Predicting risks of toxic algal blooms in California coastal region

1. Introduction

1.1 Background

While algae are important players in our ecosystem and the marine food web, under certain conditions, some algae can grow rapidly and release large amount of toxins that are harmful to human and other marine animals. Such biological events are known as harmful algal blooms. There are many types of harmful algal blooms; one of them is of particular interest to pacific coastal states of the U.S. That is *Pseudo-nitzschia* bloom, which produces the toxin, domoic acid. Domoic acid is a neurotoxin that can cause serious neurological symptoms, and sometimes death, if ingested. It is not only dangerous to swimmers and beachgoers, but it can also accumulate in seafood, especially shellfish such as crabs and clams. Therefore, high domoic acid level in coastal waters is not only a public health concern, but also an economic risk. For example, the harmful algal bloom in 2015 caused closures and delays of dungeness crab, rock crab season, which costed millions of dollars to the fishing/seafood industry.

1.2 Problem

Physical, chemical, and biological conditions such as water temperature, nutrients in the water, seasonality and other algae in the water may contribute to the formation of harmful algal blooms and the production of toxins. This project aims to use these data to predict risks of toxic algal bloom events in California coastal region.

1.3 Interest

Because of public health and economic implications of toxin levels in coastal waters, many entities will be interested in the prediction of those levels. They include government agencies in charge of public health, environmental protection and fishing regulation. The fishing and aquaculture industries, which are worth billions of dollars, will also be heavily interested in the prediction. Downstream industries such as restaurants and packaged seafood industry might also be interested.

2. Data acquisition and cleaning

2.1 Data sources

Southern California Coastal Ocean Observing System (SCCOOS) has been monitoring domoic acid concentrations and accompanying physical, chemical, and biological data since 2008. Weekly data of 8 stations from San Francisco to San Diego were downloaded via their website. Other oceanography data were added to the analysis. They include daily ocean upwelling indexes provided by Pacific Fisheries Environmental Laboratory and the Oceanic Niño Index (ONI), which monitors climate variability in the ocean, provided by the National Oceanic and Atmospheric Administration.

2.2 Data cleaning

The main dataset, the SCCOOS harmful algae dataset, contained 3,490 samples. However, it had a major problem, which was that about two thirds of the samples (2,342 out of 3,490) had the dependent variable (domoic acid concentration) labeled as NaN (missing). That would mean losing two thirds of the data, which was not acceptable to me. I contacted a colleague, Dr. Jayme Smith, who collaborates with SCCOOS. She told me that many of the domoic acid data on the SCCOOS website were mislabeled as NaN, but instead should be 'n/d' (not detected) or 'b/d' (below detection), which essentially means close to zero. I obtained the correctly labeled domoic acid data from her (included in GitHub repository).

Her data were in several different files and formats. I extracted and re-formatted her data and merged them with the SCCOOS dataset. I used her domoic acid data to replace those missing values in the SCCOOS dataset. This exercise saved more than 1,500 samples. The remaining entries with truely missing dependent variable data were removed.

Two other datasets, ocean upwelling index, and Oceanic Niño Index were downloaded from websites. Data of relevant dates/months and locations were extracted and merged with the main dataset based on sample collection dates/months, and proximity between the sampling locations of different data.

Some independent variables were removed for different reasons. For examples, a few of them had missing values in more than half of the samples. Location data was also removed, because the prediction is intended for region as a whole, not for individual sites. Volume data were also removed because they reflect detection limit, which is not relevant to the model.

Outliers in the data were examined. There were three cases in which I was sure the outliers were created by mistake. Two were negative chemical concentrations, which are not possible, and one was temperature of 0 degree, which was also not possible for the ocean. They were changed to missing values.

Most of the remaining independent variables had some missing values, around 10%, up to 30% in some. I noticed that in some, if not all of them, there were seasonal patterns, for example, water temperatures were higher in the summer. Therefore, to better impute missing values, I used monthly means instead of means of entire dataset to fill in those missing values.

In the end, there were 2,750 samples and 18 variables (1 dependent, 17 independent).

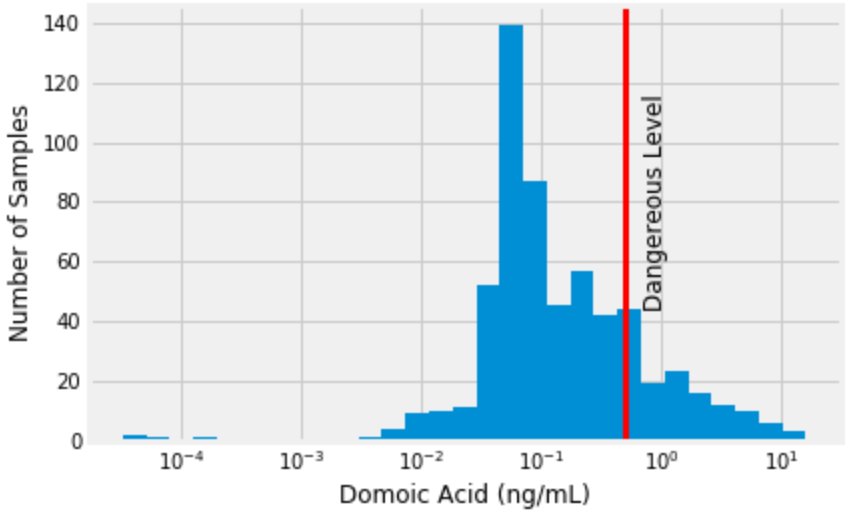
3. Exploratory data analysis

3.1 Distribution of the dependent variable

The distribution of the dependent variable (domoic acid concentration) was very skewed. Out of 2750 samples, 2155 had a value of zero. Non-zero domoic acid values also had a skewed distribution. According to domoic acid expert Dr. Jayme Smith, domoic acid concentration of 0.5ng/mL or above can be considered dangerous. Out of 2750 samples, only 120 samples had dangerous levels of domoic acid (Figure 1). Given the nature of this dataset, a classification model was probably more suited than a regression model. It was also important to note the

imbalanced nature of this datasets and the small number of samples belonging to the positive class.

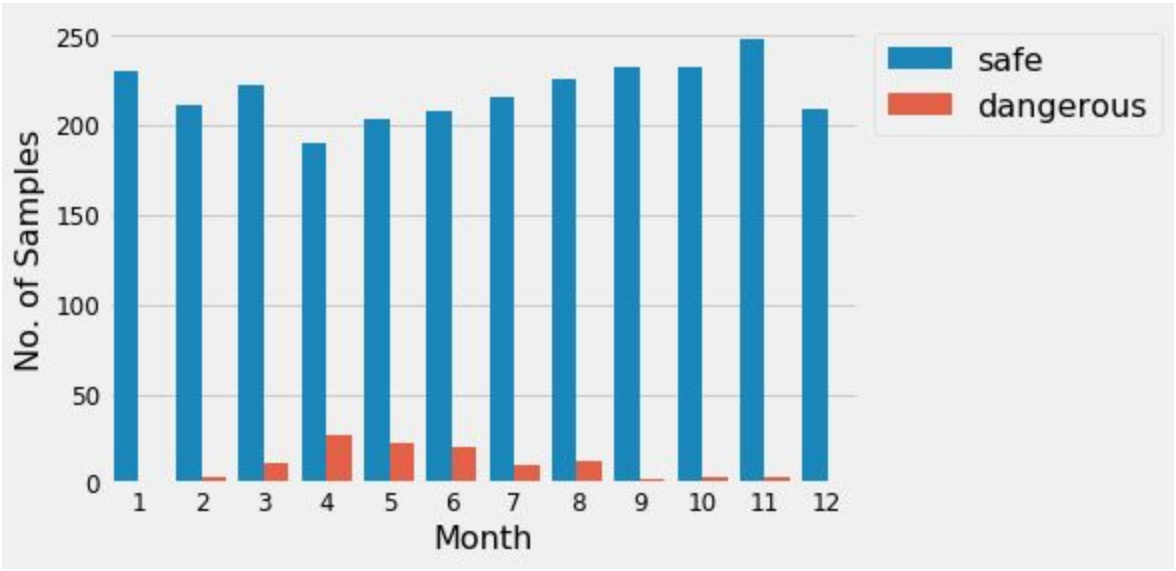
Figure 1. Distribution of domoic acid concentrations (non-zero values only)



3.2 Monthly or seasonal patterns of toxic events

Domoic acid concentrations above dangerous threshold generally occur during spring and summer months. The distribution of toxic events across different months were statistically different (chi-square test, $p<0.001$, Figure 2).

Figure 2. Numbers of samples with safe and dangerous domoic acid levels in different months

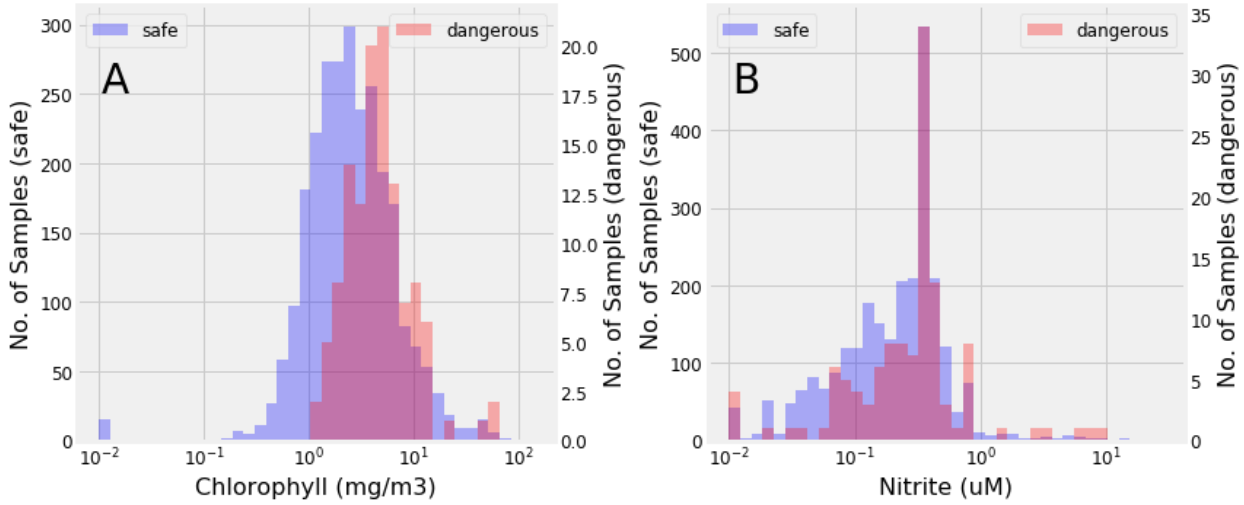


3.3 Correlations between chemical data and risks of toxic events

This dataset had measurements for chlorophyll concentration, an indicator of abundance of algae in the water. Domoic acid is produced by some but not all algae. Therefore I hypothesize a positive but non-linear relationship between chlorophyll concentrations and the risk of toxic

events. Chemical data in this dataset all had very skewed distribution, with lots of small values, and few large values. It is better to use mean of logarithms of the data than simple arithmetic mean for comparison of populations. Mean of logs of chlorophyll concentrations were higher in dangerous samples than in safe samples (z-test, $p<0.001$, Figure 3A). This dataset also had concentrations of several nutrients such as nitrate, phosphate, etc. The growth of algae depends on these nutrients, and different algae require different amounts of nutrients. I hypothesize the concentrations of some nutrients would have positive relationships with toxic event risks. Statistical tests showed that the means of logs of nitrite concentrations statistically higher in dangerous samples than in safe samples (z-test, $p=0.006$, Figure 3B). Concentrations of others were not statistically different between the two categories.

Figure 3. Distribution of chlorophyll (A) and nitrite (B) concentrations in samples with safe and dangerous domoic acid levels.



3.4 Correlations between cell counts and risks of toxic events

Concentrations of seven different algae were also part of the datasets. They include two Pseudo-nitzschia species, and five others. Pseudo-nitzschia are the known main producers of domoic acid, therefore I expect a strong, possibly linear relationship between their concentrations and the risk of toxic events. Indeed, the concentrations of both species were significantly higher in dangerous groups than in safe groups (t-test, $p<0.001$). Though, the differences were larger for Pseudo-nitzschia seriata than in Pseudo-nitzschia delicatissima (Figure 4). In fact, the concentrations of Pseudo-nitzschia seriata and domoic acid were somewhat linearly correlated (Figure 5).

Other five species don't produce domoic acid, but they may have biological interactions with pseudo-nitzschia, such as competition. Their concentrations may or may not have any impact on the risk of toxic events. The differences in their concentrations were not statistically significant between safe and dangerous groups.

Figure 4. Distribution of Pseudo-nitzschia delicatissima (A) and Pseudo-nitzschia seriata (B) cell concentrations in samples with safe and dangerous domoic acid levels.

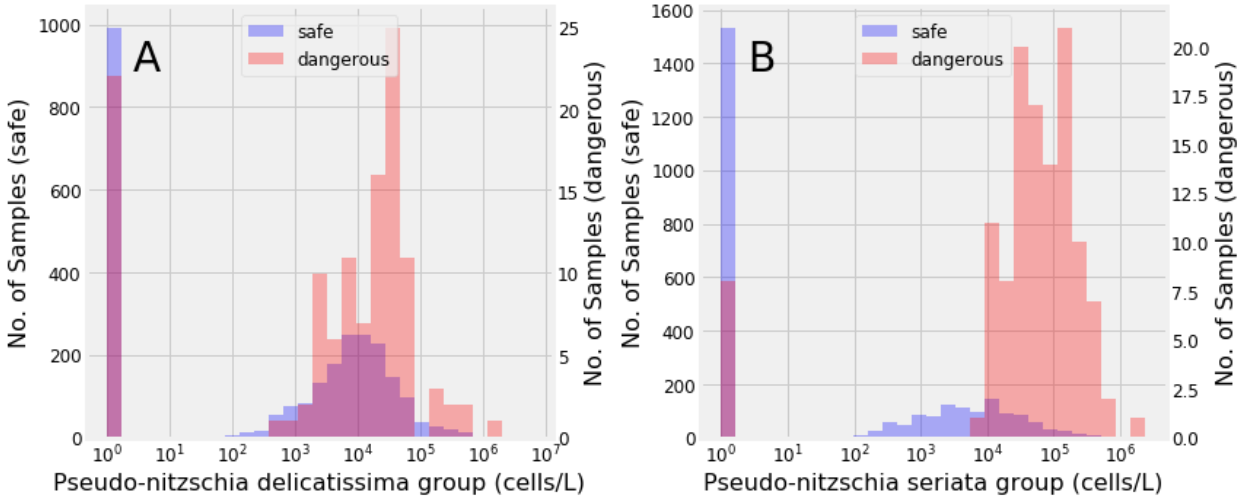
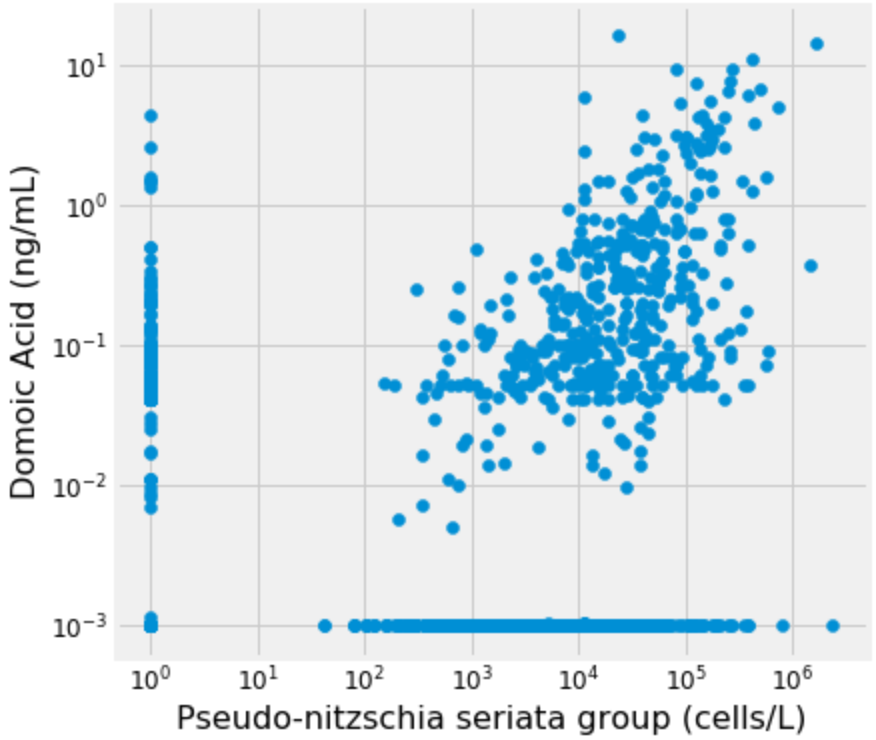


Figure 5. Scatter plot of concentrations of Pseudo-nitzschia seriata and domoic acid.



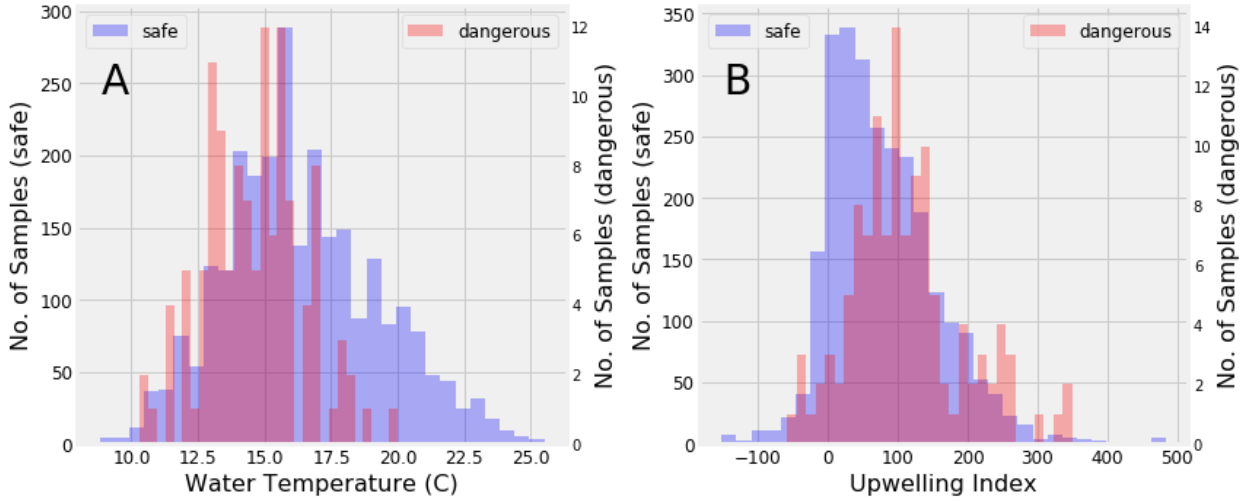
3.5 Correlations between physical data and risks of toxic events

Temperature usually has a significant impact on microbes. Therefore I expect there will be an optimal temperature range for growth of algae and production of algae. Indeed, the mean temperature of dangerous samples was significantly lower than that of safe samples (z-test,

p<0.001). There were no dangerous samples above 20 degrees celsius. Also, most of the dangerous samples had temperatures between 12 and 18 degrees celsius (Figure 6A).

I also observed higher upwelling index among dangerous samples than safe samples (z-test, p<0.001, Figure 6B). This is because ocean upwelling brings nutrient-rich water up to the surface, which is suitable for algal growth.

Figure 6. Distribution of water temperature (A) and upwelling index (B) in samples with safe and dangerous domoic acid levels.



3.6 Correlation between independent variables

I calculated the pearson correlation coefficients between all independent variable pairs (some variables were log-transformed before calculation). The strongest correlations were among nitrate, nitrite, phosphate. Their correlation coefficients were approximately 0.5. They were also negatively correlated with water temperature with correlation coefficients around -0.4. This is a known phenomenon in that nutrient-rich waters are generally colder. These correlations were not strong enough to consider discarding any of the variables.

4. Machine learning

4.1 Preprocessing and overall design

After exploratory data analysis, I decided that classification models based on whether domoic acid concentrations were above the dangerous threshold would be most suitable for this problem. Therefore, domoic acid data were extracted and converted into binary variables. Dates were transformed into four seasons.

Skewed distributions and varying scales in the datasets were expected to have negative impacts on algorithms that use distances in their calculations, such as logistic regression and support vector machines. Therefore, to mitigate this problem, variables that had skewed distributions, which were all the chemical data and algae cell counts, were log-transformed. All numerical variables were scaled to mean of zero, and standard deviation of 1. Data were then

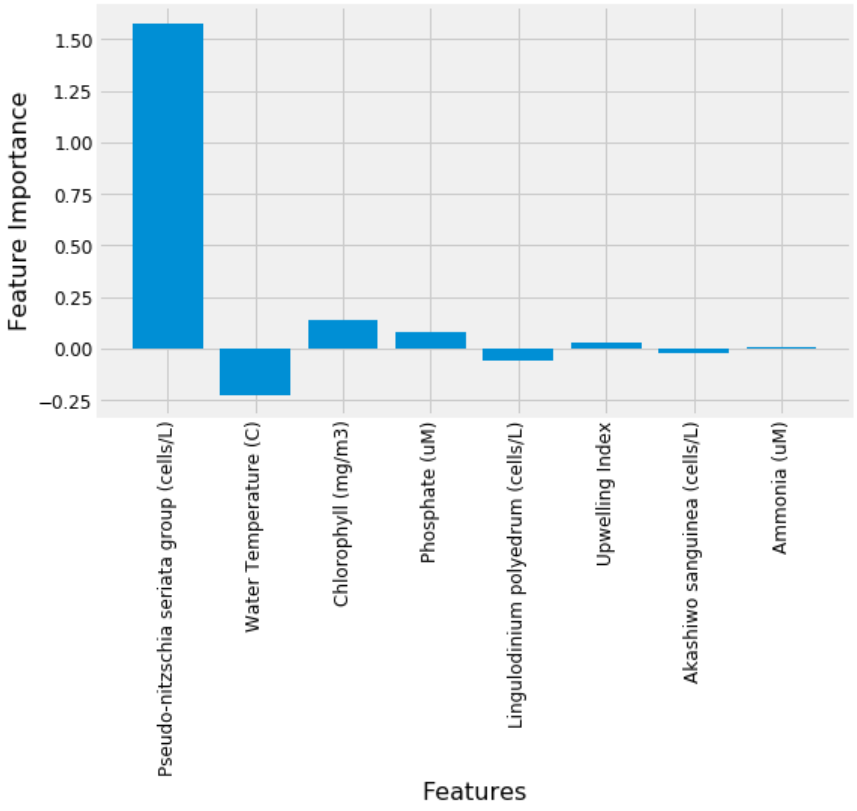
randomly split into a test set consisting of 25% of the data, and a training set that contained the rest of the data.

This dataset was very imbalanced in that positive samples only accounted for less than 5% of all samples. This imbalance is known to have negative impacts on classification models. To combat this, I added an oversampling step using the Synthetic Minority Over-sampling Technique (SMOTE) before training the models. Also, the models were evaluated using average precision, which is more sensitive to changes in imbalanced datasets, instead of accuracy or area under the Receiver Operating Characteristic curve.

4.2 Logistic regression

Using grid search with cross-validation, I found that using L1 regularization and regularization parameter C of 0.01 yields the best logistic regression model. The resulting model had high recall (correctly predicting 27 out of 29 positive samples), but had low precision of only 0.14 (Table 1). Among all the features, cell counts of Pseudo-nitzschia seriata had the largest coefficient, water temperature and chlorophyll concentration had the next largest coefficients. These were consistent with observations made during exploratory data analysis. However, several other features that significantly impacted toxic event risks, such as season, upwelling index, etc., had very small or zero indices (Figure 7). This was a result of L1 regularization. It limited the numbers of features the model considered. Maybe other models that consider more features would perform better.

Figure 7. Coefficients of logistic regression model. Coefficients of zero were not plotted.



4.3 Support Vector Machines

Hyperparameters C and gamma were tuned using grid search with cross-validation. Support vector classifier using the best hyperparameters had higher precision (0.26 vs. 0.14) than logistic regression model at the cost of recall (0.69 vs. 0.93) (Table 1). Overall performance in terms of average precision was similar to that of logistic regression.

Table 1. Performances of models evaluated using the hold-out test data set. Best performance in each category is colored red.

	Logistic Regression	Support Vector Machine	Random Forest	Gradient Boosting	Voting Classifier
Average precision	0.34	0.34	0.50	0.52	0.53
Accuracy	0.75	0.91	0.97	0.97	0.91
No. of true positives	27	20	20	18	22
No. of false positives	173	56	15	12	20
No. of false negatives	2	9	9	11	7
No. of true negatives	486	603	644	647	639
Precision*	0.14	0.26	0.57	0.60	0.52
Recall*	0.93	0.69	0.69	0.62	0.76

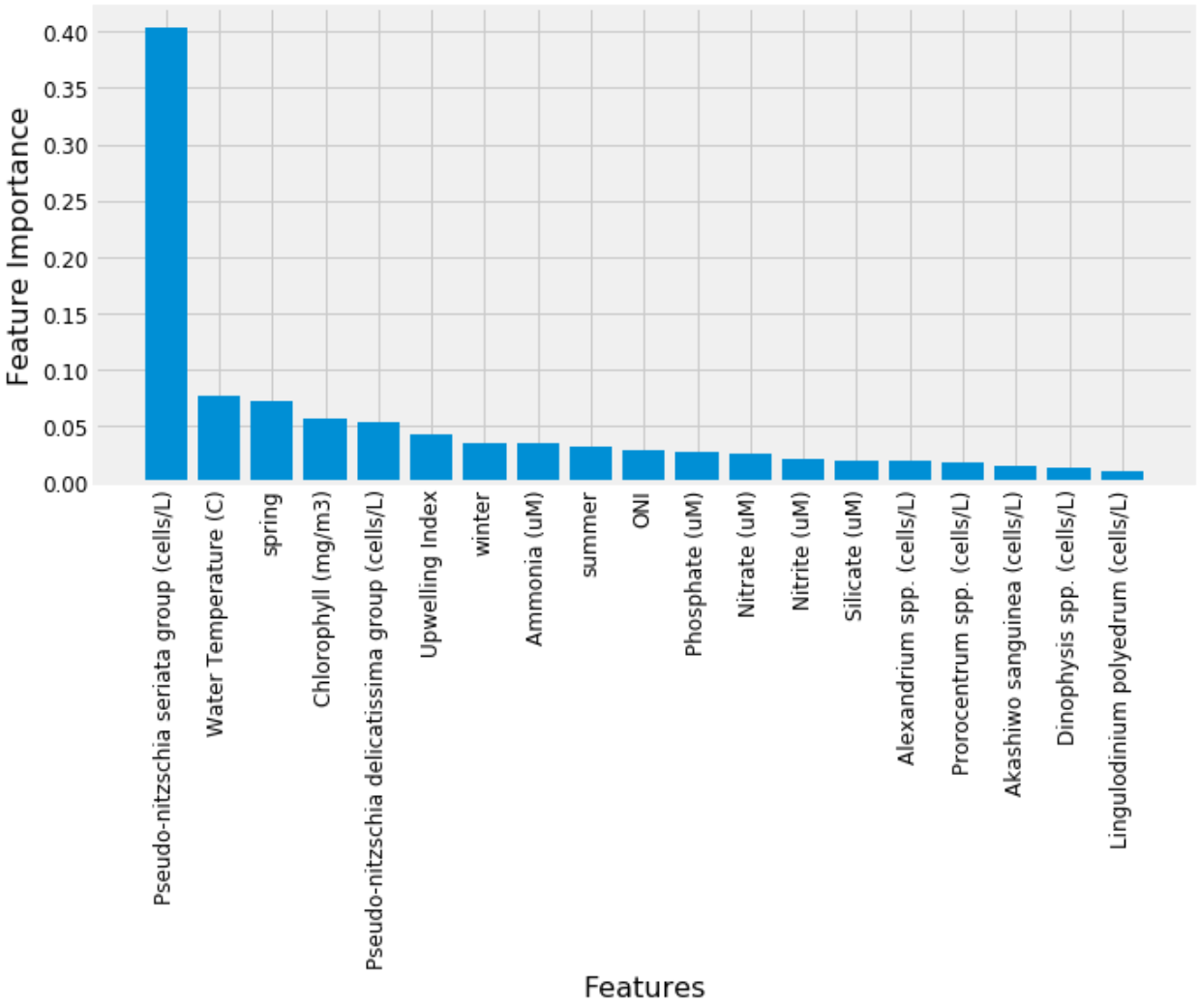
* of the positive class.

4.4 Random Forest

Several hyperparameters including the number of trees, maximum depths of trees, minimum numbers of samples required to split a node and form leaf nodes, were tuned using grid search and cross-validation. The random forest model with the best parameters had much better performance than support vector machines. It had the same recall (0.69), but much better precision (0.57 vs. 0.26). The random forest model was also better in terms of average precision and accuracy (Table 1). Pseudo-nitzschia seriata was by far the most important feature. Similar to logistic regression, water temperature and chlorophyll concentration were the next most important features. However, random forest model also considered season, upwelling index, and

Pseudo-nitzschia delicatissima to be similarly important, unlike in logistic regression model where these features were ignored (Figure 8).

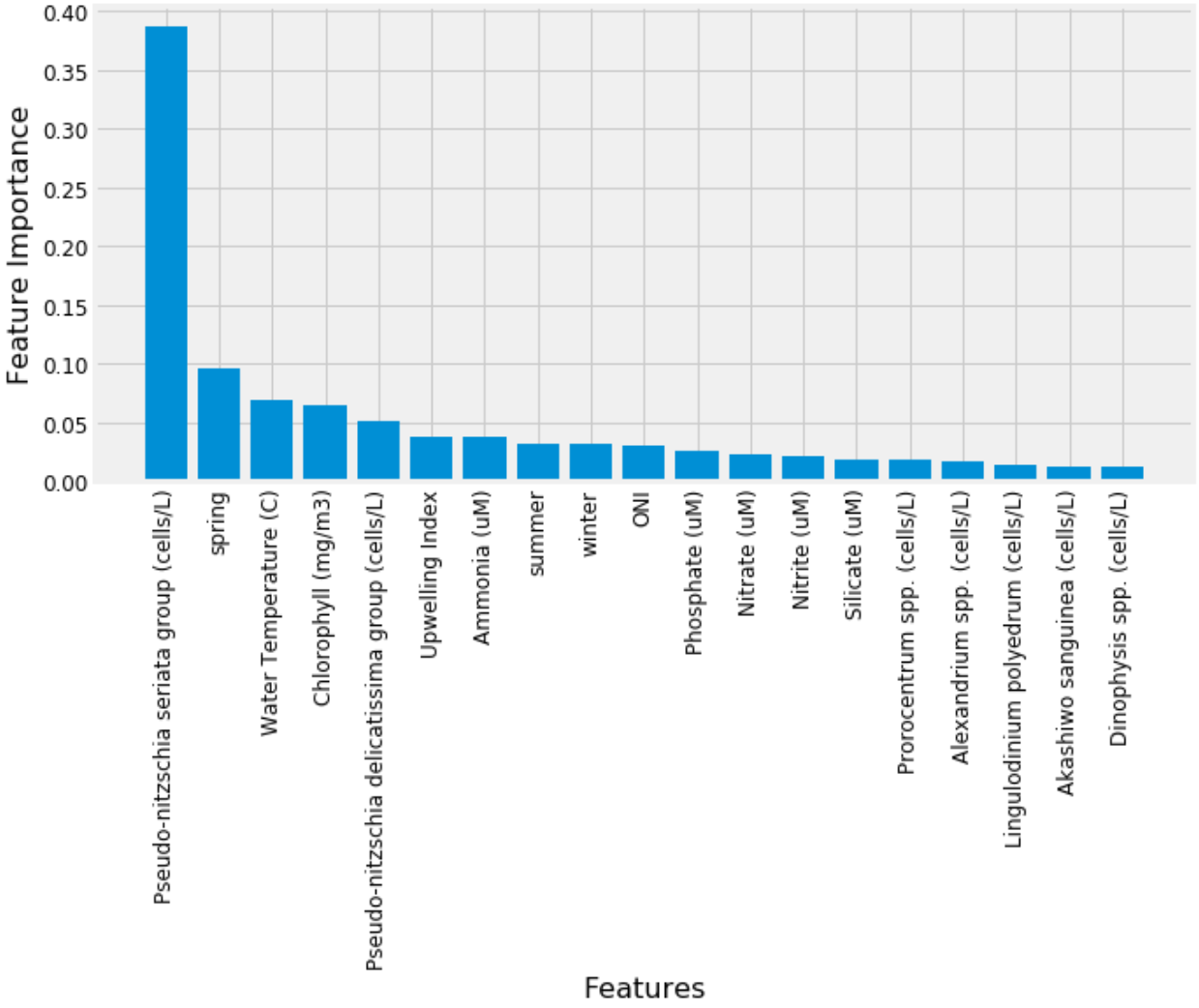
Figure 8. Feature importances of the random forest model



4.5 Gradient Boosting

Parameters used in the random forest model, except the number of trees, were also used in the gradient boosting model. The number of trees and the learning rate, were tuned using similar approach as above. The gradient boosting model had similar if not slightly better performance than the random forest model. It predicted the fewest number of false positives, and most true negatives. It had the highest precision, slightly lower recall and slightly higher average precision than random forest model (Table 1). In terms of feature importance, it was very similar to the random forest model, with slightly less importance for Pseudo-nitzschia seriata and slightly more importance for seasons (Figure 9).

Figure 9. Feature importances of the gradient boosting model



4.6 Voting Classifier

Overall, the tree based models (random forest and gradient boosted trees) performed better than the distance based models (logistic regression and support vector machines). However, the logistic regression model still had its unique strength in its higher recall than the tree based models, even though its precision is much lower. An ensemble model could potentially combine strengths of these single models and perform better.

A soft voting classifier was built by tuning the weights of different models using grid search and cross-validation. The best parameter was combining logistic regression with gradient boosting models with 1:1 weights. The voting classifier had the highest average precision of all models (Table 1).

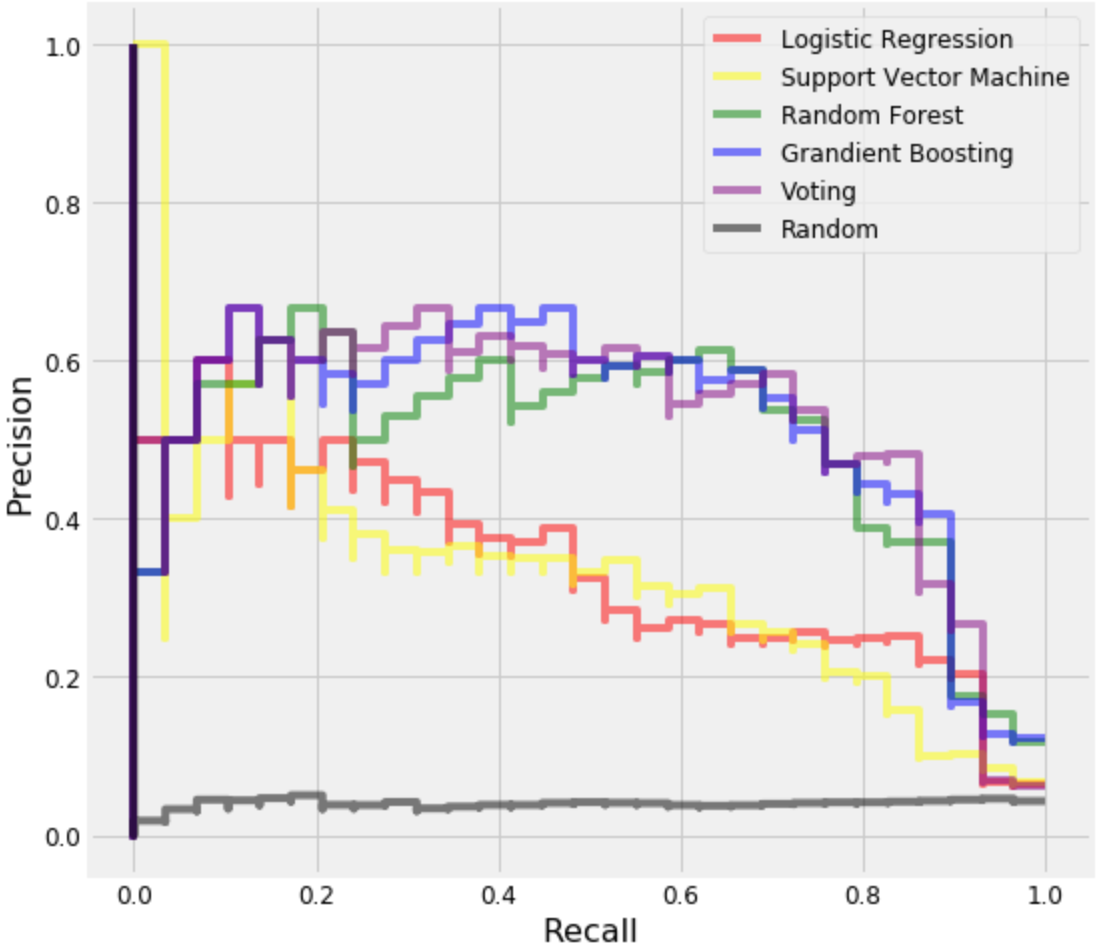
4.7 Model evaluation

In addition to the metrics listed in Table 1, a precision-recall curve was plotted for all models (Figure 10). It was clear from the curve that the logistic regression and support vector machines

models did not perform as well as the other three models. The differences in performance among the random forest, gradient boosting, and voting classifier were small.

In this problem of predicting risks of algal toxic events in the coastal region, higher recall is much more important than higher precision. A false positive prediction would probably result in unnecessary testing of domoic acid in water and seafood and maybe limitation on swimming in the ocean, etc. The consequences are mostly economic. A false negative prediction, on the other hand, could potentially have catastrophic consequences, such as public health crisis, or even death. Therefore, when evaluating this model, it is important to look at the performances of models when high recall is required. The voting classifier performed the best with recall between 0.8 and 0.9. All models had poor precision when recall was higher than 0.9, though the random forest model performed the best in this area (Figure 10).

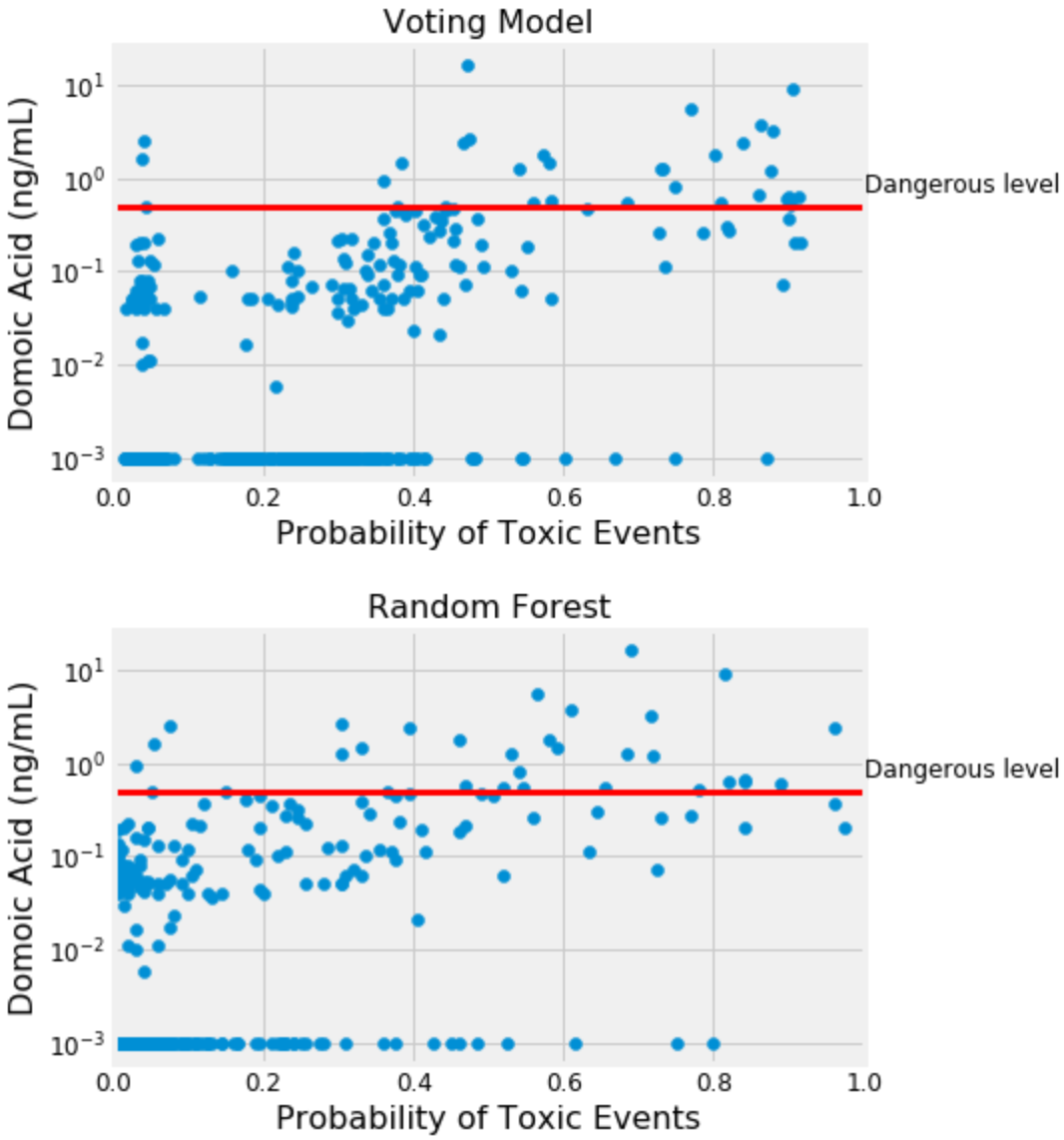
Figure 10. Precision-Recall curve for all models in this study using the test dataset



In practice, the probability of toxic events rather than binary dangerous/safe predictions would be used in decision making. In addition, different domoic acid concentrations above the dangerous threshold also carry different amounts of risks. Therefore, it is important to look at how predicted probability matched up with domoic acid concentrations. For example, it is somewhat acceptable if a borderline dangerous sample had a predicted probability smaller but

close to 0.5. On the other hand, it would be very bad if a sample with very high domoic acid concentration had a probability less than 0.2. I examined two of the best models, the random forest model and the voting classifier, using a plot of predicted probability and domoic acid concentration. The voting classifier predicted probabilities above 0.4 for all but 4 samples while the random forest model predicted probabilities less than 0.4 for 7 dangerous samples (Figure 11). It was evident that the voting classifier would be more useful than the random forest model in practice.

Figure 11. Scatter plot of predicted probabilities and domoic acid concentrations for the voting classifier (top) and the random forest model (bottom).



5. Future directions

The biggest challenge in this study was the small number of positive class samples. The small number resulted in more randomness in validating and testing the models. Collecting more data

would be the most obvious next step in improving the models. Potential data sources include similar data collected by other states or countries, data collected by remote sensors at offshore locations, and/or encouraging more frequent data collection from California collaborators. It is important to inform collaborators about the most important features learned from this study (Figure 8 & 9), so that at least data of these features are collected.

There were a few dangerous samples that none of our models got right (Figure 11, the leftmost two dangerous samples). This suggested that they were caused by reasons not reflected in this dataset. Further exploring what might have caused the toxic events in those samples might lead us to other important features of which we are not aware yet.

It would also be interesting to build models for predicting risks of future toxic events, for example, one or two weeks in advance, since data were collected weekly. Such future models would allow decision makers more time to assess risks and take proactive measures for potential toxic events.