Ensuring Clean and Abundant Water Sources for Public Health

Yiyang Wen, Nicky Lim, Ruqi Zhang, and Haoqi Yao

November 10, 2018

Abstract

This paper aims to inform the authorities of the state of water management in the US to make data-driven decisions. In our findings, we found nitrates and arsenic to be the leading chemicals contributing to a poorer quality of life and shorter lifespan. We investigated the BP Oil Spill in 2010 which contributed to the higher arsenic levels in the Gulf of Mexico and affecting the surrounding states. The high nitrate output was found to be from agricultural activities using SHAP, a unified approach to explain outputs of our model. We identified at-risk counties to unclean for the authorities to quickly respond and implement the appropriate programs. We also found that a lower quality of life necessitates lower income and education level. In our analysis of droughts, we found that life stock water usage and domestic water usage have stronger correlation with the droughts in each county comparing with other features.

1 Introduction

Access to clean drinking water is essential for numerous daily activities such as irrigation of crops, basic hygiene and medical care. The World Health Organisation (WHO) Millennial Development Goals 7: Ensure Environmental Sustainability (MDG 7) aims to help people gained access to improved drinking water sources. The world has met the target of halving the proportion by 2015 and WHO is still trying to helped 2.1 billion people worldwide to gain access to clean water sources [1]. With such important uses, it is important for the US authorities to be cognizant of the water situation in the US so that they can respond to any threats and make informed choices in public planning. This leads us to the topic question: **Where are the pain points in water within the US that requires support and intervention, and how can we address it?**

To answer this question, we first sought to answer the smaller questions:

1. Which chemicals are the one threatening the community and which are the results of our industries?

2. Which industries cause these chemicals and what can we do about them?
3. Which areas will face threats to clean water supply in the future?
4. How does earnings and education influence quality of life?
5. Which areas are affected by droughts and what are the intervention to avoid droughts? Which areas are threaten of the water supply due to the droughts?

We used regression and machine learning models to infer causality and observe trends to provide actionable insights that authorities can control.

2 Key Findings

The data revealed that Southern and Midwestern states have a poorer standard of health. There is correlation that it is a result of poorer water supply (in particular, the concentration of arsenic and nitrates) as a result of the agricultural activities there.

Knowledge of this communities without clean water can help the authorities to redirect sanitation efforts to clean up the community's water supply. Identifying the industries can also help inform lawmakers and the public that agricultural activities (Nitrates usage) and Arsenic requires more regulations as they are a huge contributor to the state of the water system.

With detailed research, we found that agriculture strongly affects the nitrates levels and Arsenic levels a result from the BP Oil Spill in 2010.

In addition, Earnings and education attainment is higher for those with better quality of life and this reflects the importance of public health in the society.

Lastly, droughts are largely affected by live stock water usage and domestic water usage comparing with other features.

3 Data manipulation and modelling methods

3.1 Missing Value Imputation

- K-Nearest Neighbours (KNN): A non-parametric method whose decision is made based on the k closest samples in the training set. We impute missing values using the their nearest k numbers.
- Low Rank Matrix Completion: A missing entries filling method assuming low rank of the matrix

3.2 Feature Engineering

- Genetic Programming: It is a style of computer programming in which algorithms are written in terms of types to-be-specified-later that are then instantiated when needed for specific types provided as parameters.

3.3 Feature Selection

- Feature Importance and SHAP: It is a unified framework for interpreting predictions which has been introduced in [5]. We utilize it for feature importance measures.

3.4 Hyperparameters Tuning

- Bayesian Optimization: We use Bayesian optimization for hyper-parameters tuning. Bayesian optimization introduces a distribution over functions and uses an acquisition function to query points for finding the optimal.

3.5 Machine Learning Models

- Linear Regression: It is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables.
- SVM: It is a machine learning technique that constructs a hyper-plane or set of hyper-planes in a high- or infinite-dimensional space, which can be used for classification, regression
- Random Forest: It is an ensemble learning method for classification, regression.
- Extra Trees: It is a tree method that drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random.

4 Data Analysis

4.1 Identifying toxic chemicals

We first explored the data by visualizing health data from `http://www.countyhealthrankings.org`. The reason why we use this data is because it directly represents public health information in each county. The two columns we focused on here were ranking value of `length of life` and ranking value of `quality of life` (Figure 1). Quality of life and length of life is highly correlated (80%), hence we plotted based on the quality of life.

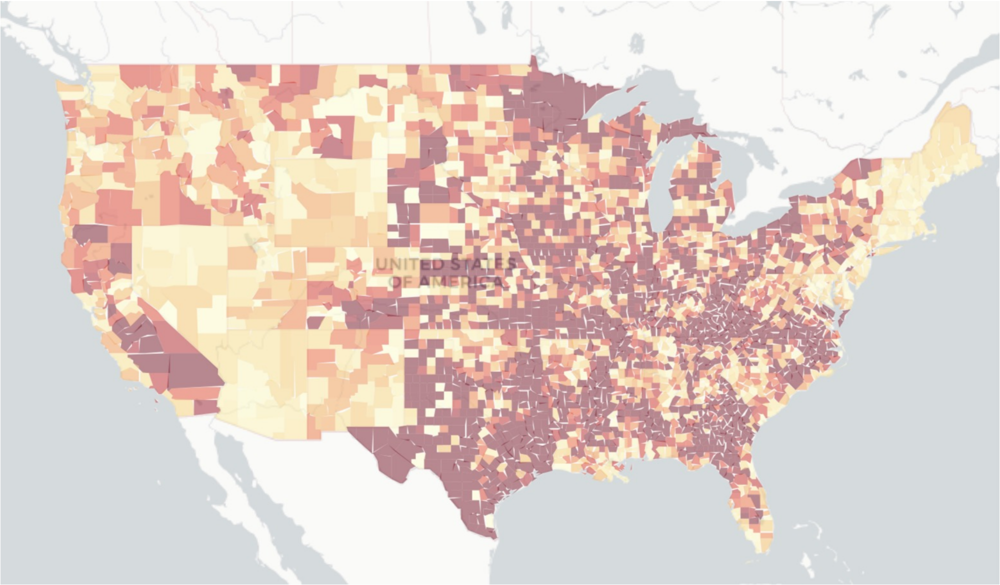


Figure 1: Heat map of quality of life

From the heat map we can see that the lighter areas have a lower score due to the better quality of life. This gives the initial deduction that water improvement efforts should be focused on at the southern and mid-western states like Texas, Georgia, Virginia, Kentucky, and Missouri. On the contrary, the states with the best quality of life are District of Columbia, Delaware, Hawaii, Rhode Island, and Connecticut.

First, we analyzed the `chemicals` data set using a regression model with the health data. We used k-NN imputation and low rank model to infer missing values it takes into context the correlation structure of the chemical data (Figure 3).

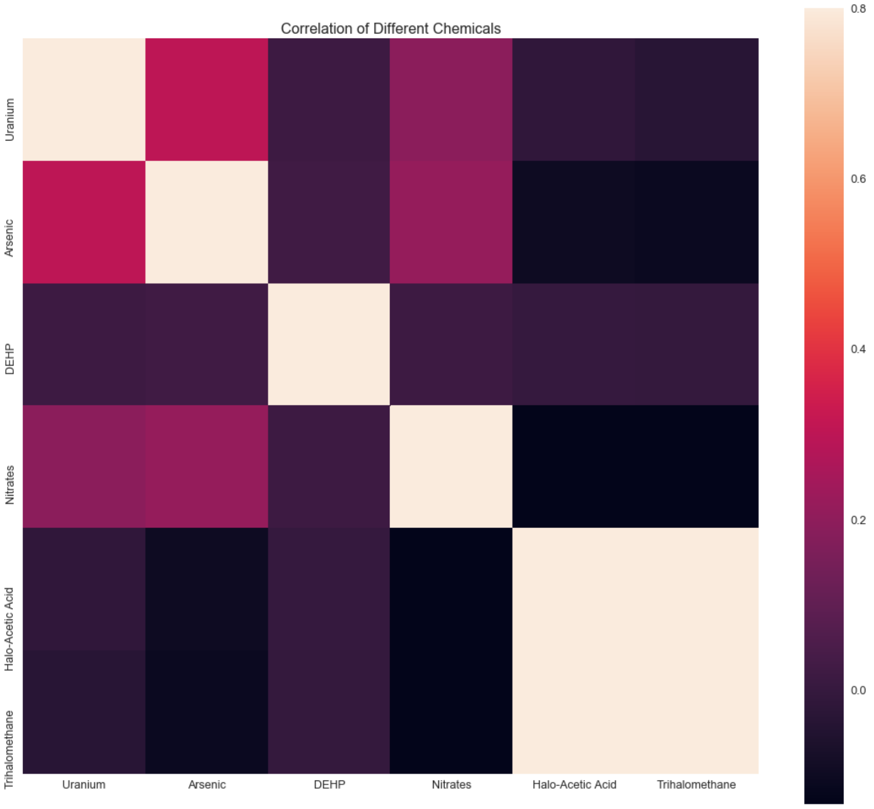


Figure 2: Correlation heat map of `chemicals`

From the correlation heat map we can see that Halo-Acetic Acid and Trihalomethane have relatively high correlation. Other chemicals seem to be less correlated with each other.

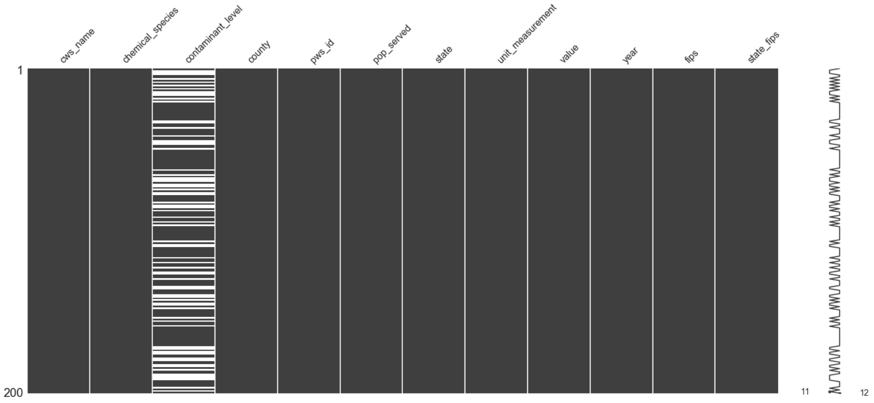


Figure 3: Missing values in `chemicals` dataset

We subsequently performed linear regression to revealed that Nitrates and Arsenic were the most significant variables in the water that correlates with health issues (Table 1) due to their most negative coefficients in the regression after data standardization. We observed that Nitrates is the most significant variable with smallest P-value. So in the later part we mainly just tried to use other variables in industry tables to predict the Nitrates level.

For arsenic, we further investigated the chemicals and found out that the recent BP oil spill in 2010 explains the increase levels of toxic arsenic in the open which created long-term threat to the marine ecosystem. [3].

We proceeded to investigate the arsenic levels in the surrounding states of the BP Oil spill in 2010 and plotted a factor plot that confirms the impact of the oil spill on the Arsenic levels (Figure 4). To try to combat nitrates concentration, we analyzed the top companies that oversee the community's water supply and picked out the ones with high concentration levels of nitrate (Figure 5). This list provides a starting point for authorities to approach in bringing the toxic nitrate concentration down.

variable	coeff	$P > \|t\|$
x1	-0.1209	0.612
Arsenic	-0.7292	0.163
x3	-0.1866	0.716
nitrate	-1.6903	0.036
x4	-0.2097	0.117
x5	0.0140	0.884

(a) length of life

variable	coeff	$P > \|t\|$
x1	-0.1222	0.608
Arsenic	-0.7368	0.159
x3	-0.1866	0.714
nitrate	-1.7079	0.034
x5	-0.2119	0.113
x6	0.0141	0.883

(b) quality of life

Table 1: Linear regression of Chemicals in community water system

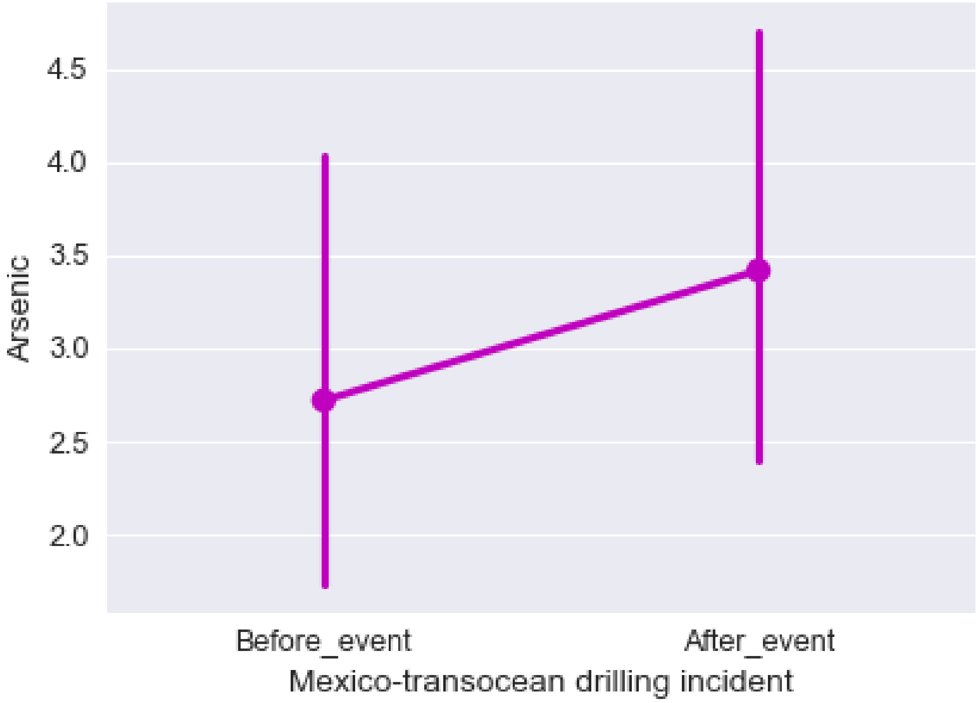
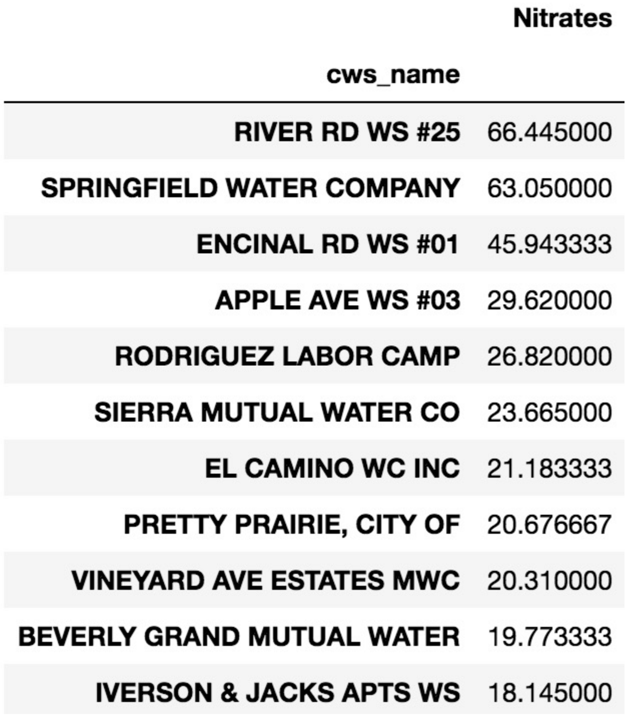


Figure 4: Arsenic levels in Louisiana community water system before and after the 2010 BP Oil Spill.



	Nitrates
cws_name	
RIVER RD WS #25	66.445000
SPRINGFIELD WATER COMPANY	63.050000
ENCINAL RD WS #01	45.943333
APPLE AVE WS #03	29.620000
RODRIGUEZ LABOR CAMP	26.820000
SIERRA MUTUAL WATER CO	23.665000
EL CAMINO WC INC	21.183333
PRETTY PRAIRIE, CITY OF	20.676667
VINEYARD AVE ESTATES MWC	20.310000
BEVERLY GRAND MUTUAL WATER	19.773333
IVERSON & JACKS APTS WS	18.145000

Figure 5: Nitrate levels of worst community water systems.

4.2 Toxic Chemicals impacted by industry

To further understand the reasons for the toxic chemicals, we suspected that certain counties had higher levels of toxic chemicals due to industrial activities. Furthermore, we wanted to know which community's water source will be threatened in the future with high levels of nitrates. So in this report, we used the industry ratio to predict the next year's nitrates level, which is most important chemical in water to judge the public health.

To evaluate which industry has the most influence on Nitrates. We mainly used the feature importance in random forest and SHAP method. Figure 6 and 16 reveal most prevalent industry is agriculture, and it is consistent with Figure 1 as the top agricultural states are Texas, Nebraska, Illinois, Minnesota, Kansas, Indiana, Wisconsin and North Carolina [6]. Furthermore, a big source of nitrates in water sources come from agriculture fertilizer [4].

As seen from Figure 9, the red represents its contribution to increasing the nitrates level (undesirable) and the blue represents the opposite. This plot gives us clarity on the agriculture industry's negative effects on the water quality by increasing the level of nitrates. This suggests authorities to relook at agricultural practices.

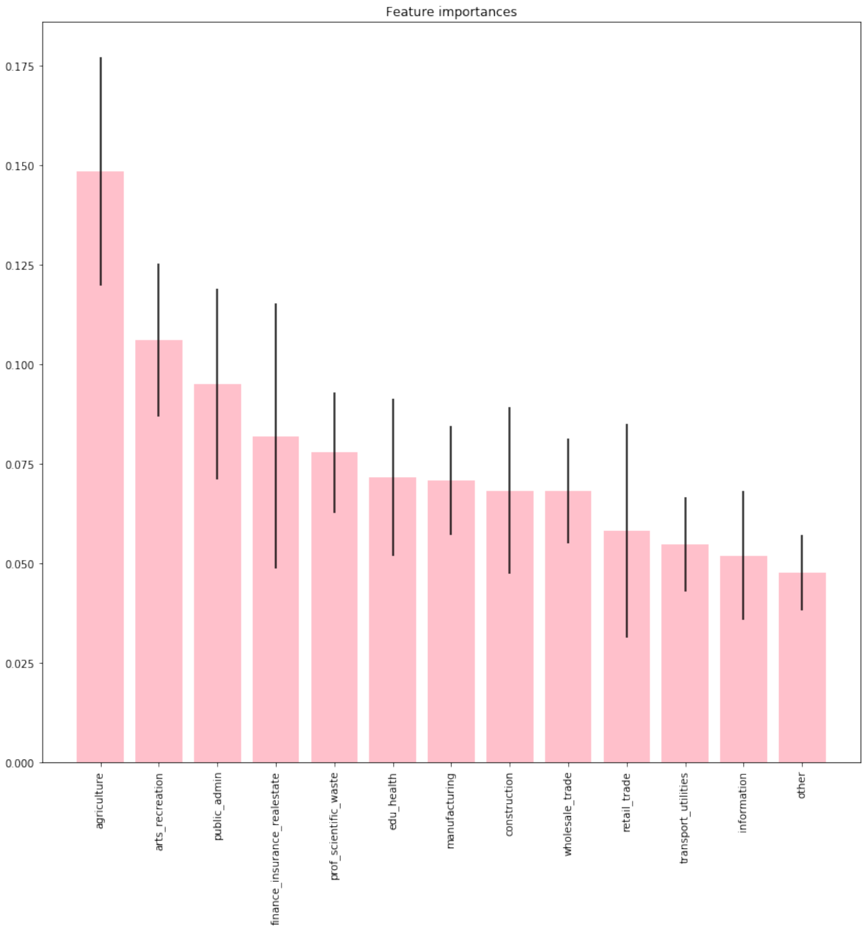


Figure 6: Feature Importance plot

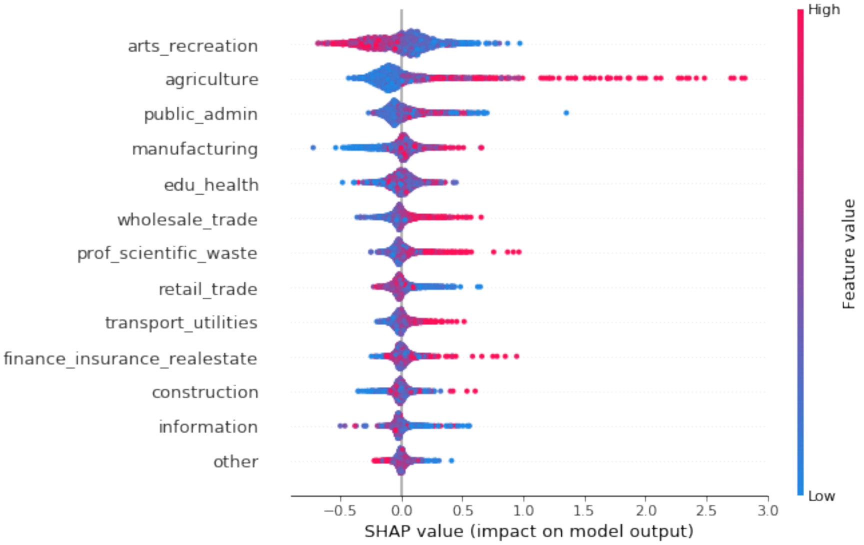


Figure 7: SHAP plot of its impact on toxic chemicals

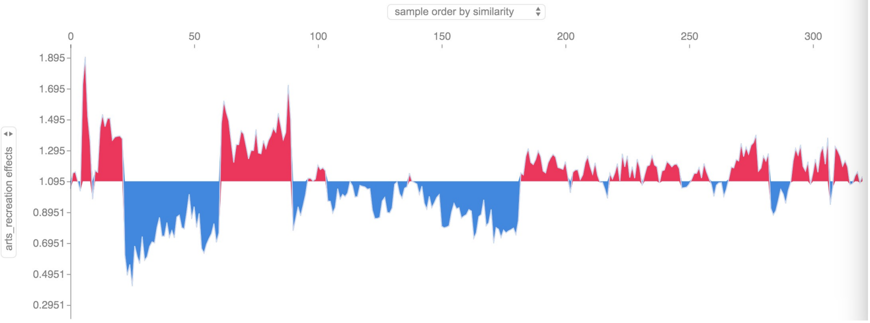


Figure 8: SHAP plot of recreation industry.

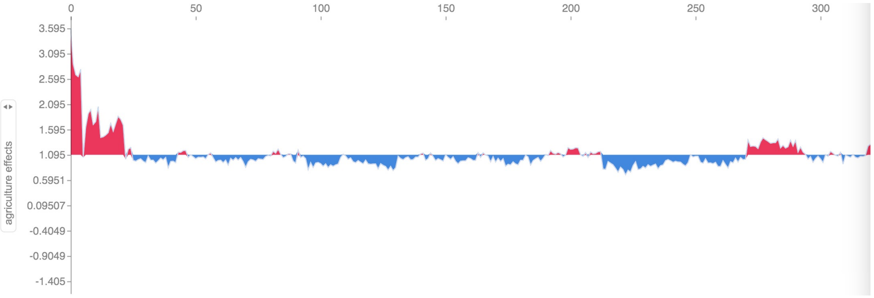


Figure 9: SHAP plot of agriculture industry.

4.3 Predicting the danger areas

The next natural question to ask is: where's the next location that could be in danger? This is important because it takes time for preventing this incident, and if necessary, to set up the infrastructure to control and respond to the any possible situations.

We divided the datasets into training data and test data. Then we used the Bayesian optimization [2] to find the best parameters of each model by using cross validation. Table 2 shows 3 algorithms, support vector machine, random forest and extra trees.

Model	Test MSE	Train MSE	Best parameters
Extra Trees	0.548	-0.534	$n_{estimator}$ = 200 $max_{features}$ = 1, max_{depth}=30
Random Forest	0.612	-0.612	$n_{estimator}$ = 235, $max_{features}$ = 0.95, max_{depth} = 17
SVM	0.826	-0.766	$C = 10, \gamma = 10$
LightGBM	0.435	0.0.475	bagging fraction = 0.9, feature fraction = 0.8, lambda l2 = 0.03, max_{depth} = 10, learning rate = 0.05,

Table 2: Machine learning models with bayesian-optimized hyperparameters. Extra Trees gave us the best results.

LightGBM was the best predictor for us and we used the model to forecast which counties will be affected. Table 3 reveals the counties that will be affected by high nitrates concentrations and where the authorities should prioritise.

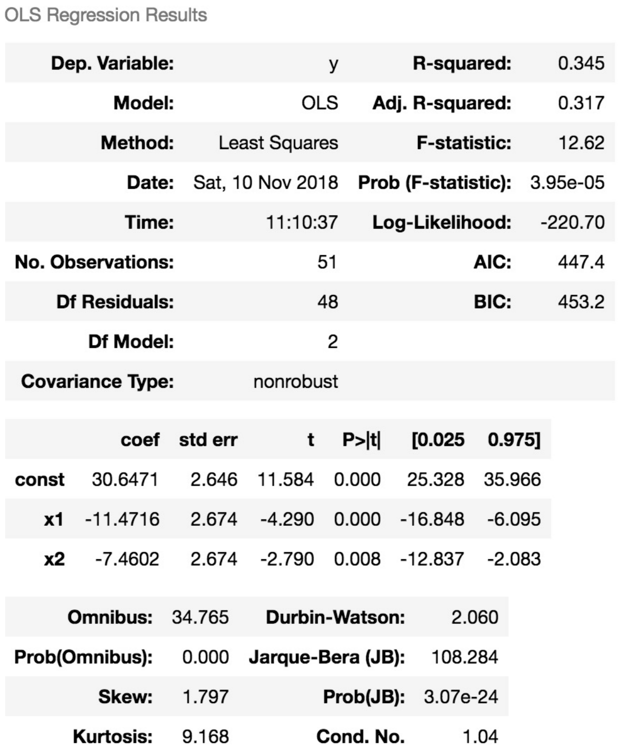
county
Kern
Madera
Monterey
Tulare
Ventura
Frederick
Howard
Nassau
Marathon

Table 3: Machine learning models with bayesian-optimized hyperparameters. Extra Trees gave us the best results.

4.4 Labor force and health

While agriculture is constrained to the weather and the land, labor and resource is also a determining factor in deciding where to set up farms and food processing plants. Our team hypothesizes that availability of cheap, lower-skilled worker reinforces the accumulation of agricultural activities in some states and the high agricultural activities result in the pollution of the water system.

Using the `health`, `earnings` and `education_attainment` dataset, we are able to evaluate that having a a poor quality of life is always a results of lower total median income (Figure 11) which is logically because of the lack of money contributes to poorer health. Furthermore, a lower quality of living is also a results of fewer number of people pursuing higher education. In Figure 12, we see those with higher education attainment and median income (further away from the origin) will have a better quality of life (red)



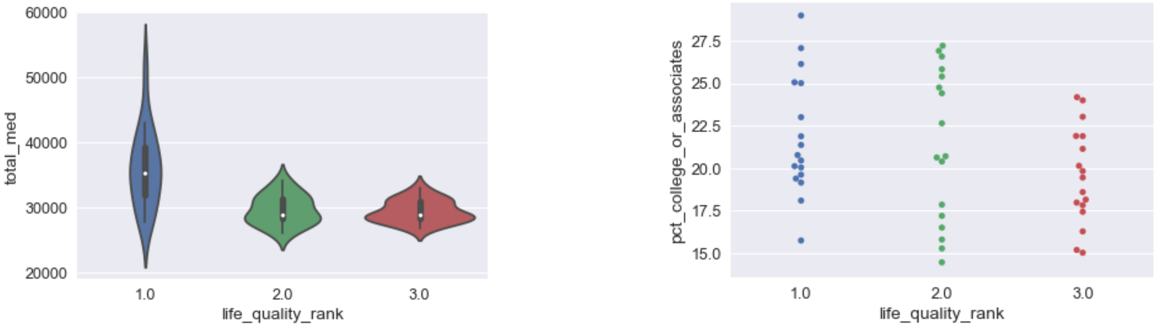
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.345
Model:	OLS	**Adj. R-squared:**	0.317
Method:	Least Squares	**F-statistic:**	12.62
Date:	Sat, 10 Nov 2018	**Prob (F-statistic):**	3.95e-05
Time:	11:10:37	**Log-Likelihood:**	-220.70
No. Observations:	51	**AIC:**	447.4
Df Residuals:	48	**BIC:**	453.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	30.6471	2.646	11.584	0.000	25.328	35.966
x1	-11.4716	2.674	-4.290	0.000	-16.848	-6.095
x2	-7.4602	2.674	-2.790	0.008	-12.837	-2.083

Omnibus:	34.765	**Durbin-Watson:**	2.060
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	108.284
Skew:	1.797	**Prob(JB):**	3.07e-24
Kurtosis:	9.168	**Cond. No.**	1.04

Figure 10: Regression results



(a) Total Median Income

(b) % of college or higher education

Figure 11: factors against Quality of life (QoL)

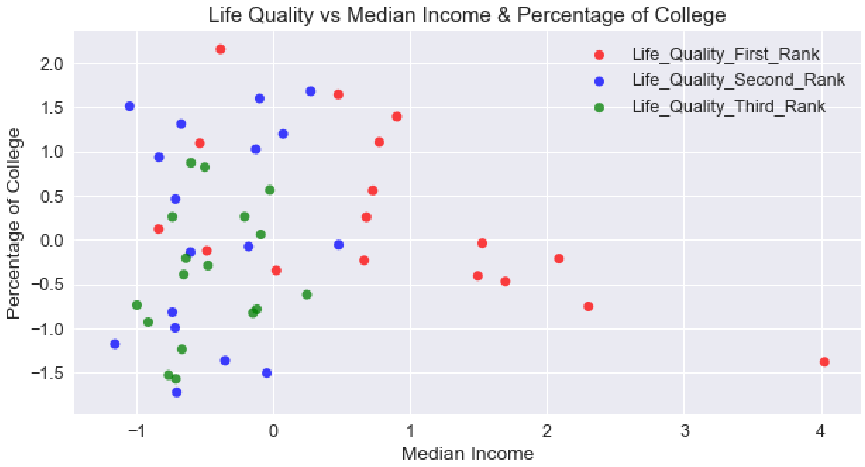
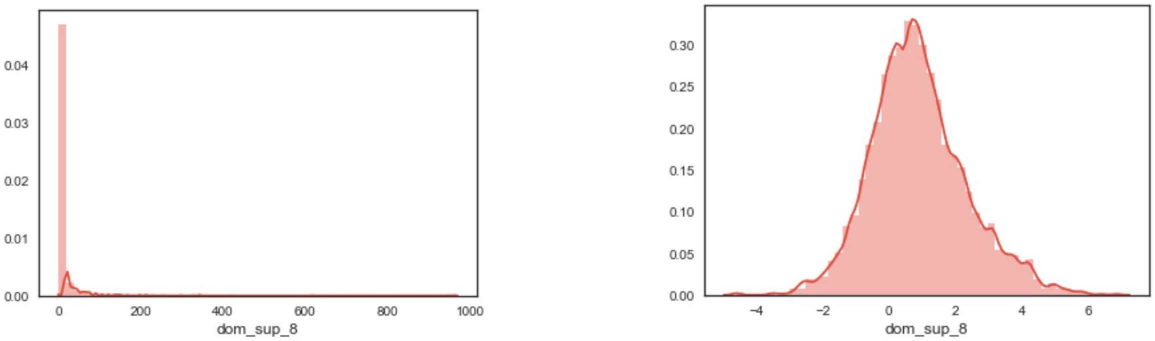


Figure 12: Nitrate levels of community water systems.

4.5 Drought hotspots

However, natural disasters such as droughts also threaten our water supplies. We analyzed the droughts data set together with the water usage, earnings and education_attainment which may potentially affect droughts. We applied natural logarithm transformation on variables which are positively skewed and used transformed variables in the following analysis(total_withdrawal, dom_sup_8 and total_med). Figure 13 shows an example of the original data distribution and the data distribution after the transformation. We could clearly see that the transformed distribution is more normal which is beneficial for the following regression analysis.



(a) raw data

(b) logarithmic transformation

Figure 13: Dom_sup_8.

Here is the heat map of droughts:

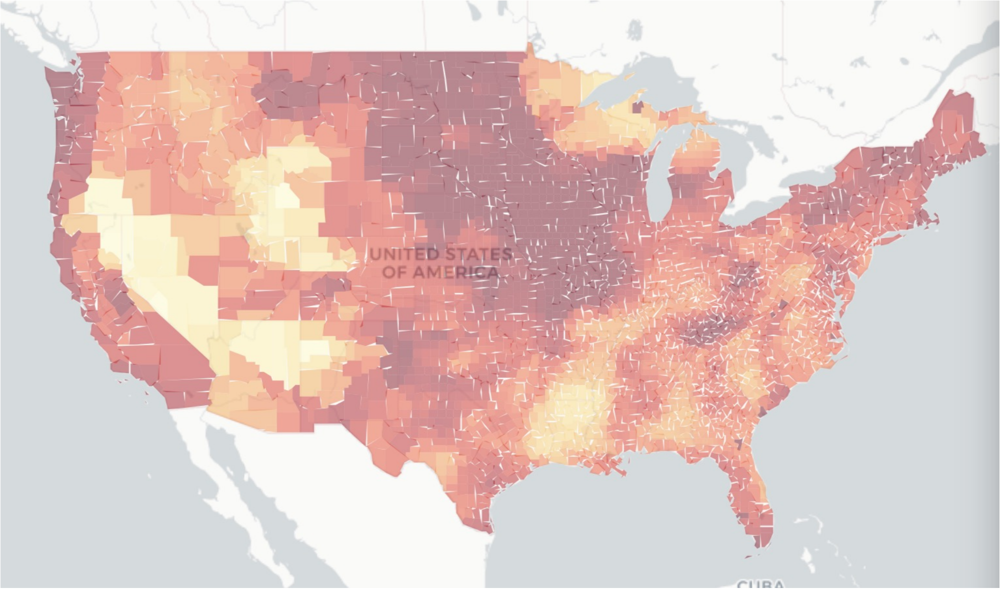


Figure 14: Heat map of droughts

From the heat map of droughts (light color suggests more severe degree of droughts), we can see that counties in Arizona, Nevada, and Utah suffered from droughts. The reason that these states have significant droughts is because they have large area of deserts and fewer vegetation comparing with others.

Then, we did feature engineering towards the variables and add new features:

irrigation_ratio	Irrigation, acres irrigated, divided by total withdrawals
livestock_ratio	Livestock, total withdrawals, divided by total withdrawals
aqua_ratio	Aquaculture, total withdrawals, divided by total withdrawals
mining_ratio	Mining, total withdrawals, divided by total withdrawals
industrial_ratio	Industrial, self-supplied total withdrawals, divided by total withdrawals
domestic_ratio	Domestic, total use (withdrawals + deliveries), divided by total withdrawals
thermo_ratio	Thermoelectric recirculation, total withdrawals, divided by total withdrawals
public_ratio	Public Supply, total withdrawals, divided by total withdrawals

From the missing values bar chart and nullity correlation we can see that there are missing values in the `water usage` data. The bar chart shows the columns with missing values, and the nullity correlation shows when one type of data missing, which other types will be likely to be missing too. We drop those columns accordingly.

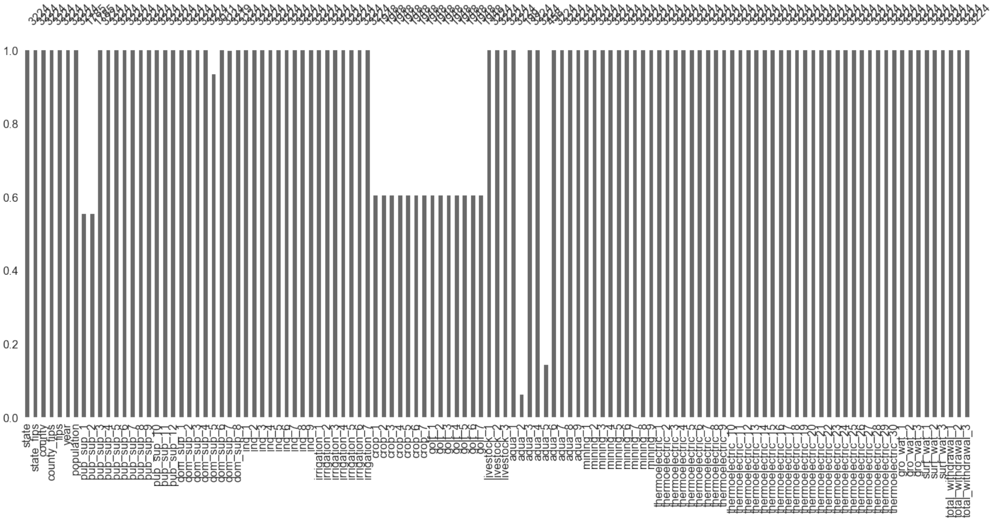


Figure 15: Bar chart of the missing values in droughts

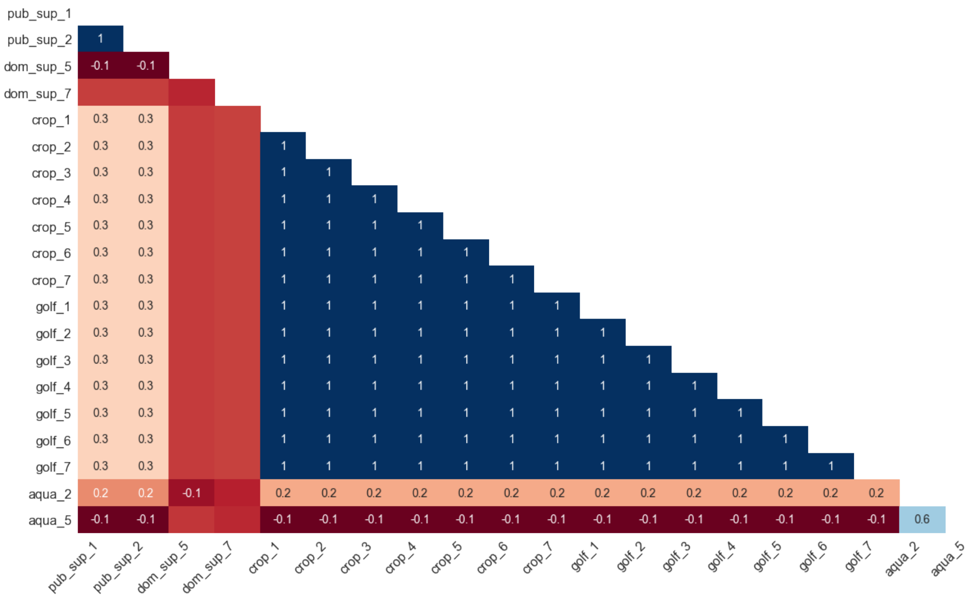


Figure 16: Nullity correlation of the missing values in droughts

We first apply SHAP to analyze the importance of the features on droughts. From Figure 17, we could see that livestock, domestic supply and irrigation have the most important impact on the droughts. Large domestic public supply and irritation may cause the area to be droughty while developing the livestock will help reduce the situation of the drought. We hypothesis that this is because large usage of water by domestic public supply and irrigation will make the droughts severe while livestock could help improve the environment to avoid the drought climate.

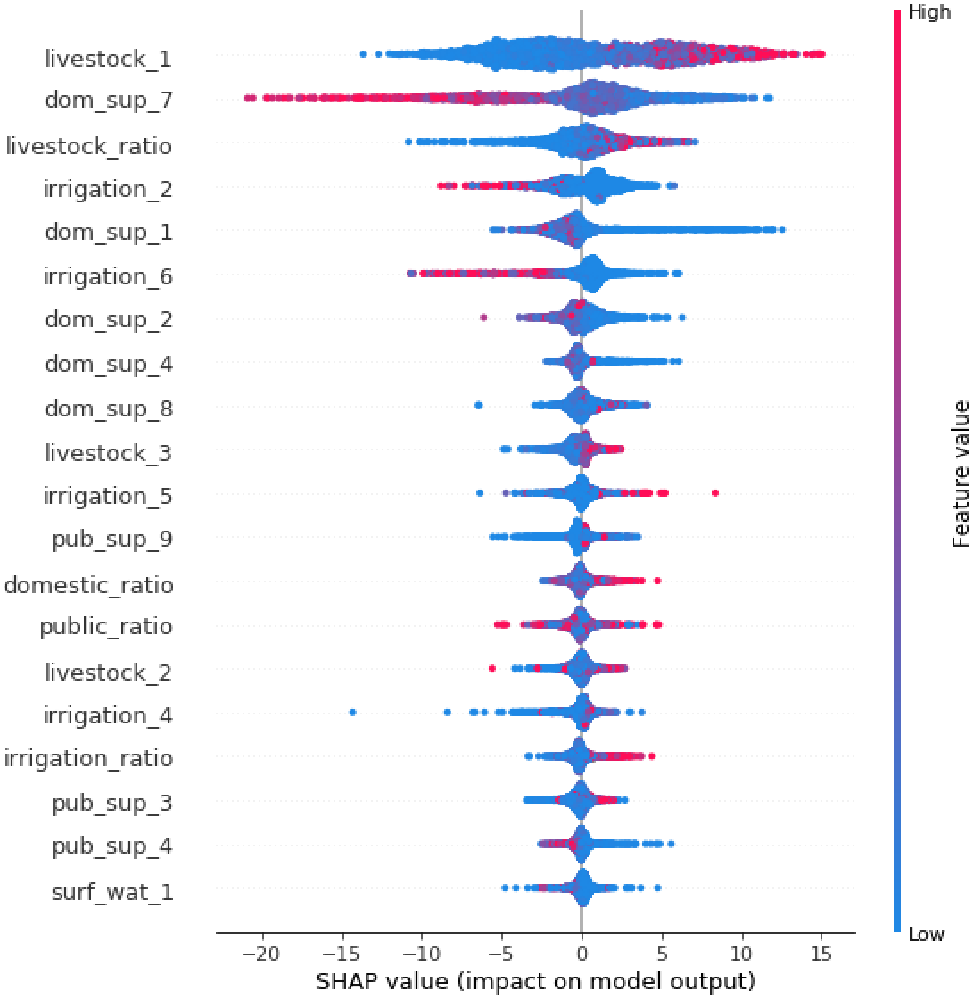


Figure 17: The feature importance by SHAP

5 Conclusions

From our analysis towards the chemicals, we found that Nitrates and Arsenic can severely damage water quality and therefore affect the quality of life for us. The reason is that agricultural activities will suggest high usage of Nitrates, and oil production and accident (Gulf of Mexico oil leakage) can increase the density of Arsenic. Earning and education level can also affect life quality of human being.

As for the droughts, we made a conclusion that live stock water usage and domestic water usage will affect the droughts.

References

[1] Millennium Development Goal drinking water target met, available at https://www.who.int/mediacentre/news/releases/2012/drinking_water_20120306/en/.

[2] Github- BayesOpt: A Bayesian optimization library, available at https://github.com/rmcantin/bayesopt.

[3] Oil Spills raise Arsenic levels in the ocean, available at https://www.imperial.ac.uk/news/91182/oil-spills-raise-arsenic-levels-ocean/

[4] Nitrate Contamination and the Sources of Nitrate Pollution , available at https://www.betalabservices.com/nitrate-test/

[5] S. Lundberg, S. Lee, A unified Approach to Interpreting Model Predictions, available at https://arxiv.org/pdf/1705.07874.pdf

[6] What US states produce the most food? , available at https://www.westernfarmpress.com/management/what-us-states-produce-most-food-ranking-1-50