Predicting Harmful Algae Blooms

Team 15

Anushka Wani, Rachel Weiss, Helena Fu, Ashley Bao, Ajibola Falade

Meet our team!

Anushka Wani

Rachel Weiss

Helena Fu

Ashley Bao

Ajibola Falade

Our AI Studio TA and Challenge Advisors

Divya Nori

Mia Maksin

Alyssa Long

Agenda

1. Our Challenge
2. Data Preparation
3. Models and Evaluation
4. Next Steps

Our Challenge

Scientific Background

Many factors of HABs are known, but how they contribute to “blooms” is less understood

What are HABs?
When colonies of bacteria grow out of control, leading to harmful impacts on living things and their ecosystems

What are the main causes?
Nutrient pollution, climate change impacts, human activity, sometimes post natural weather storms

Important **factors?**
Warmer waters, slower and shallower water, increase of phosphorus and nitrogen, sunlight, pH, turbidity

Impacts and Implications
Ecosystem damage, biodiversity loss, pollution, disease/deaths

Our Goal

Harmful Algal Blooms (HABs) cause devastating impacts to surrounding ecosystems.

- Use AI/ML to determine which factors alone or in combination if any, can be early predictors of HABs in various environments.
- Present a comprehensive review of the chosen factors and whether they can be used as early predictors of HABs.

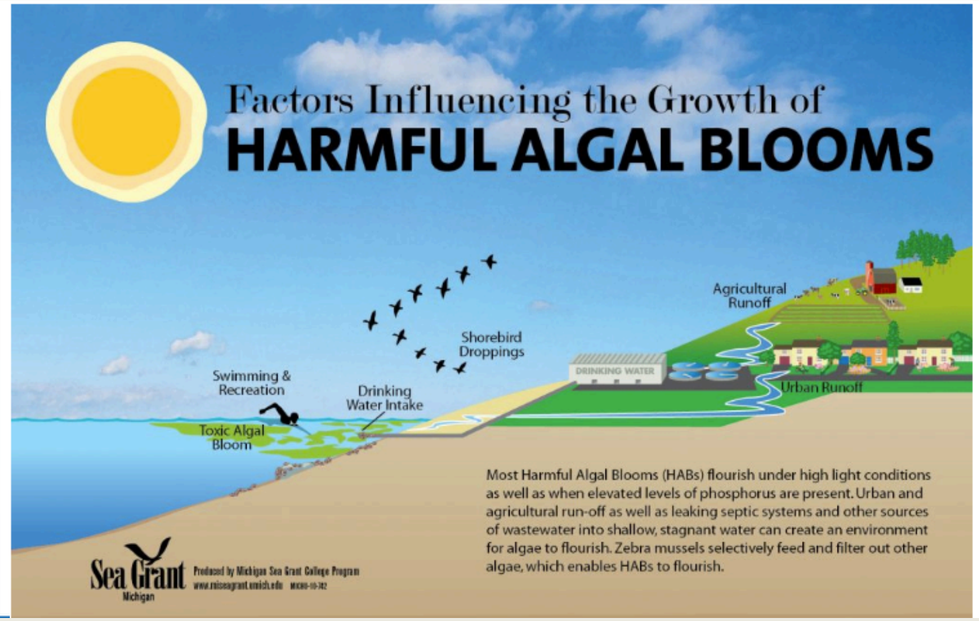
Business Impact

- Biointerphase wants to use the information we gather to make an impact on HAB mitigation
- They seek external funding through a variety of sources and is actively pursuing projects related to HAB mitigation
- Advancements in predictive tools could be beneficial to understand when and where HAB mitigation tools should be deployed
- This project will help Biointerphase show how AI/ML can be deployed in the HAB realm

Data Preparation

Data Understanding

- Built domain knowledge to inform data cleaning, preprocessing, and model-building.
- Consulted resources like NOAA for deeper insights into our topic area.
- Identified limitations in our initial dataset and sought additional data to enhance our models.



Datasets

Original Dataset

- From the National Oceanic and Atmospheric Administration (NOAA)
- Specifically around the Gulf of Mexico
- Only data on algae concentration, location, water temperature, and salinity
- Not enough data for a model

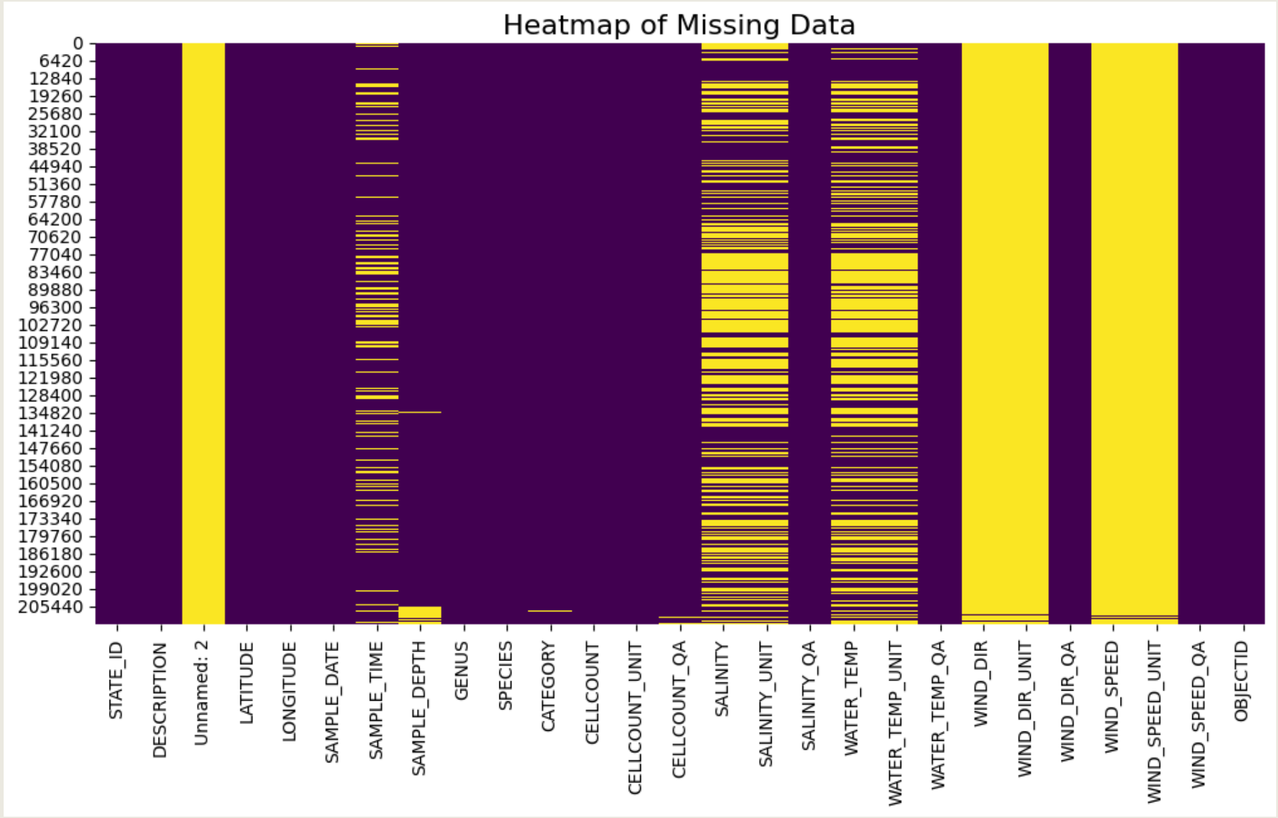
Additional Dataset

- From Biogeochemical (BGC) Argo
- Information from inside the ocean using drifting profiling floats
- Data on location, water temperature, salinity, and nutrient levels
- No information about algae concentration

Data Preprocessing Attempts

Why we added Additional Data

- NOAA dataset was discarded due to excessive missing data in key features (shown on right)
- Key feature columns were also not included in the NOAA dataset, so additional data was needed to create an accurate model
- Sourced additional datasets with more complete water quality data.



Our Plan

Create Model 1 from BGC ARGO data that predicts water nutrient data using salinity, location, and water temperature

→

Use Model 1 to populate artificial nutrient data for the NOAA dataset

→

Train Model 2 with augmented data to predict algae blooms

Model 1: Preparation

- Normalized numerical data
- Date-time encoding
- Features: Latitude, Longitude, Depth, Pressure, Temperature, Salinity, Date, Time
- Labels: Sigma_Theta (seawater density), chl_a (Chlorophyll a - allows algae to photosynthesize), Oxygen, Oxygen Saturation, Nitrate

Model 1: Results

RandomForestRegressor Model, Multioutput

- The model explains a large proportion of the variability in the data
- Possible concerns about overfitting

```
R-squared for pressure: 0.9999998892540899
R-squared for sigma_theta: 0.9991498469094865
R-squared for chl_a: 0.966093672588821
R-squared for oxygen: 0.9929825134024218
R-squared for oxygen_saturation: 0.9956214176103647
R-squared for nitrate: 0.9979924010278399
R-squared for b_bp700: 0.8646724010142063
R-squared for POC: 0.8633131011883625
R-squared for CDOM: 0.9884055662439148
R-squared for pH_insitu: 0.9954018692435777
R-squared for pH25C: 0.9977732250004491
R-squared for TALK_LIAR: 0.9972880178234131
R-squared for DIC_LIAR: 0.9959995255741582
R-squared for pCO2_LIAR: 0.993117023365271
Overall R-squared: 0.9748436050175983
```

We then used Model 1 to predict nutrient, oxygen, etc. levels for our HABSOS data.

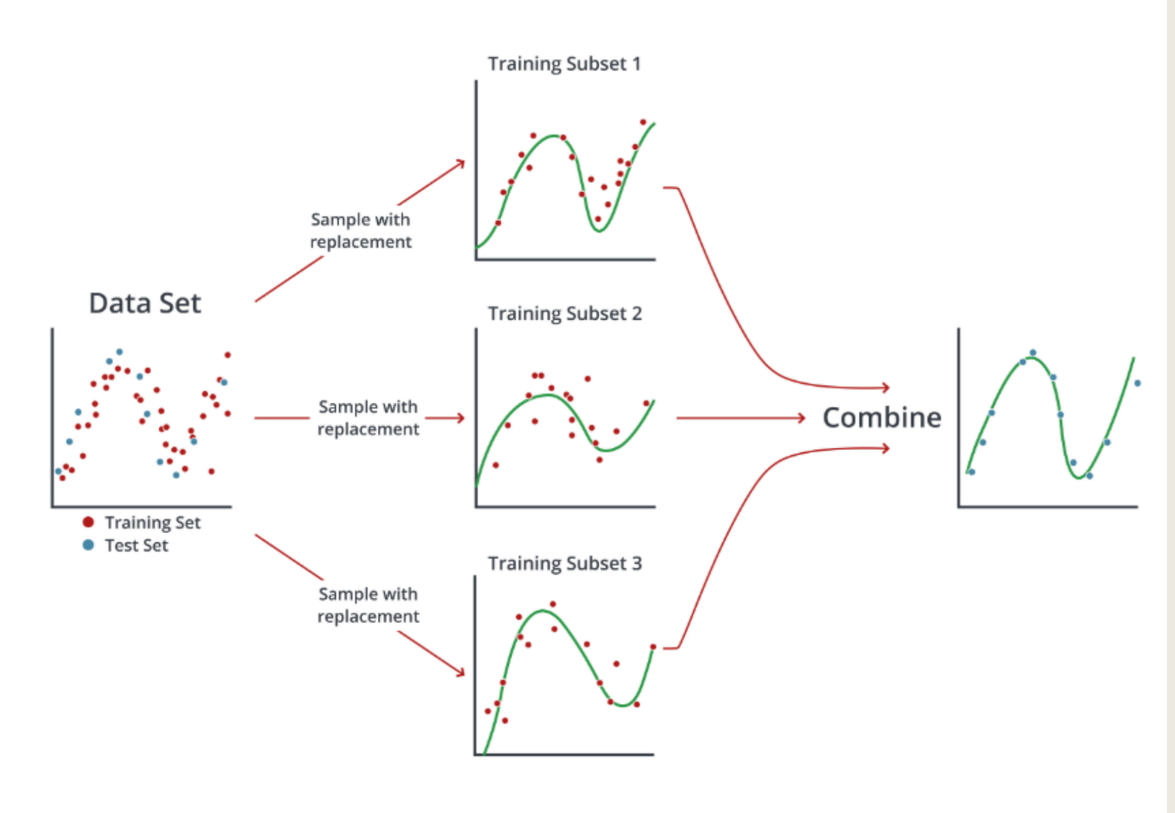
Model 2: Preparation

- Removed text-based and unnecessary columns to ensure only relevant features were included.
- Used one-hot encoding for categorical variables (e.g., region and species names)
- After scaling features, we split the dataset into training and test sets, setting up a robust evaluation pipeline

Model 2: Model Selection cont.

Looked at RandomForestRegressor, GradientBoosting, and KNeighborsRegressor

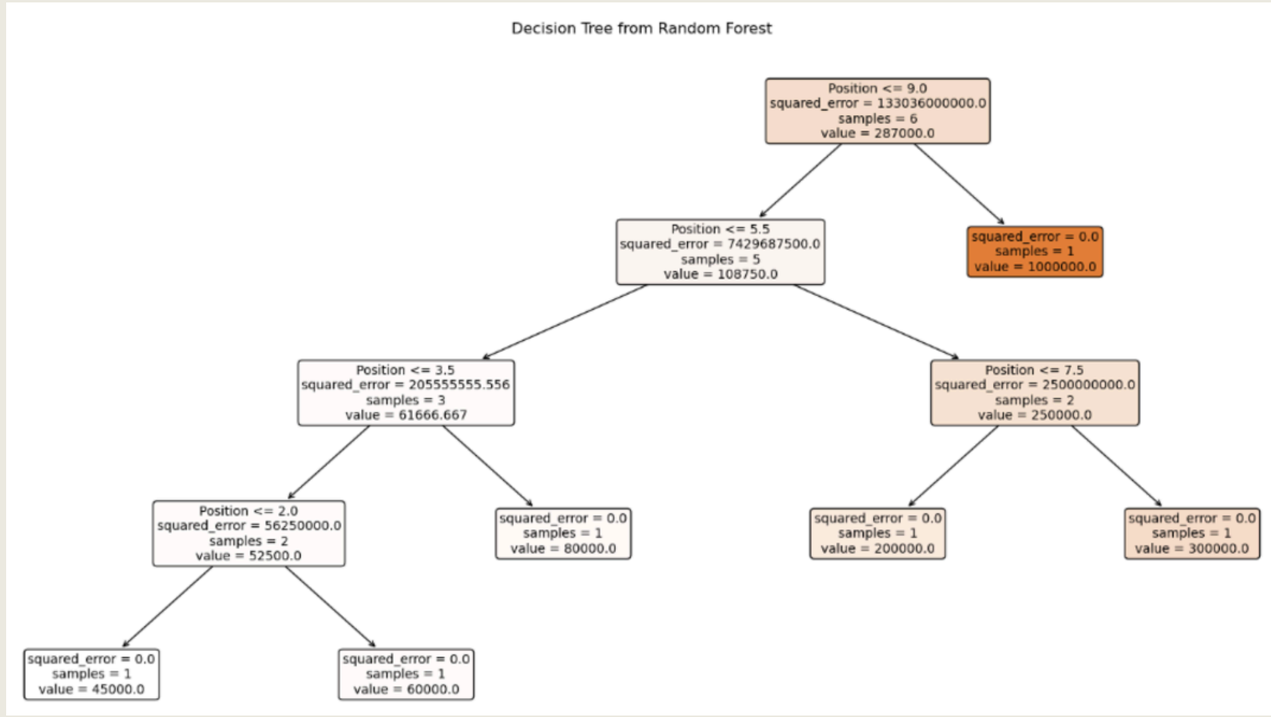
- RandomForest had the best results: Initial results showed an R^2 of around 0.26 and an MAE of ~140,000, indicating some predictive power but room for improvement



Model 2: Model Selection

For Random Forest, we conducted hyperparameter tuning to improve results further, testing values like the number of trees, max depth, and sample splits.

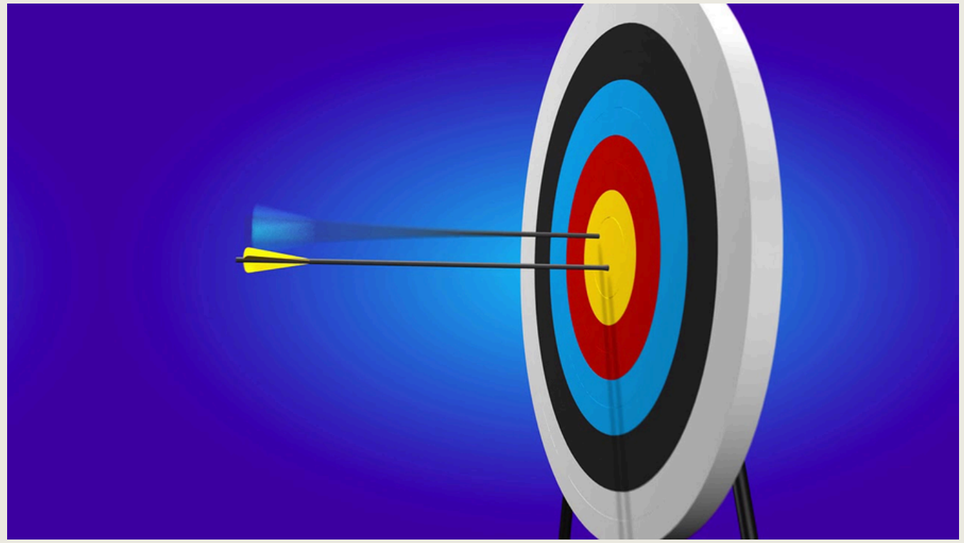
- Tuning provided incremental improvements but kept R^2 around 0.2 - 0.26.



Model 2: Improvements

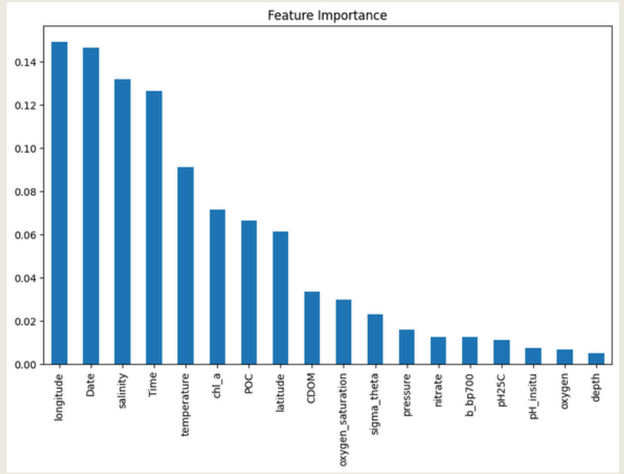
After applying the log transformation and refining the dataset, the Random Forest Model achieved:

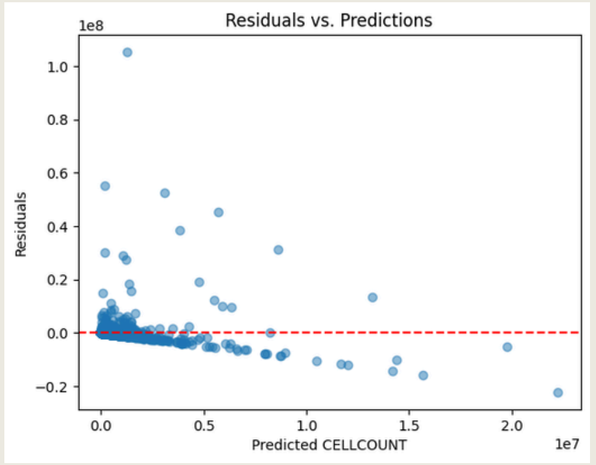
- Mean Absolute Error (MAE): 1.09
- R^2: 0.766 indicating a significant improvement in predictive accuracy compared to prior iterations.



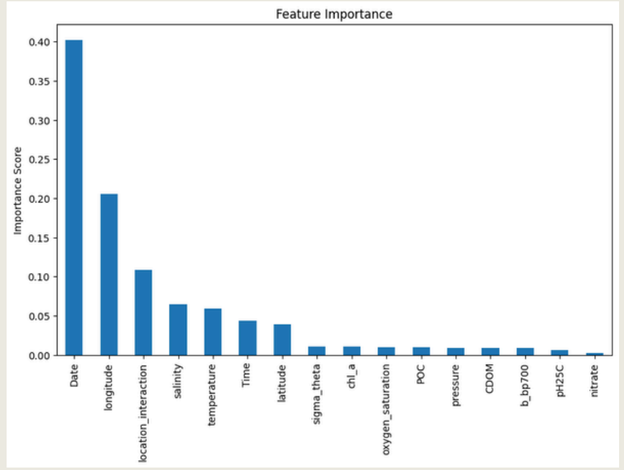
Model 2: Improvements

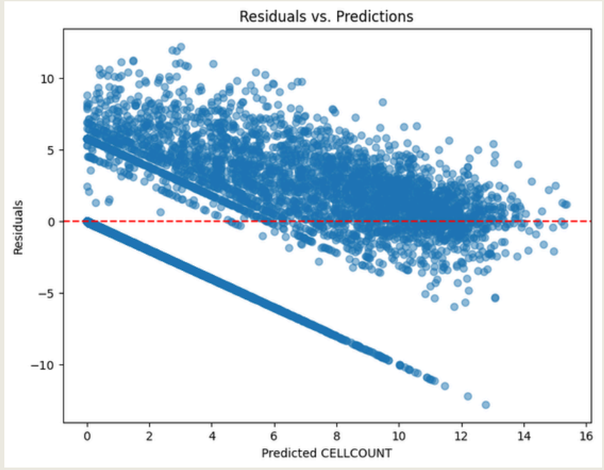
Initial Graphs (Before log transformation):





Final Graphs(After log transformation):





Model 2: Feature Importance

Top 3 Features

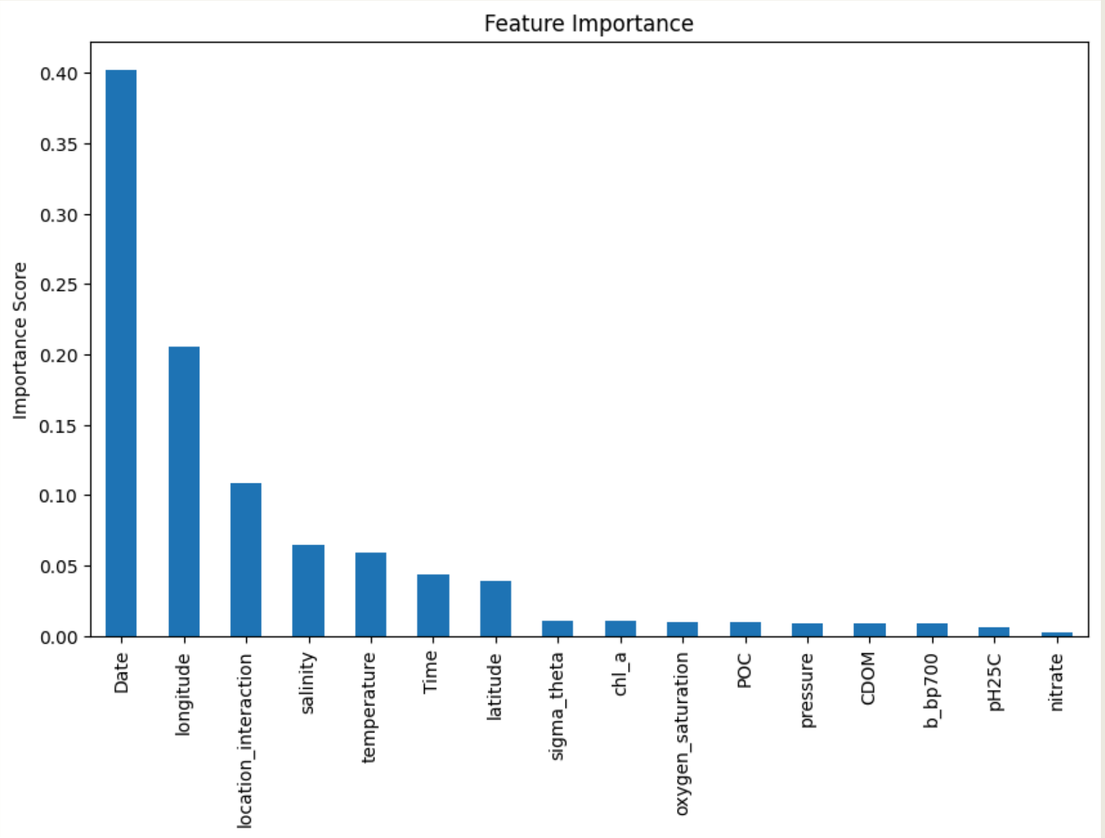
1. Date - Sample Collection Date
2. Simple Longitude
3. Location Interaction - Effects of longitude and latitude combined

Meaningful Predictors

1. Salinity
2. Temperature
3. Location Interaction

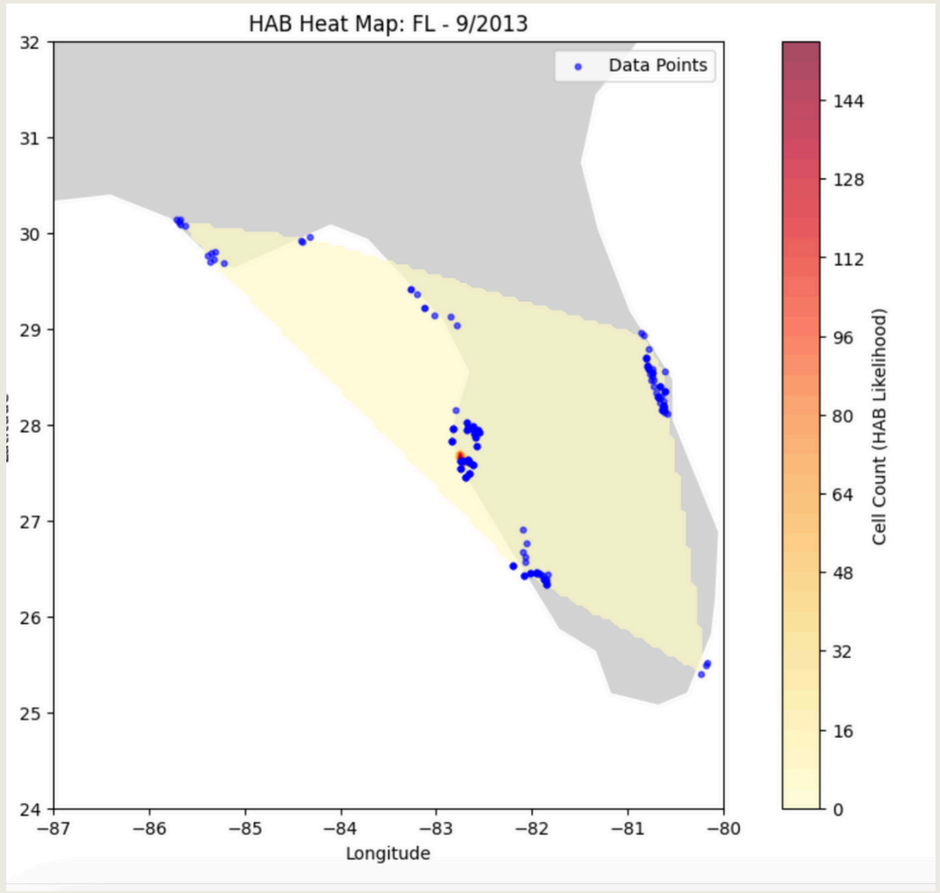
Less Impactful Features

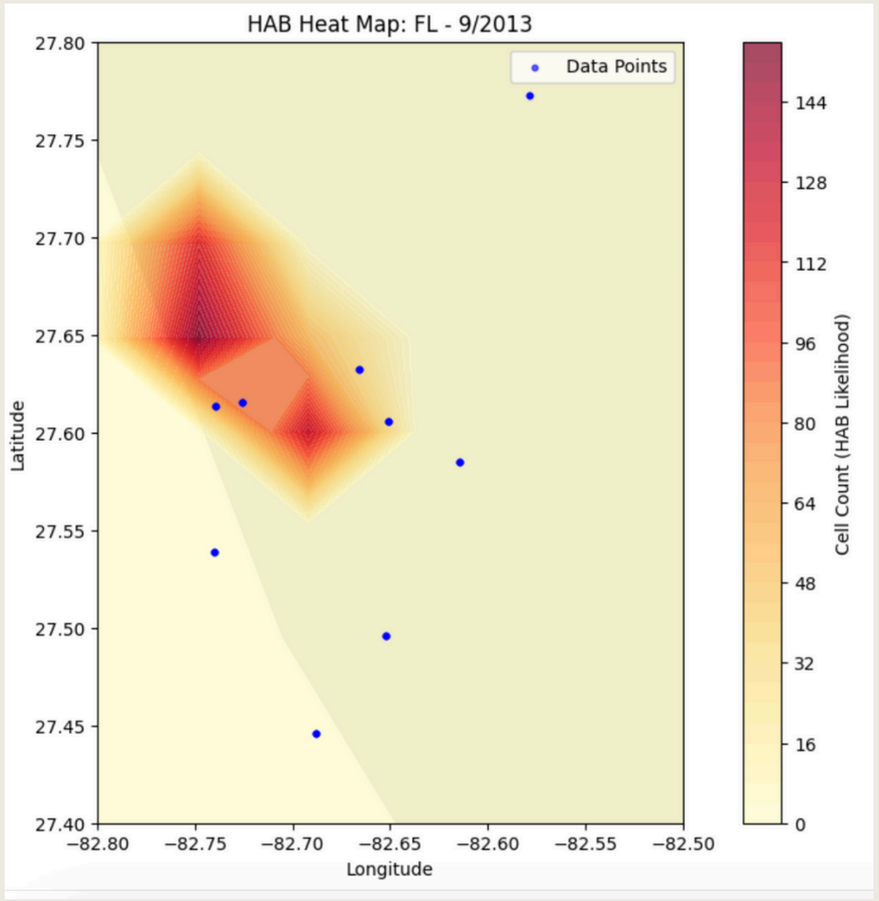
Features like pH25C and nitrate had minimal influence, suggesting potential redundancy



Date and Longitude combined contributed ~60% of feature importance

Heatmaps - Florida





Future Directions

- Improving the accuracy of our model
  - More comprehensive datasets
  - Comparing different types of models
- A web interface for the heatmap
- Usage by Biointerphase as a way to identify where HABs are likely

Questions?