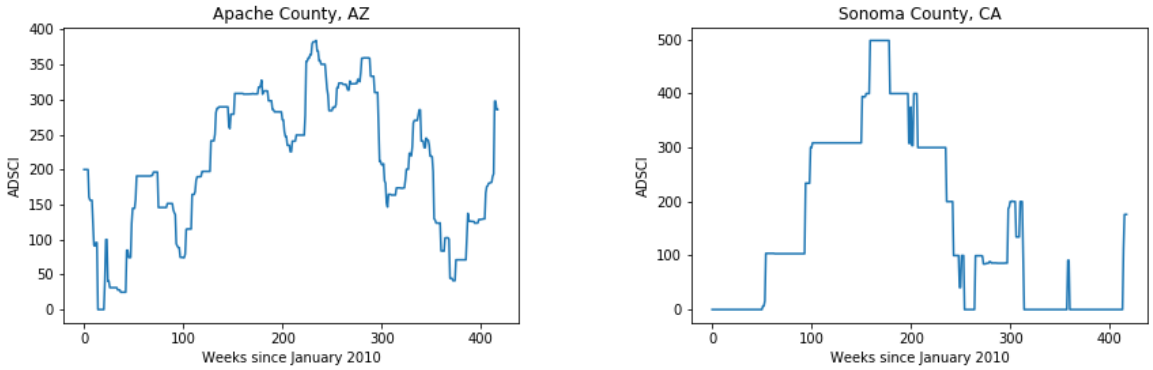


Figure 4: Average State Industry Growth vs Average National Industry Growth

3.3 Time Structure of Drought Severity

We notice that fluctuations in drought intensity naturally forms a collection of time series data. In a later section, we further explain a drought severity statistic that is useful in quantifying percentage of population affected drought. In this subsection, we briefly explore the behaviour of this statistic. We sample a few counties from relatively more drought-stricken areas like California and Arizona.



(a) Apache County, AZ

(b) Sonoma County, CA

Figure 5: Examples of ADSCI Time Series

We fitted simples models like low order ARIMA and regime-switching models, like self-exciting threshold autoregression (SETAR). The Box-Jenkins identification gave very vague signals as to the order of the model, and SETAR models resulted in much lower Bayesian Information Criterion (BIC). This is not unexpected as droughts are largely related to long-time-range climate patterns. Since the focus our study is economic loss estimation, and

given that future climate patterns are difficult to foresee with provided data, we decided that forecasting methods are easily incorrectly specified or contain high variance.

3.4 Final Decision

After performing feature engineering, we used Pearson's Correlation Coefficient to explore correlations between annual changes in education factors, annual changes economic factors, and drought severity. We noticed strong correlations between economic loss in agricultural and manufacturing industries and drought related measures, and decided to further analyze relationships between droughts and the aforementioned industries.

4 Feature Engineering and Feature Interpretation

4.1 Drought Severity Classification Index

The Drought Severity Classification Index $DSCI$ is a measure used by United States Drought Monitor to classify drought severity.

$$DSCI = D_0 + 2D_1 + 3D_2 + 4D_3 + 5D_4$$

The higher the $DSCI$, the more severe the drought. The provided *drought.csv* dataset details the percentage of time that each county is experiencing a D_0, D_1, D_2, D_3, or D_4 event within any specified week between 2010-2016. We defined an arbitrary defined weeks with an average $DSCI > 200$ as time periods with significant droughts. Taking all continuous weeks with $DSCI > 200$, we aggregated drought events by average $DSCI$ throughout the drought period to build our later models.

4.2 Feature Engineering & Model Data Generation

Since we wished to investigate the economic loss in the manufacturing and agriculture industry caused by drought, we sought to rigorously quantify this value. We defined a drought event, $x^{(i)}$, to be a contiguous time frame of ADSCI values exceeding 200, such that the drought lasts between 180 days (6 months) and 1,800 days (5 years). We have chosen these criteria to identify significant droughts and study economic loss upon termination of drought. This provided us with 1,700+ data points.

For the i-th drought event, $x^{(i)}$, we compute the total manufacturing and agriculture income loss rate, $I^{(i)}$, from the start year to the end year of the drought. We also compute the total income loss rate, $T^{(i)}$ of other industries we believe are mildly or not affected by drought, such as arts & entertainment, informatics, health & education, and retail sale. We compute total income by multiplying the number of people employed in the industry and the industry's median income. We treat the median income as an approximate realization

of the mean income estimator. We have that

$$I^{(i)} := \frac{\text{total manu. \& agri. income at end} - \text{total manu. \& agri. income at start}}{\text{total manu. \& agri. income at start}}$$

$$T^{(i)} := \frac{\text{total non-drought income at end} - \text{total non-drought income at start}}{\text{total non-drought income at start}}$$

We claim that $T^{(i)}$ is representative of the normal economic trend during the drought. We "de-trend" the manufacturing and agriculture loss rate by subtracting $T^{(i)}$ from $I^{(i)}$. We have that the loss rate explained by drought is

$$R^{(i)} = I^{(i)} - T^{(i)}$$

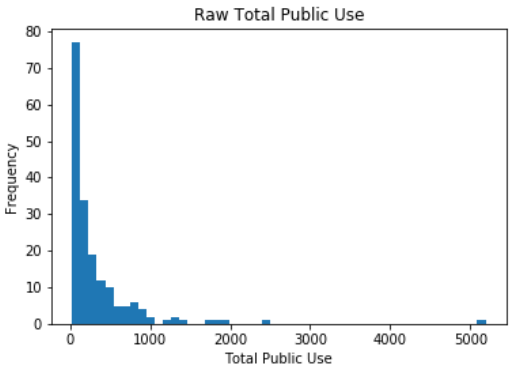
and that the total loss caused by the drought is

$$L^{(i)} = R^{(i)} \times \text{total manu. \& agri. income at start}$$

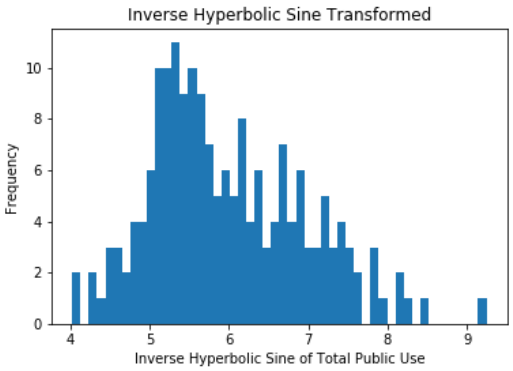
Due to missingness in datasets, we are left with 350 data points with valid $L^{(i)}$ values. Out of these, we have that 183 drought events with negative loss, and 167 drought events with positive loss, but are less than one standard deviation (not significantly different from zero). The average loss, $\bar{L}$ is -29 million. Therefore, we conclude that droughts exert a strong influence in shifting the manufacturing and agriculture total incomes away from the economic trend.

We proceed by investigating the 183 data points with negative loss. For each drought event, we generate the following features.

- Severity, the average ADSCI value of the county during the drought.
- Length of drought in years.
- Percentage of the county's workforce employed in agriculture at start of drought.
- Percentage of the county's workforce employed in manufacturing at start of drought.
- The total number of people employed in county at start of drought.
- The median income of county at start of drought.
- Water usage statistics of the county in 2010 (domestic total use, public total use, irrigation use, thermal power generated).
- Average and variance of the county's seasonal (4 season) Palmer's index over the period of the drought. Palmer's index is a summary of temperature and precipitation.



(a) Distribution of Total Public Use



(b) IHS Transformed Distribution

Figure 6: Effect of IHS on Heavy Right Tailed Distributions

For the water usage variables, we notice a heavy right tail of small number of counties using a large amount of water. The empirical distribution of these variables resemble a Pareto distribution. To lessen any undue influence caused by outliers and extreme values onto our later models, we apply an inverse hyperbolic sine function to these variables, and obtain a more symmetric and tighter distribution. The inverse hyperbolic function is defined as

$$y = arcsinh(x) = \ln(x + \sqrt{1 + x^2})$$

We obtain a 19-dimensional feature vector for each of the drought events, producing a 183×19 data matrix.

5 Modeling Approach

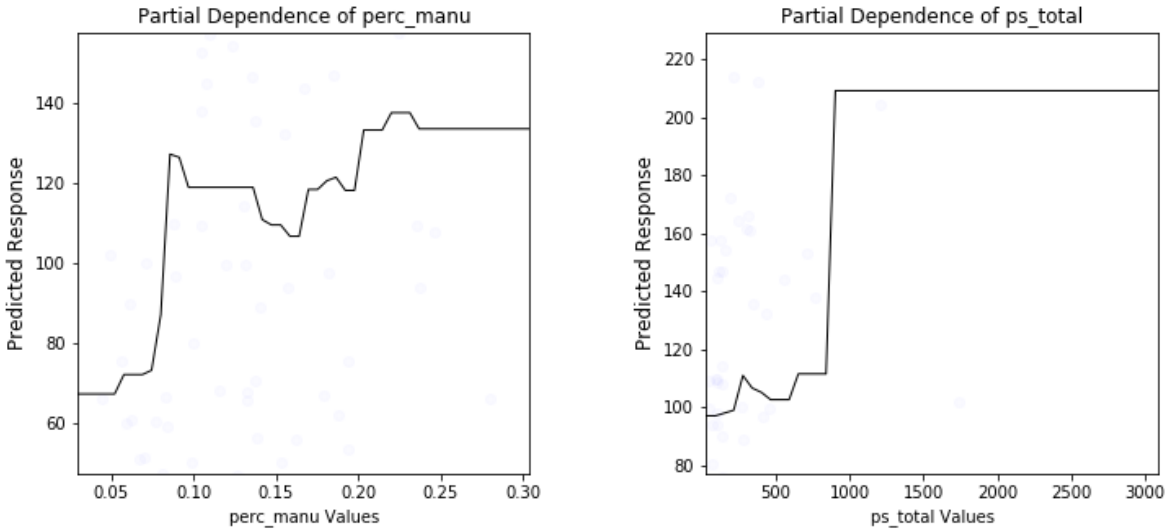
We employ a gradient boosting regressor to model total loss using the 19 covariates identified in the section above. We obtain a 30% training and test split from the data matrix, and use early stopping on test set to determine the number of boosting iterations. Other model hyperparameters are fine-tuned using 5-fold cross validation.

We use the regression model mainly as a method of identifying important driving factors of economic loss and variable interaction. Using the univariate and bivariate analysis in Section 6, we build another parsimonious model with reduced features, and found similar performance.

6 Analysis of Results

6.1 Partial Dependence Plots

Partial dependence plots (PDP) are frequently used to analyze feature importance in gradient boosting machines (GBM).(3) By fixing the values of selected features in set S, we can



(a) Percentage of Manufacturing Jobs in County

(b) Public Water Use

Figure 7: Effect of IHS on Heavy Right Tailed Distributions

marginalize out the impact of all other features in $\bar{S}$. For a fitted model $\hat{f}$, we define partial dependence as follows:

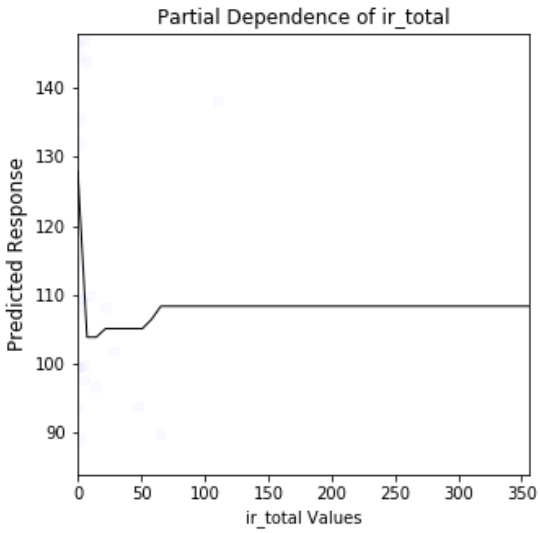
$$\hat{f}(S) = E_{x_{\bar{S}}}\Big[\hat{f}(x_S, x_{\bar{S}})\Big] = \int \hat{f}(x_S, x_{\bar{S}}) dP(x_{\bar{S}})$$

By integrating out the effects of all other variables, we see the sole impact of variables from set S in our predictor function. A large range of prediction values indicates that a feature is important. Unfortunately, for this dataset, the integral is intractable, so we approximated each PDP using Monte Carlo Simulation.

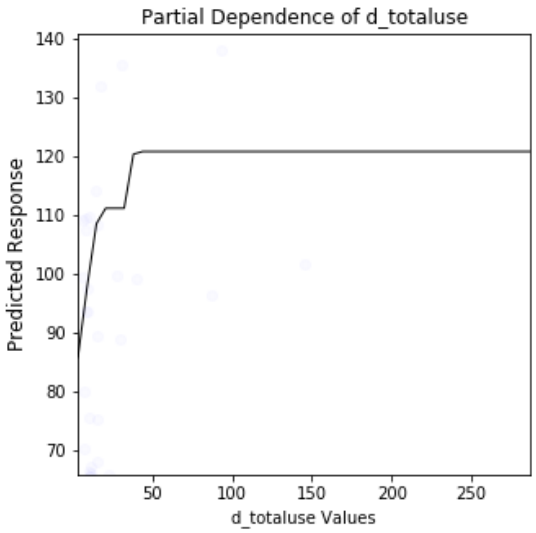
$$\int \hat{f}(x_S, x_{\bar{S}}) dP(x_{\bar{S}}) \approx \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_{\bar{S}}^{(i)})$$

From the Partial Dependence Plots in figure 7, we notice that features such as percent of manufacturing jobs in a state and the amount of public water available heavily influence drought impact on loss in manufacturing and agricultural income. Conversely, from figure 8, we see that features such as non-crop irrigation and state domestic use have rather flat partial dependence plots, and hence are less significant features.

Figure 9 shows that *totalemployed*, *pstotal*, *percmanu*, *severity*, and *fallvar* are most indicative of income loss in agriculture and manufacturing. Influence from *severity* confirms our hypothesis of the impact of droughts on our two industries of interest. The influence of *totalemployed*, *percmanu*, and *fallvar* are quite intuitive. The more employees and manufacturing jobs in a state, the more losses should occur. Furthermore, a large proportion of American produce is grown during fall months; weather variance during these months



(a) Total Couty Non-Crop Irrigation



(b) Total County Domestic Water Use

Figure 8: Effect of IHS on Heavy Right Tailed Distributions

	range	variable
5	214.368439	total_employed
8	112.226486	ps_total
1	70.415085	perc_manu
0	52.070465	severity
4	37.812241	fall_var

Figure 9: Top 5 Most Important Features from PDP Analysis

(*fallvar*) should influence income loss in during a fall drought. On the other hand, the impact of *pstotal* and is quite surprising. One would think that the higher public water supply available, the more water there exists to mitigate the drought. However, people in locations with more *pstotal* may become more reliant on a plentiful supply of water, and are more wasteful of a limited supply during times of drought.

PDPs make inherent assumptions on the independence between variables, which is often unwarranted without sufficient niche knowledge. To mitigate these shortcomings and supplement the findings from our PDPs, we analyzed feature interactions.

6.2 Feature Interactions

Features within a fitted GBM model often interact with one another in a complex, non-linear manner. Beyond adding interpretability to our PDPs, uncovering the effect of significant feature interactions achieves two important objectives. Firstly, we can improve model fit by adding interaction terms, but more importantly we can gain a deeper understanding behind phenomenon that mitigate income loss following droughts droughts.

We modelled all possible two-way feature interactions between variables using the Friedman and Popescu's H statistic (4). We formulate H on variates x_j and x_k as follows:

$$H_{jk}^2 = \sum_{i=1}^{n} \Big[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \Big] / \sum_{i=1}^{n} PD_{jk}(x_j^{(i)}, x_k^{(i)})$$

The statistic can range from 0 to 1, with 0 indicating no interaction between x_j and x_k, and 1 indicating that both x_j and x_k have constant partial dependence plot curves with all interaction stemming from influence from the other variable.

We see surprisingly strong correlation between domestic water use and Palmer's Average in fall (a proxy for variance of weather in fall). This implies that high domestic water use or unpredictable fall weather on their own may not be detrimental, but can cause severe damage to agricultural income in conjunction. For $\frac{143}{\binom{19}{2}} = \frac{143}{171} \geq 83\%$ of pairwise interactions, the Friedman's H value was less than 0.1, which signifies high degrees of independence between variable influence on the prediction variable (income loss in agriculture and manufacturing industries). This indicates that the partial independence assumptions used in our PDPs and the conclusions drawn about feature importance is quite reasonable.

7 Discussion

As droughts continue to rampage across the US, American industries have suffered as a result. By quantifying the severity and frequency of US droughts through the Drought Severity Classification Index across, we were able to identify regions that were most frequently undergoing severe droughts. Then, by honing into the industry outputs of these regions, we discovered through a gradient boosting regressor that the effect of long-lasting droughts

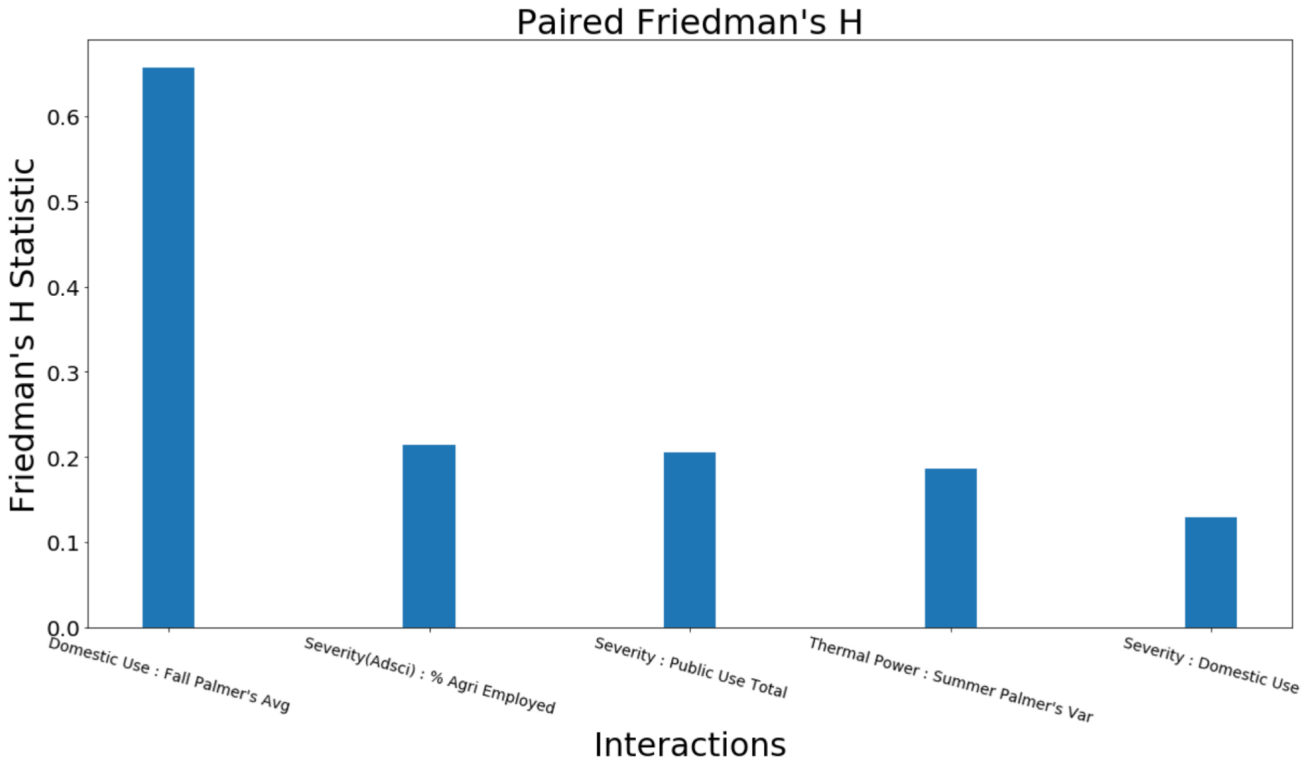


Figure 10: Interesting Paired Interactions from Friedman's H Analysis

on the agriculture and manufacturing industries were most severe and impactful, where features importance was then measured using Partial Dependence Plots. We concluded that the most important features impacting the economic loss in these industries included: total industry employees, public water supply, percent of manufacturing employees and fall weather.

It is common knowledge that global temperature averages have been increasing in the last decades as a result of climate change. Although nations have begun discussing solutions to combat this phenomenon, the instability in climate has already began to take form. In recent summers, lengthy heatwaves spanning North America have become a common occurrence. If global measures are not implemented in the upcoming years to combat this phenomenon, the frequency and severity of natural disasters can only be expected to increase. Then in order to maintain current levels of societal growth and standards of living, the United States of America and the rest of the world need to adapt to withstand the negative effects of droughts.

If provided with long term data (10+ years), we can perform temporal analysis on time series to forecast expected short or long term drought losses in different regions. Using measures similar to Value at Risk (VaR), governments and businesses can be informed of expected economic impacts of drought. With the proper preparation and warnings, they can minimize the economic suffering of counties ridden with drought.

8 Appendix

8.1 Gradient Boosting Hyperparameters

- max-depth: 3
- learning_rate: 0.1
- min_child_weight: 1
- gamma: 0.1

References

[1] National Association of Manufacturing. Top 20 Facts about Manufacturing. NAM. 2018 http://www.nam.org/Newsroom/Top-20-Facts-About-Manufacturing/

[2] United States Department of Agriculture. Ag and Food Sectors and the Economy. USDA. 2018 https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/ag-and-food-sectors-and-the-economy/

[3] Friedman, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 2001 JSTOR, 1189–1232.

[4] Friedman, Jerome H, and Bogdan E Popescu.Predictive Learning via Rule Ensembles. The Annals of Applied Statistics. 2008 JSTOR, 916–54

[5] Miller, Brandon. Deadly heat waves becoming more common due to climate change. CNN. 2017 https://www.cnn.com/2017/06/19/world/killer-heat-waves-rising/index.html.

[6] National Centers for Environmental Information. State of the Climate: Drought for August 2018. NoAA. 2018 https://www.ncdc.noaa.gov/sotc/drought/201808.

[7] National Centers for Environmental Information. DROUGHT: Monitoring Economic, Environmental, and Social Impacts. 2018 https://www.ncdc.noaa.gov/news/drought-monitoring-economic-environmental-and-social-impacts

[8] E., H. North American Drought Monitor. Drought Indices and Data. 2018 https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/indices