Drew Adamski

Final Report:
NYC Water Quality Network Analysis

Problem Statement

The safety of our water supply can change rapidly, and accurate detection of poor water quality can be the difference between swift mitigation and full blown public health crisis. Many of the issues found upstream can be detected in the harbors where waterways ultimately end. But is it possible to formulate a network of water quality where measurements in one location give information about those at another? Could that network be leveraged to reduce monitoring costs?

By using New York City Harbor Water Quality data, I created a tool to help local governments shrink monitoring costs and predict poor water quality readings by reducing the amount of sampling needed to draw conclusions. Utilizing a variety of exploratory techniques like Network Analysis, Geostatistics, Frequentist Statistics, and Supervised and Unsupervised Machine Learning, I developed a model to solve this problem.

After reducing the dataset from 100 to 30 features, my tuned XGBoost Classificaiton model was able to achieve an average precision of 0.88 for three separate sites. This process can be repeated for other sites as well to enhance monitoring throughout the New York Harbor.

Data Wrangling

The raw dataset from NYC EPA contained 88,366 rows with 100 columns, so while it was usable, it needed serious size reduction. I started by looking at the sample locations. The dataset came with a map showing the different sampling locations used in 2016. There were many more sample sites in the dataset than listed on the map, so I only included data measured at sites present on the 2016 map. I also wanted to limit the data to this century, so I dropped all samples done before 2000.

With 100 parameters to consider, reducing dimensionality was a critical step if I wanted to pull stronger insights from the dataset, so I dropped any column that had less than 10,000 entries or any columns that measured the same thing.Outlier data points were looked at individually to determine whether the number was an incorrect entry or a legitimate measurement. The former were corrected or made null while the latter were kept in the dataset.

Null values were filled using forward filling with the previous site measurement being used. This would preserve the temporal differences in measurements that would be lost by imputing the overall or site specific mean.

The final shape of my dataset was 30,989 rows with 43 columns.

Exploratory Data Analysis

Different parameters will serve different purposes in my analysis. Some measurements have standard limits that must not be exceeded while the rest of the columns will be used as potential indicators of poor water quality. The following columns have set limits:

- Dissolved Oxygen = 3.0 mg/L minimum
- pH = 6.0 - 9.5
- Fecal Coliform = 2000 cells/100mL maximum
- Enterococcus = 104 cells/100mL maximum
- Ammonia = 2.0 mg/L maximum
- Ortho-phosphate = 0.20 mg/L maximum
- Kjeldahl Nitrogen = 10.0 mg/L maximum
- Chlorophyll a = 30 µg/L maximum

The limit that I am most interested in looking at is Chlorophyll A. Chlorophyll A signals high presence of algae which can mean blooms that kill all aquatic life in the area. It also is closely linked to other measurements and can act as a proxy measurement for high nutrient loads in the area.

I looked at correlations between every column and my target parameter, and I found some strong correlations which I then plotted (Figures 1 and 2). In order for something to qualify as a “strong” correlation, the absolute value of it’s correlation had to be greater than 0.30. I began by plotting Chlorophyll A against it’s two most highly correlated parameters: pH and Top Silica.

Chlorophyll A

- Top pH 0.3288
- Top Silica -0.307083

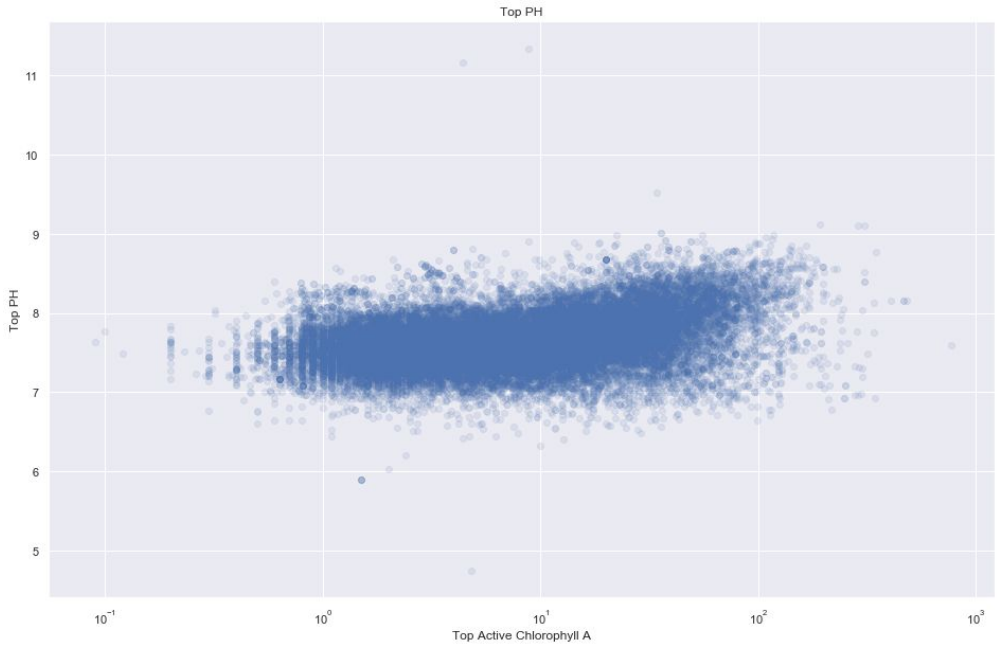

Figure 1: Scatter Plot of Chlorophyll A vs pH

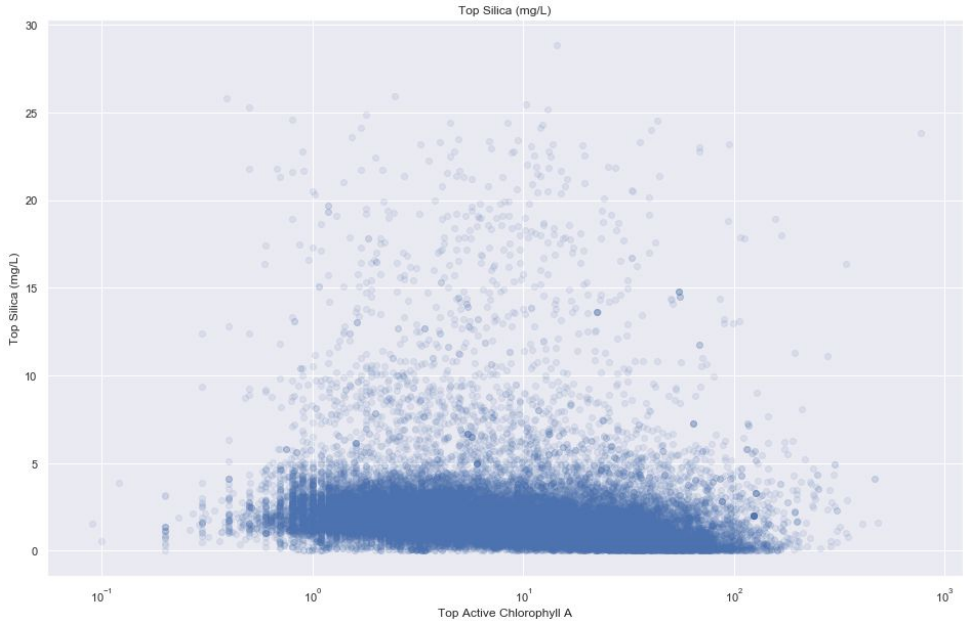

Figure 2: Scatter Plot of Chlorphyll A vs Silica

Both of these graphs appear to corroborate the correlation measurements. There may be a chemical component to chlorophyll A that increases the pH of the water, or it may be a single factor in the complex chemical composition of natural waters.

In-Depth Analysis

Now that I knew what information my data held and understood the relationships between different parameters, I needed to determine which sites I would build my model around. If I wanted to reduce the number of sampling locations, I needed to identify sites that had the best predictive ability. To do this, I used geography based scatter plots, graph analysis with NetworkX, and Inverse Distance Weighting plots.

Figure 3 plots each site based on their geographical position (latitude and longitude). The size of each dot is related to the probability of a site having an above limit chlorophyll reading. I took the frequency of failed tests and divided it by the total number of samples taken in the area.

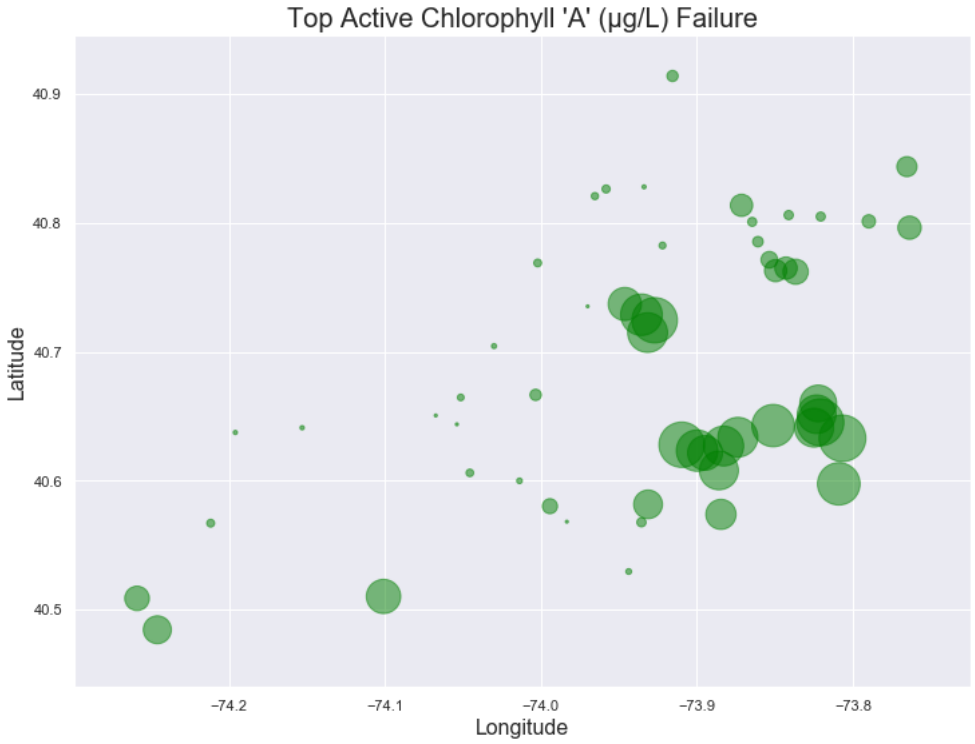

Figure 3: Geographic Scatter Plot with marker size representing probability of high Chlorophyll A reading

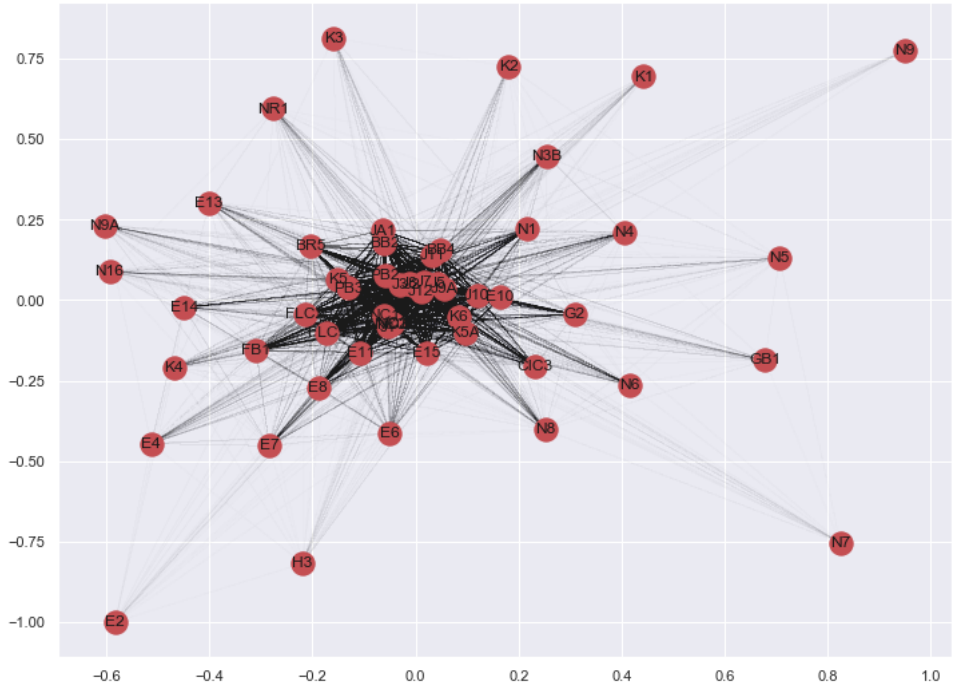

Figure 4: Multigraph representing overlap between high Chlorophyll A readings

Figure 4 depicts the relationships between the different sampling sites. Each node represents a site, and the edges connecting the nodes represent every time two sites have a chlorophyll measurement above the EPA limit in the same month as each other. It is a multigraph meaning that nodes can be connected multiple times, so darker lines represent sites with more overlap. Sites located in the center are connected to a greater diversity of locations while those on the fringe may only have a couple overlapping records with a handful of sites. I measured the degree centrality for each site which is a mathematical representation of how a sites connectivity to other nodes in the network.

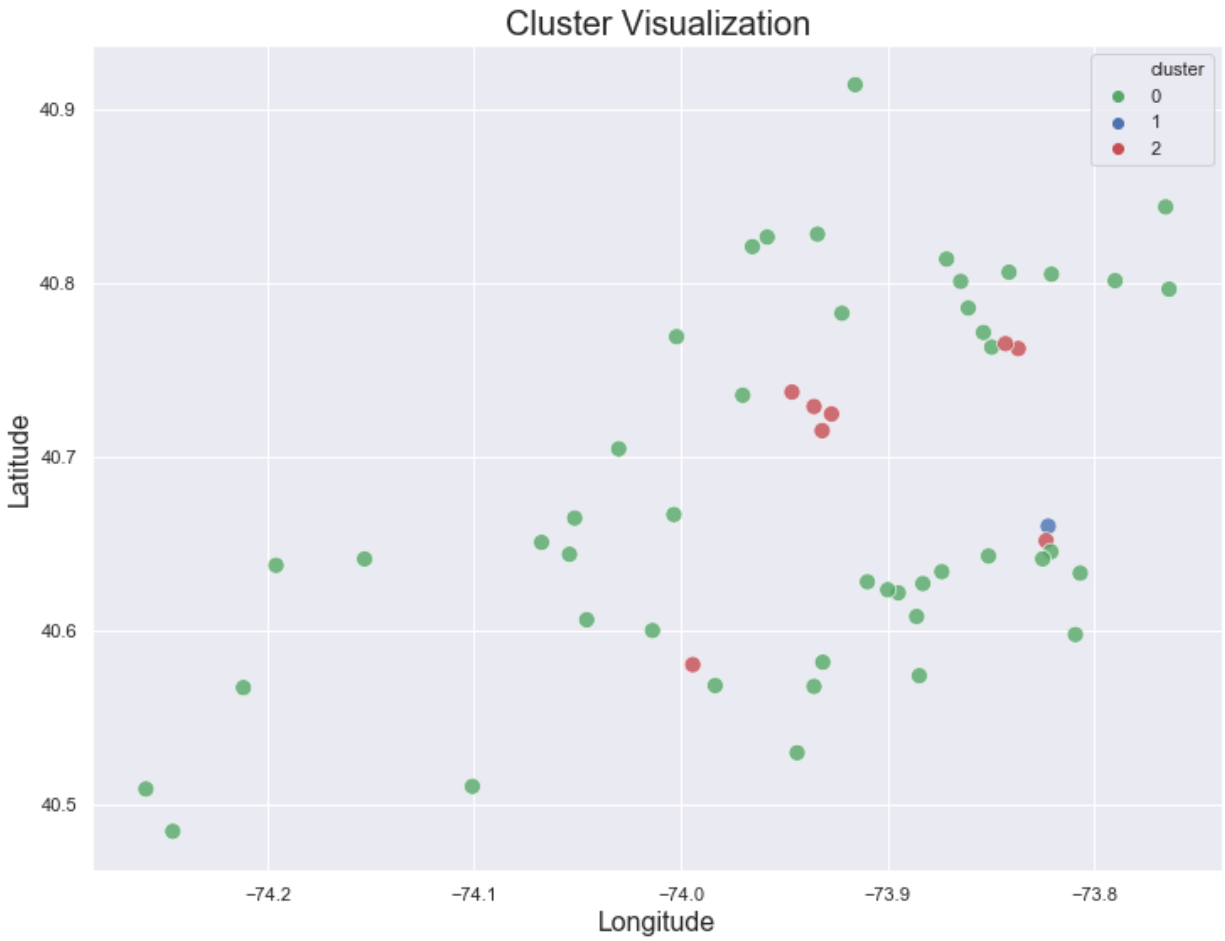

Figure 5: Plotting clusters geographically

Another method I used to determine sites of interest was an unsupervised K-Means clustering algorithm. I tested multiple numbers of clusters to identify the best option and plotted the distortion present for each model. Using the Elbow Method, I decided on 3 groups, however after plotting the resulting clusters (Figure 5), the clusters did not offer useful information as they were very uneven. Cluster 1 (blue) contains a single point, Cluster 2 (red) contains 8, and Cluster 0 (green) contains all other sites.I tried clustering again on points within Cluster 0, but the results did not add more information.

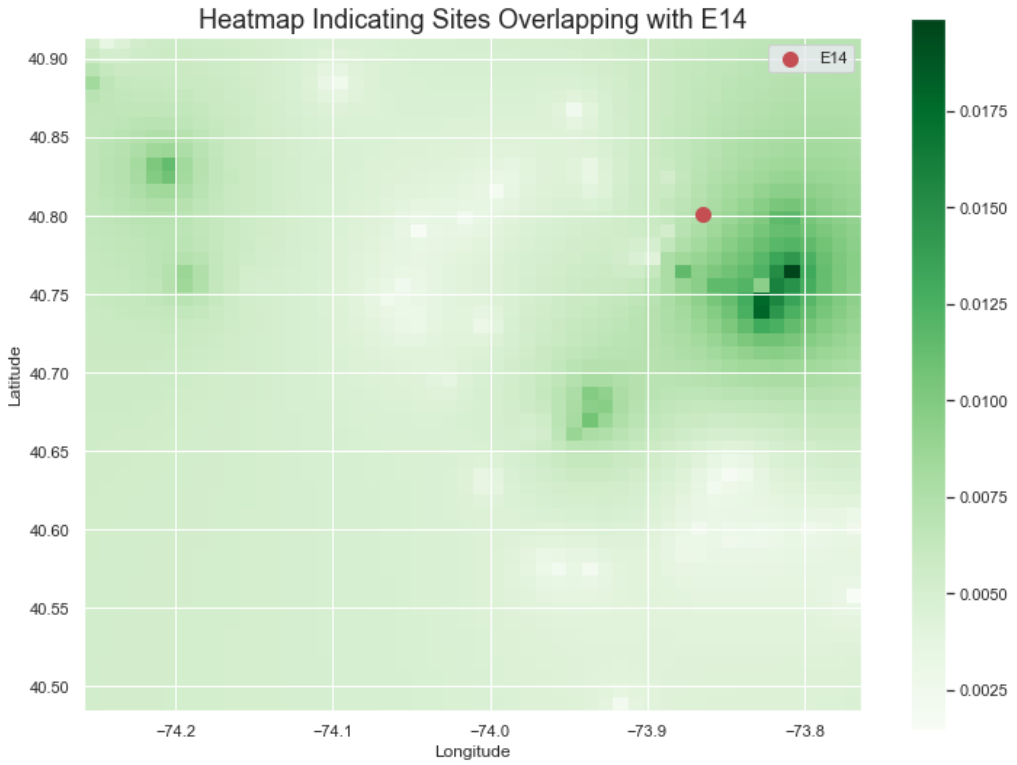

Figure 6: Inverse Distance Weighting Heatmap

Figure 6 is a heatmap showing the strength of overlap between different sites. The target site is marked with a red circle. The darkness of the map is determined through Inverse Distance Weighting. The probability of overlap between the target and another site is then put through an algorithm that considers the distance between the two sites. The map is then shaded according to these calculated values. Areas that are completely white did not have any overlap. This helped give an idea of the effect geography has on these relationships.

The combination of the scatter plot map, multigraph, and geographic heatmap helped me decide on 6 sites to focus my research on: BB2, E14, FLC1, J8, K6, and PB3. There was a wide variation in the ability to successfully model each site.

Model Selection

For each site, I tested 3 different machine learning classification models: Logistic Regression, Random Forest Classifier, and XGBoost. The metric I focused on when building my models was precision. I wanted my model to predict areas with high chlorophyll

Before building the models, however, I needed to do some feature selection. To do this, I performed Recursive Feature Elimination to determine the proper number of features to use. There was an even split between how many features to use for each site. Half only needed 15 features to make accurate predictions while the other half required all of the available columns. I

used an out-of-the-box Random Forest Classifier and Recursive Feature Cross Validation where the parameter with the lowest feature importance was dropped for each iteration until there were none left (Figure 7).

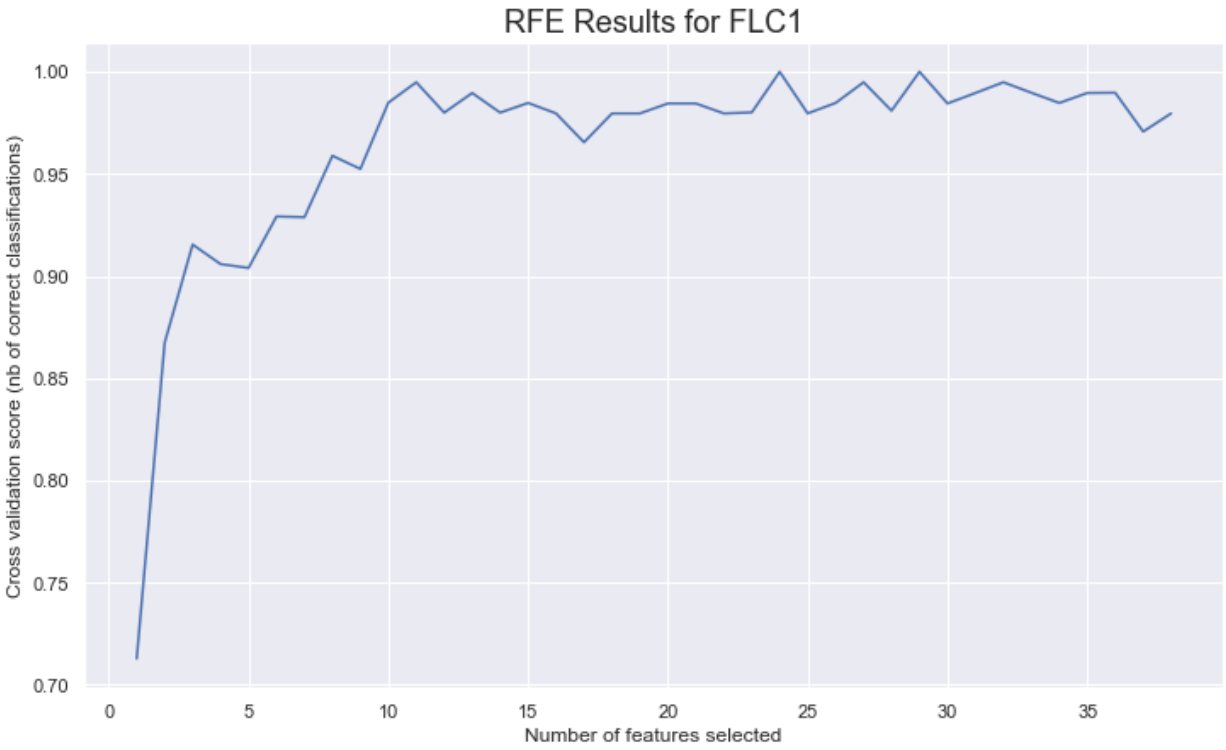

Figure 7: Recursive Feature Elimination Graph illustrating model effectiveness with different number of features

There was a wide variety in model performance. Some sites were able to be predicted with fairly good precision such as J8 while others were essentially a coin flip or worse (Figures 8 and 9).

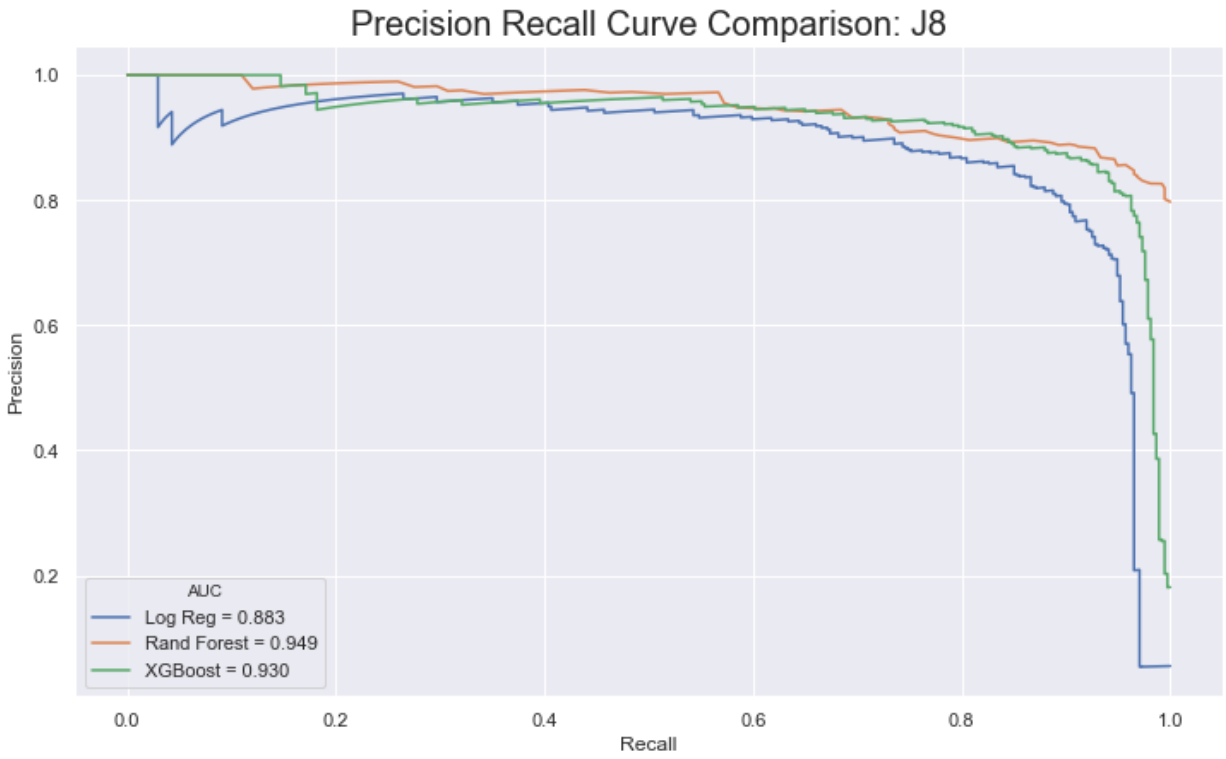

Figure 8: Precision Recall Curve for site with high predictability, J8

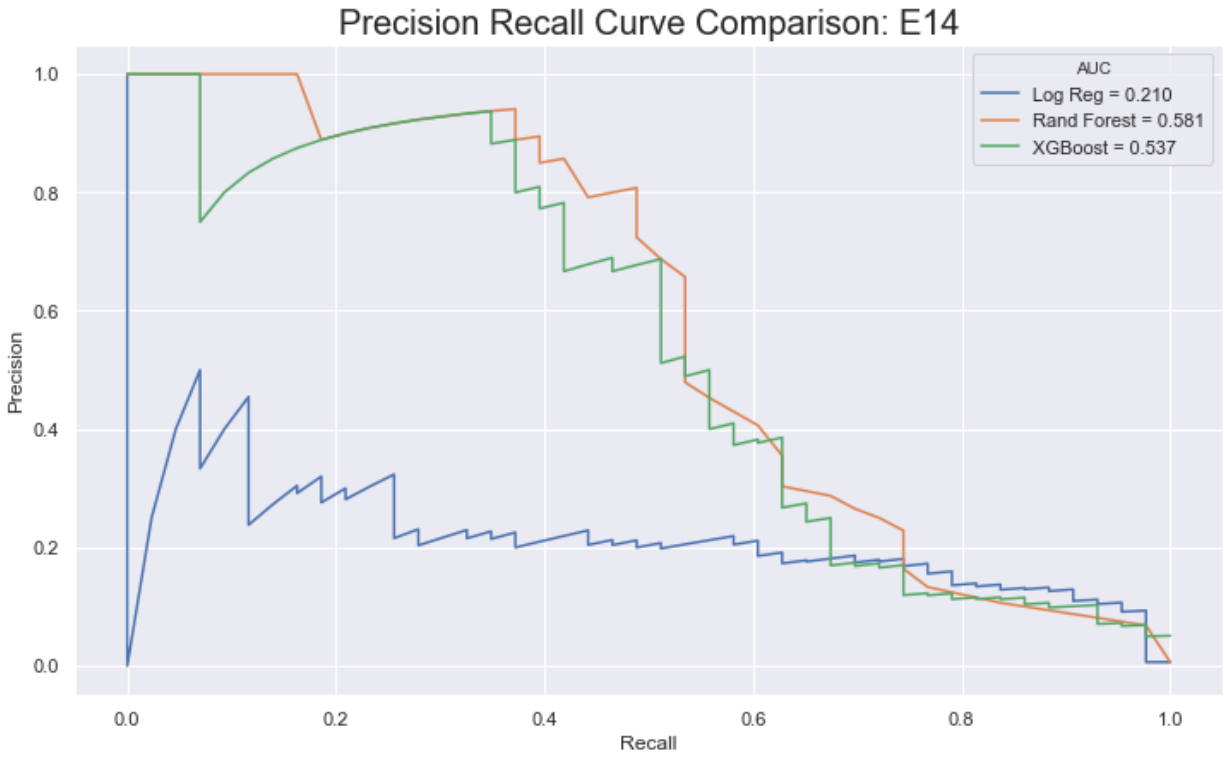

Figure 9: Precision Recall Curve for site with low predictability, E14

When it came to the models, most sites performed best with an XGBoost or Random Forest Classifier. Logistic Regression consistently performed the worst. I did hyperparameter tuning for the XGBoost model at site J8 and BB2 using Grid Search Cross Validation. I was then able to improve the models slightly more to a precision of 0.89 at J8 and a precision of 0.87 at BB2.

Takeaways

Out of the box, XGBoost is best model for chlorophyll. Consistently outperforming Random Forest Classifiers and Logistic Regression models, XGBoost gave the highest precision which is the most important consideration. False positives cost the regulatory agencies money to perform more testing which defeats the purpose of the modeling in the first place.

Ortho-Phosphorus matters Ortho-Phosphorus Failure was consistently the most or one of the most highly correlated features with high chlorophyll levels at different sites. Perhaps using this test as a preliminary measurement and then depending on whether the reading counts as a pass or fail, then it would be worth looking at other sites.

A lot of features are not important. The redundancies in measured parameters are unnecessary. Most are collinear so they do not add to the model, and there are also parameters that have no impact on the target variable and can be removed.

Sites have different abilities to be predicted. When building these models, there were certain sites that were not very predictable for one reason or another.

Future Research

This project gave me a lot to think about as far as ways to improve environmental monitoring. While my models were valuable to this specific project, it is difficult to say how useful they would be in another place like Wyoming or San Francisco. Different ecosystems have different inputs and outputs. However, the process that I followed could be replicated elsewhere. There are quite a few areas for further research that could be worthwhile delving into.

I would like to expand these models to include every single site. This would give an even more complete picture of the water system. Due to time constraints, I had to be selective with which sites I tested, but if it was not an issue, I could be even more thorough. I could create an ensemble model that takes a single measurement and makes predictions of every other site.

There are EPA standards for a variety of parameters such as bacteria, dissolved oxygen, phosphorus and many more. Going through this process for each of those could also prove valuable to improving poor water quality detection and reducing monitoring costs.

I looked at data from the past 20 years to capture weather abnormalities that may not repeat yearly but rather on a larger cycle like 5 or 10 years. Shrinking the range to 5 years may give more accurate models.

Developing a way to incorporate physical connectivity beyond latitude and longitude could help the model. This could be an entirely separate analysis, but by running experiments on the flow of water within this system, the flow of pollutants would also be better understood. This would in turn greatly improve the modeling.