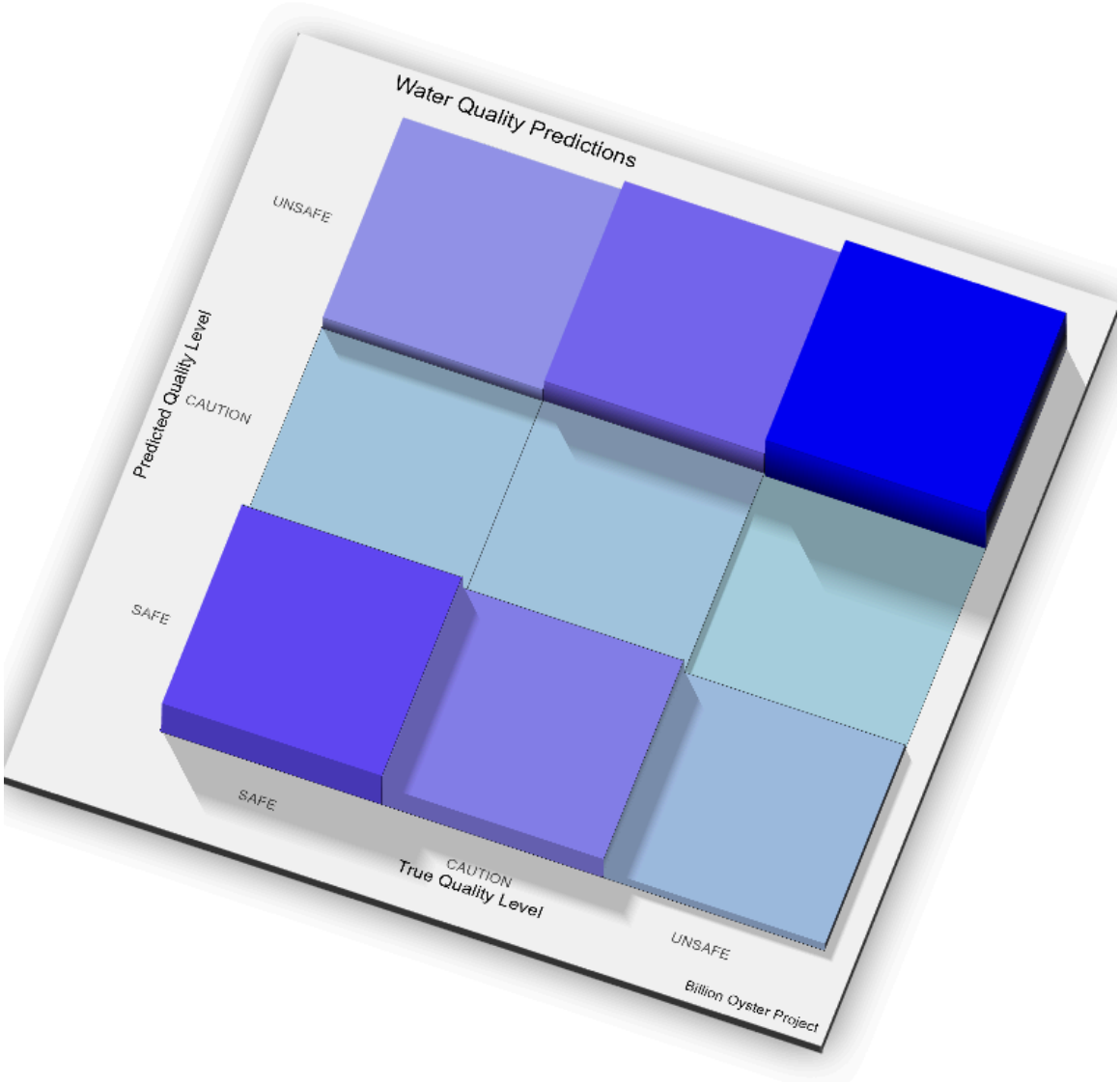
# Predicting Enterococci Levels in NYC Harbor

Art Steinmetz

2024-07-20



## Abstract

This document explores the relationship between testing location,weather, tides and water quality in the NYC Harbor. The data sources are the Billion Oyster Project (BOP), the Citizens' Water Quality Testing Program and the NOAA.

This note is an update of a previous exploration of the data where I showed that a linear regression model was a very poor explainer of bacteria levels in New York Harbor water. In this note I set a lower bar by classifying bacteria into just three classes, "safe," (<=35 colonies) "caution" (<= 104 colonies) and "unsafe" (> 104 colonies). I train a "random forest" machine learning model on a sub-sample of the data and then evaluate the model with a different test set of data.

In summary, this model works very well in fitting the training set but does much worse out of sample. The model does show good accuracy in predicting "safe" and "unsafe" water but very little accuracy in predicting bacteria levels in the "caution" range. The dominant predictor is the testing site, since several sites NEVER have "safe" water in the data set. No other variable stands out in significance.

This is not an academic-quality study. It is an exploration of the data. I am not a water quality expert or a professional statistician. Comments and criticism are welcome.


## Data

The main data source is the BOP water quality spreadsheet found here: BOP Water Quality Data[1] I also used the NOAA data site for tide, temperature and rainfall data.

## Feature Engineering

The BOP data includes time of last high tide. I thought I could get more granular by imputing the direction and strength of the tidal current at the time of the water sample. I used the NOAA tide data to find the previous slack tide time and level, then the next slack tide time and level.By determining where in the tide phase the sample was taken and the total change in water level for that phase, I impute the direction and strength of the tidal current when the sample was taken using this formula:

$$CurrentSpeed = HighLowRangeFt * sin\left(\pi * \frac{HoursSinceLastTide}{TideDurationHrs}\right)$$

So the further we are from a slack tide, high or low, the faster the current will be. The bigger the change in water level during a tidal phase, the stronger the current will be. Ebb tides are negative values, flood tides are positive. *CurrentSpeed* is an index so the units don't have a specific meaning like feet-per-second.

I get the tides from the closest NOAA tide station to each water sampling site. Where the location of the sampling site is not known, I default to the Battery tide station at the bottom of Manhattan.

[1] https://docs.google.com/spreadsheets/d/1813b2nagaxZ80xRfyMZNNKySZOitro5Nt7W4E9WNQDA/edit?gid=1924583806#gid=1924583806

This occurs when the name of the sampling site does not agree with any site name in the location meta data. **There are significant number of such cases.**

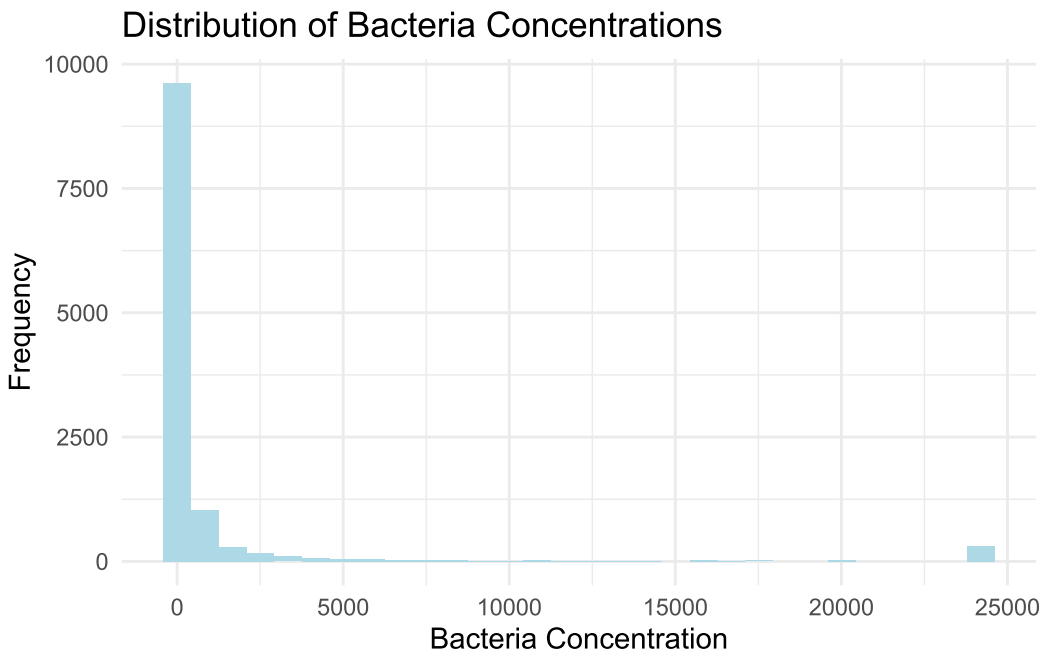
The city of New York uses 48-hour rainfall amounts in its safety criteria so that is what I use as the precipitation variable.

The BOP data does not include temperature. I used the NOAA Central Park temperature for each sample day as a data feature. This is a (not very good) proxy for the water temperature but also for seasonality. This allows seasonality to be a continuous variable. Otherwise, "month" would be a categorical variable.

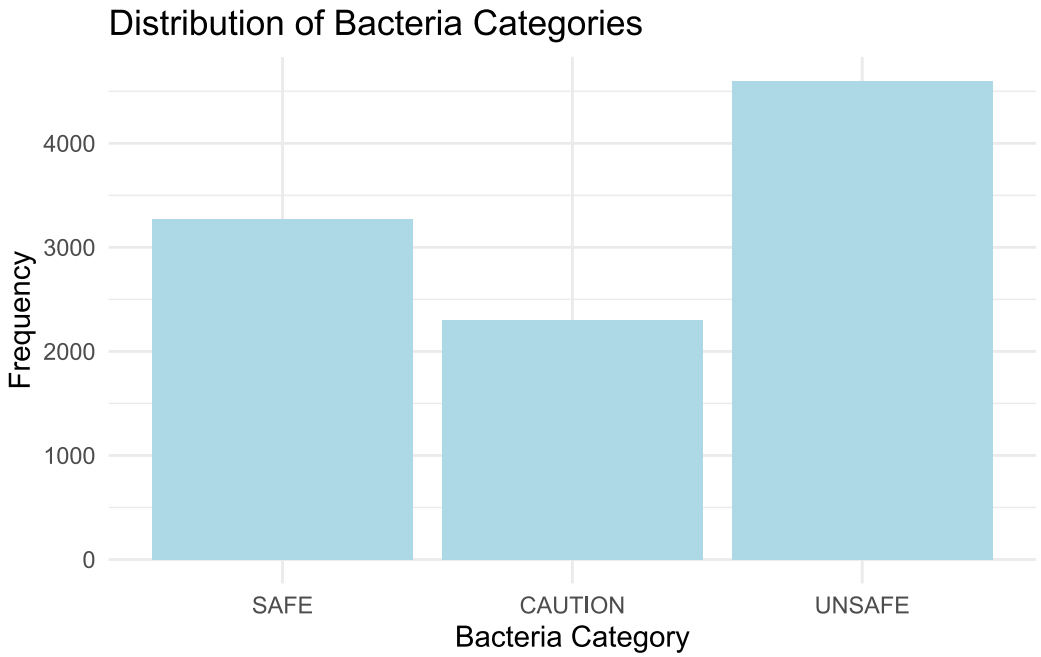
In the end I chose to the following features: `Site`, `TideHighLowRange`, `HoursSinceLastTide`, `CurrentSpeed`, `48-HourPrecip` and `Temperature`.

## Data Exploration

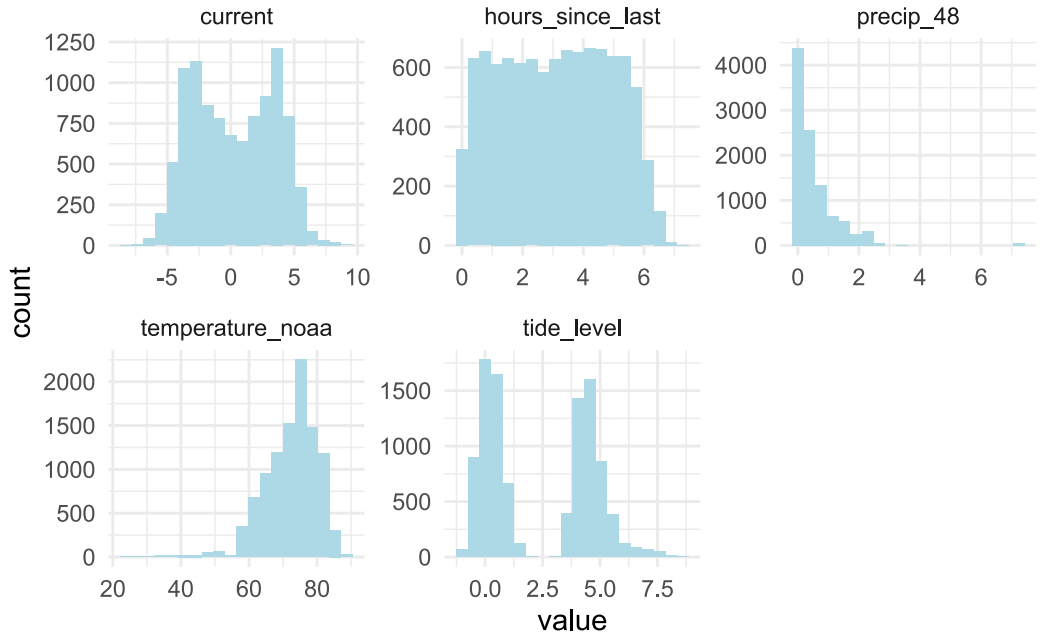
The bacteria levels are distributed in a lopsided way. The extreme high level is effectively infinity and conveys little information. Values above 5000 are only 5% of the observations and values below 500 are 82% of the observations.



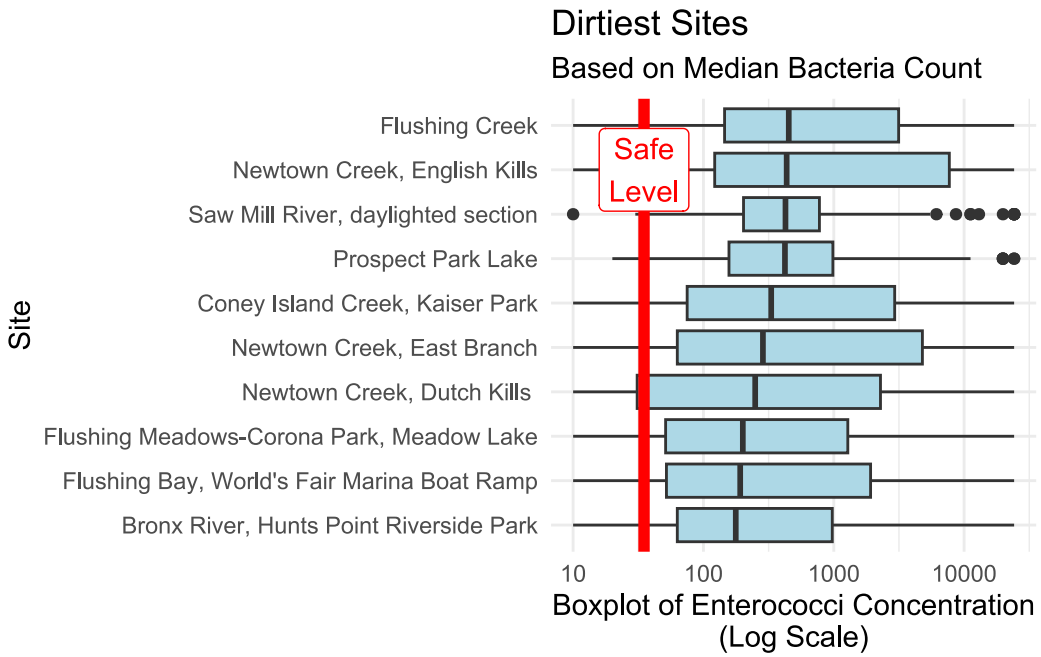

If we group the bacteria levels according to the official quality standards we get a better behaved distribution.

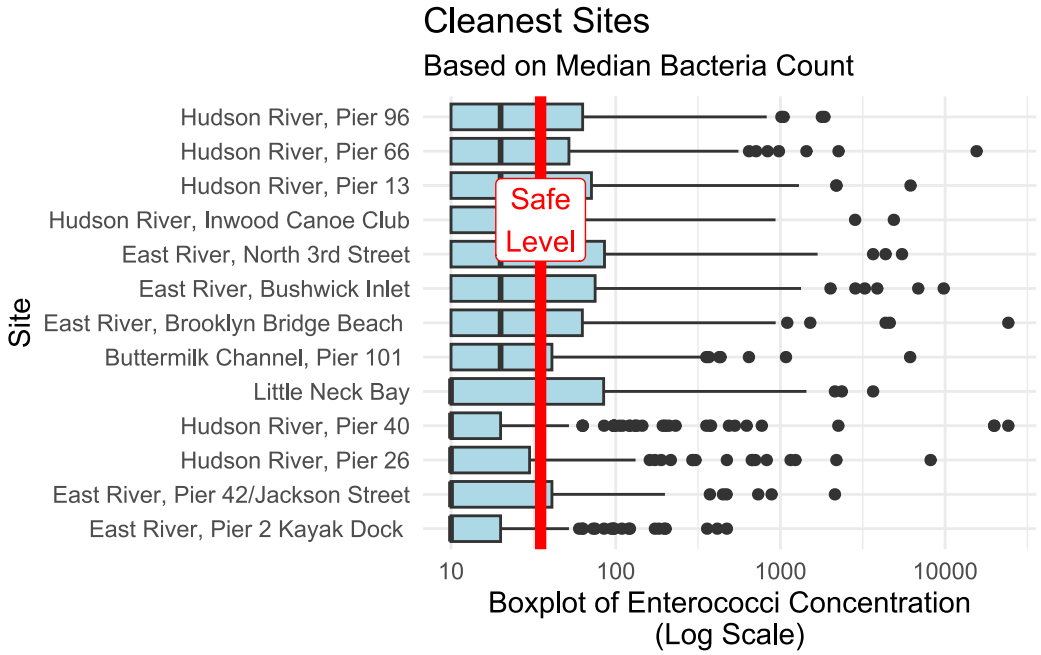



What are the distributions of all the variables? Note the tide level distribution are the levels at just the high and low tides.



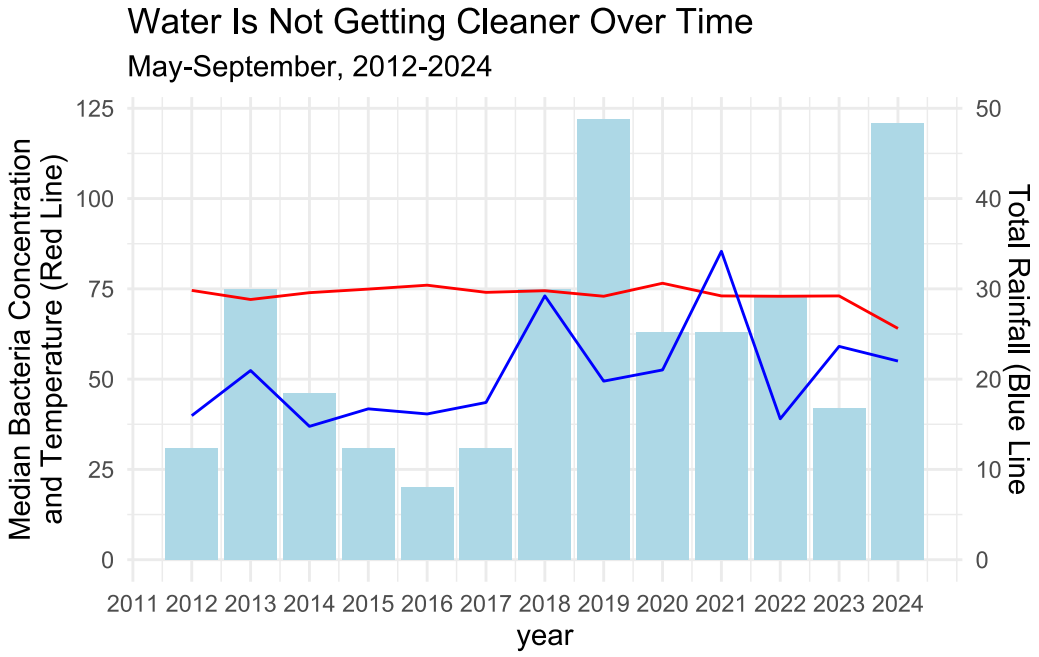

What are the cleanest and most contaminated sites?




What is obvious is that even the cleanest sites have a lot of variation in bacteria levels. This might give us some hope that environmental factors might be more important than location in predicting bacteria levels.

Now let's look at some trends over time. Sadly, the overall level of bacteria has not improved over time. Looking at temperature, there are no clear trends. There are a couple years where a lot of rainfall seems associated with more bacteria but other years contradict that.




## Modeling

We use a random forest algorithm to train a prediction model. This class of models works very well on imbalanced data like we have here. It can also handle data sets with many categorical inputs like `site` in this case. More on this technique can be found at https://en.wikipedia.org/wiki/Random_forest . To create the model we split the data randomly into a training set and a test set. 75% is used for training and the rest we hold out for testing. The sets are stratified so the same proportion of each bacteria category is in each set. The model is tuned using cross-validation on the training set and then evaluated on the test set.

## Results

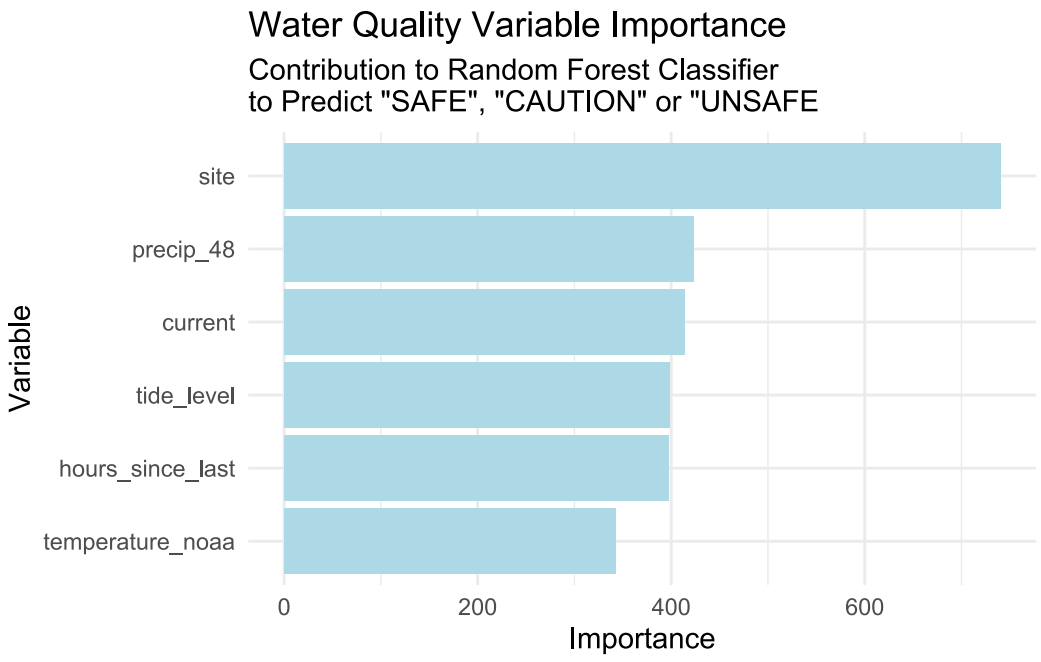
The simplest way to evaluate the model is to look at the confusion matrix. This is a table that shows the number of correct and incorrect predictions for each category. In a perfect model all the observations would lie on the diagonal and the off-diagonal counts would all be zero. The table below shows that out of 1150 "UNSAFE" observations, the model predicted 903, or 81%, correctly. This was the best result. On the other hand, in 7% of *all* the cases, the model predicted the water was "SAFE" when the actual was "UNSAFE" (176/2544). Additionally, The model is far better at predicting "SAFE" and "UNSAFE" than "CAUTION."

Truth Table

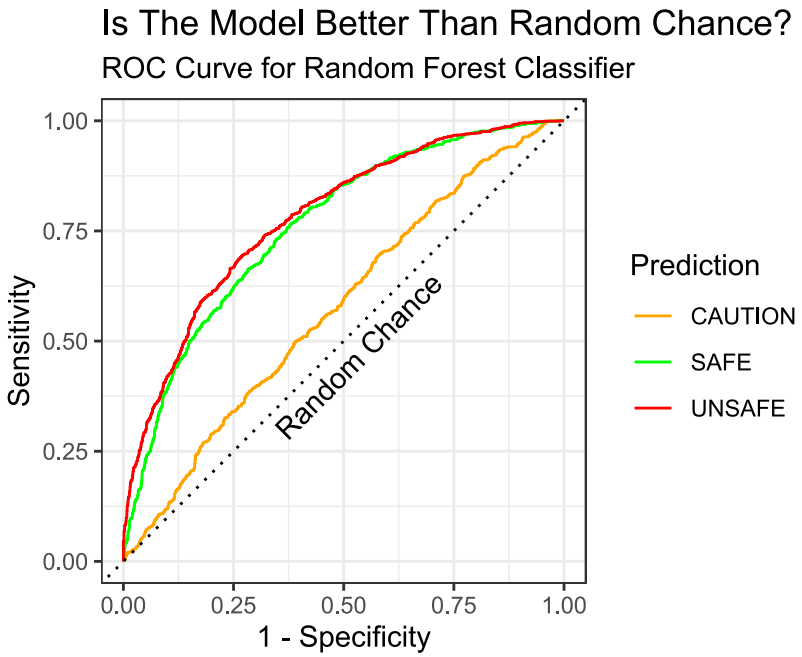
| | Prediction | | | |
|---|---|---|---|---|
| Truth | SAFE | CAUTION | UNSAFE | Total |
| SAFE | 492 | 76 | 250 | 818 |
| CAUTION | 225 | 63 | 288 | 576 |
| UNSAFE | 176 | 71 | 903 | 1,150 |
| Total | 893 | 210 | 1441 | 2544 |

In a regression analysis, we measure the coefficient of each of the input variables so there is a precise measure of how much each variable contributes to the prediction. In a random forest, we don't have linear relationships but we can measure the relative importance of each variable.

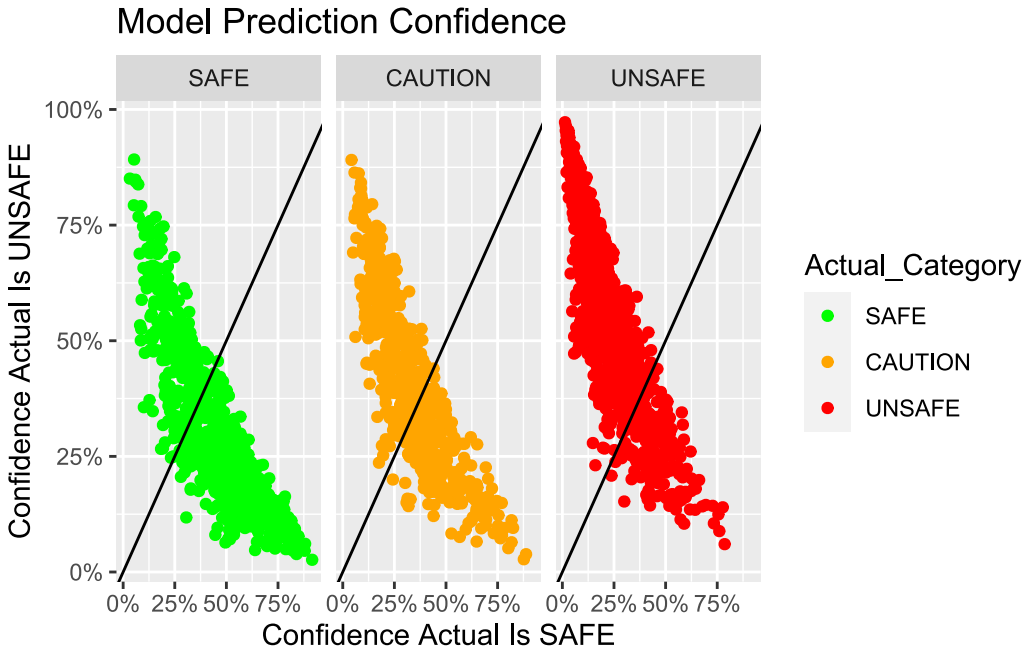
The `site` input is nearly twice as important as the other inputs. Since location doesn't change over time, modeling might seem superfluous. Overall results can be summarized with the the single *"Kappa"* statistic which indicates the model is about 30 percentage points better than random chance. If we remove the `site` input from the model, the prediction validity is reduced, as we'd expect, but it is still 20% better than random chance. In such a model 2-day rainfall becomes the most important input.




The "Receiver Operator Curve" visualizes how much better than model is than random chance. Curves that bend up and to the left are better. We can see that the "CAUTION" predictions are barely better than a coin flip.




Every prediction the model makes includes a confidence level. The classification with the highest confidence level is what the model predicts. How certain is the model that the prediction is correct? The plot below shows the confidence level for each prediction. The majority of the classifications are correct but we can see the model is highly confident in many cases where it is wrong.



Note

We can illustrate the importance of separating training and testing data. It's easy to "overfit" when including all of the data. In the example below we train on all of the water quality data. The accuracy and prediction confidence is very high but it's an illusion. We don't know anything about predictive ability in the future.

## Prediction is "Easy" When We Overfit

Truth Table

| | Prediction | | | |
|---|---|---|---|---|
| Truth | SAFE | CAUTION | UNSAFE | Total |
| SAFE | 3,152 | 11 | 108 | 3,271 |
| CAUTION | 103 | 2,016 | 182 | 2,301 |
| UNSAFE | 44 | 4 | 4,552 | 4,600 |
| Total | 3299 | 2031 | 4842 | 10172 |

## Conclusion

We have created a random forest model that shows modest accuracy in predicting whether enterococci levels will be at the extremes of "safe" or "unsafe." The model is not very good at predicting the middle "caution" levels of bacteria. Unfortunately, the site location itself is the most important predictor of bacteria levels so changing environmental factors like tide, temperature and rainfall tell us little about levels over time. Understanding point sources of pollution around each site and how they vary might be more fruitful.