MIDS Capstone Project: **Marine Microplastics**

Final Presentation

Daniel Solomon, Kayla Derman, Nathaniel Browning

Meet The Team

Daniel Solomon

Lead Machine Learning Engineer

Kayla Derman

Lead Data Engineer

Nathaniel Browning

Lead Application Developer

Mission Statement

Our mission is to provide consumers with the knowledge and capabilities to predict their potential exposure to microplastics via seafood. Our goal is to place the power of risk minimization in the hands of consumers through the power of analytics.

The Problem?

As awareness about microplastic pollution rises, consumers and health organizations are becoming increasingly concerned about the potential health risks posed by consuming fish contaminated with microplastics. Studies have shown that microplastics can accumulate in fish tissue, potentially passing through the food chain to humans. However, consumers lack an easy way to assess whether the fish they purchase contains high levels of microplastics. Currently, there is no readily accessible tool that provides this information based on real-time or historical environmental data.

Target Customers

Primary User:
Health Conscious Consumers

Our primary user will be health conscious consumers who are concerned with the safety of the types of fish they eat or fish for.

Primary User:
Individuals Looking to Learn

Our second user will be individuals/consumers who are interested in learning more about how microplastics impact the ecosystem.

Minimal Viable Product

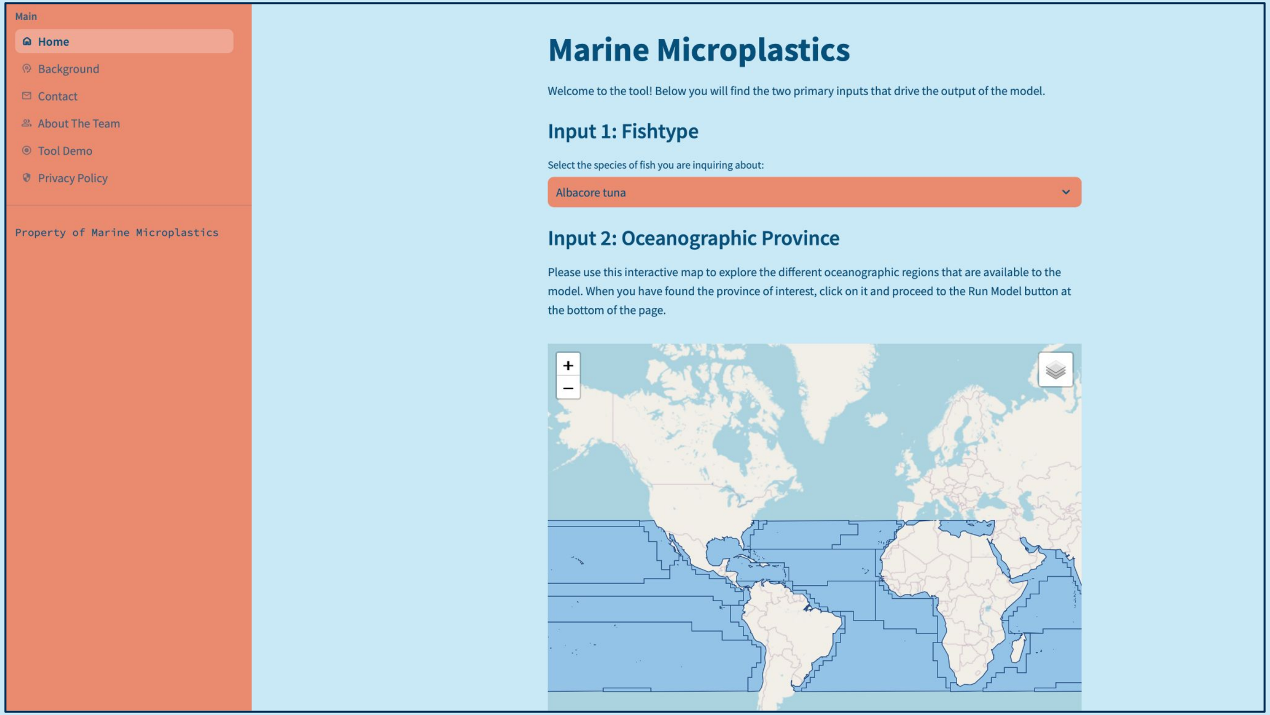
An exploration of the Marine Microplastics tool and our MVP

Minimal Viable Product

Our MVP: A web application that allows users to select from a dropdown of common store bought fish and receive predictions on the microplastic contamination based on satellite data.

1. **Function:** Users will select a fish name from a dropdown along with the general location that the fish would have been caught from (if available).
2. **Processing:** Data retrieval from a backend that uses CYGNSS satellite data to assess regional microplastic contamination.
   a. Our tool will look for correlations between the above satellite data and microplastic contamination levels via predictive models.
3. **Output:** A risk level of microplastic contamination by color (red, yellow, green).

Live Tool Exploration

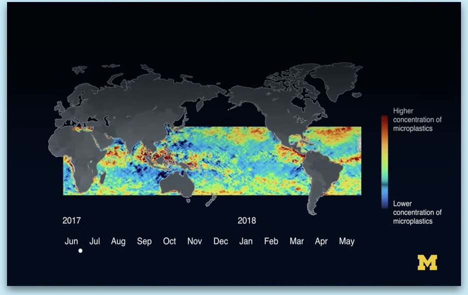


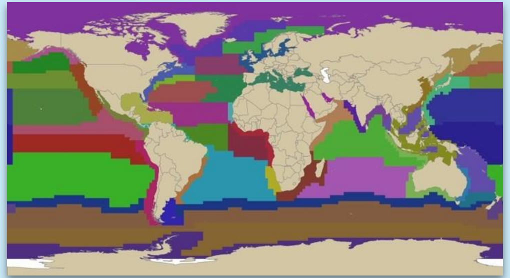
Technical Background

An exploration of the tools modeling and background functions

Datasets & Resources

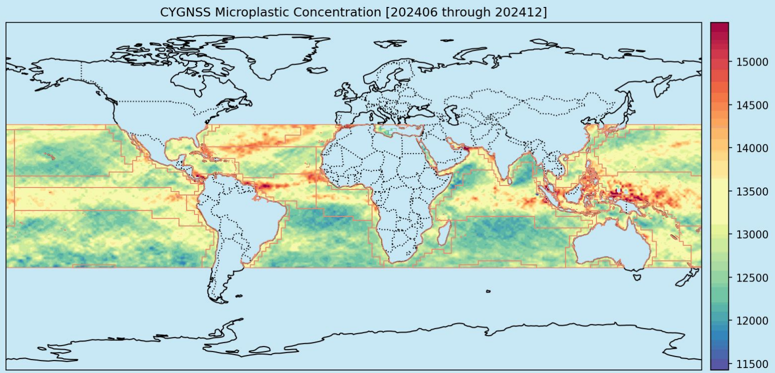
CYGNSS Satellite Data	Public satellite data monitoring microplastic concentration in oceans. Key metadata includes geographic coordinates, microplastic concentration levels, and timestamps.
Savoca Data Set	A large dataset that aggregates a series of studies around microplastic accumulation in fish (used for the training model).
Ecological Geography of the Sea - Longhurst	Textbook by Alan R. Longhurst that identifies and defines oceanic provinces that are used by Savoca dataset.
AquaMaps*	A public dataset mapping the longitude and latitude of thousands of fish species around the globe along with their concentration.





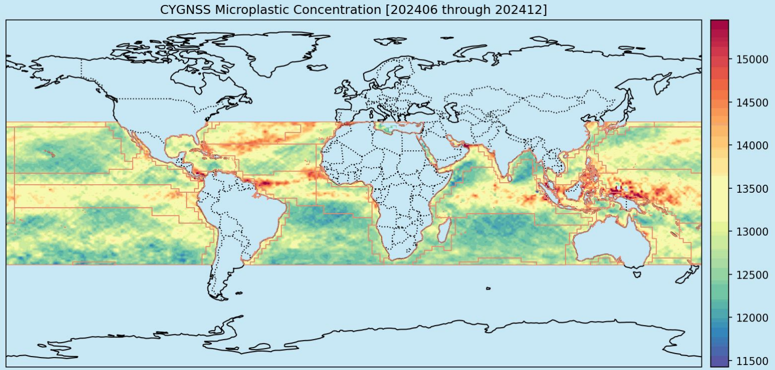
*Indicates a Stretch Goal Resource

Dataset Details



Dataset	Details
CYGNSS Satellite Data	○ 2245 Satellite Image Arrays ○ August 18, 2018 through Present Day ○ -37 through +37 Degree Latitude ○ 360 Degree of Longitude
Savoca Data Set	○ An aggregation of microplastic research results ○ Primary Features ■ Species, trophic level, longhurst province code, mean microplastic counts
Ecological Geography of the Sea - Longhurst	○ Includes 56 geographic ocean regions ○ Encompasses unique biochemical environments

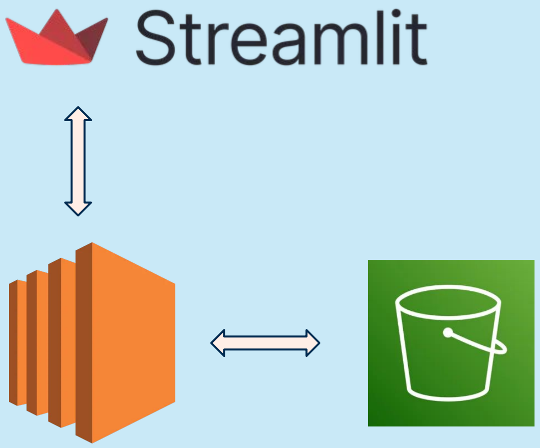
Dataset Details



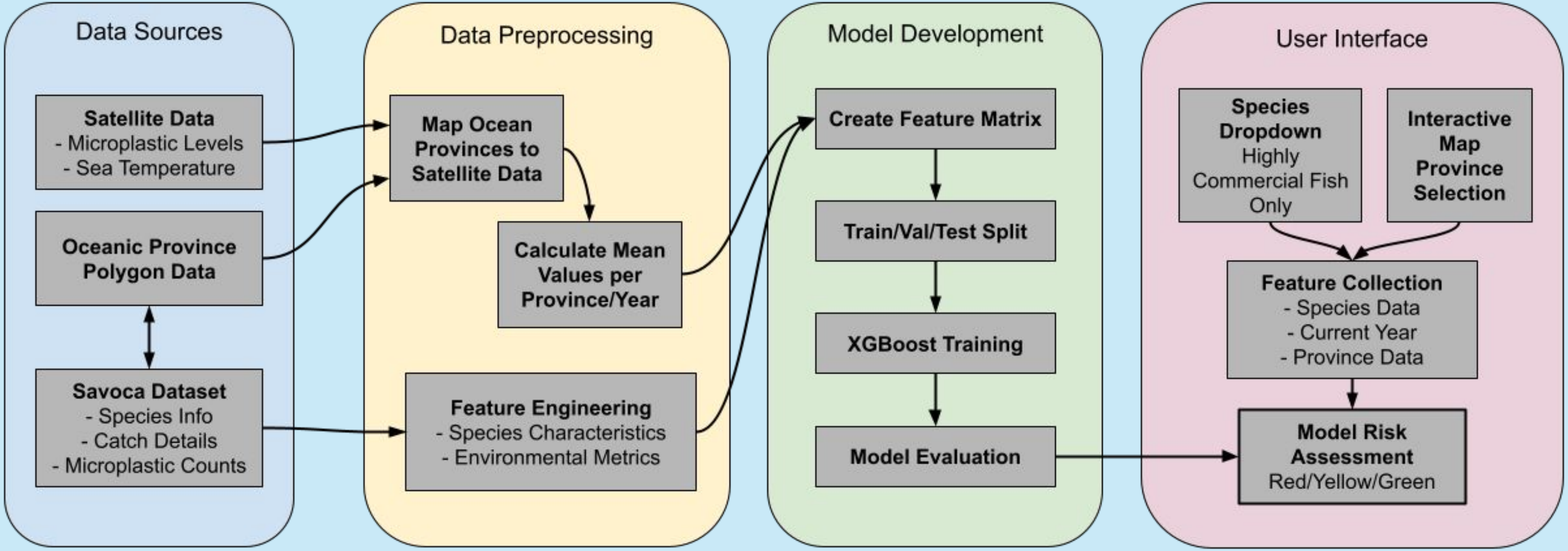
Dataset	Details
CYGNSS Satellite Data	○ 2245 Satellite Image Arrays ○ August 18, 2018 through Present Day ○ -37 through +37 Degree Latitude ○ 360 Degree of Longitude
Savoca Data Set	○ An aggregation of microplastic research results ○ Primary Features ■ Species, trophic level, longhurst province code, mean microplastic counts
Ecological Geography of the Sea - Longhurst	○ Includes 56 geographic ocean regions ○ Encompasses unique biochemical environments

Front-End Workflow

Streamlit	EC2 Instance	S3 Bucket
Streamlit serves as our source for the front-end of our application. We run the model and UI interface through this service.	We utilize an EC2 instance to run the Streamlit web application off of. The EC2 instance is linked to the S3 bucket to pull data.	An S3 bucket with the updated CYGNSS data is accessed through our EC2.



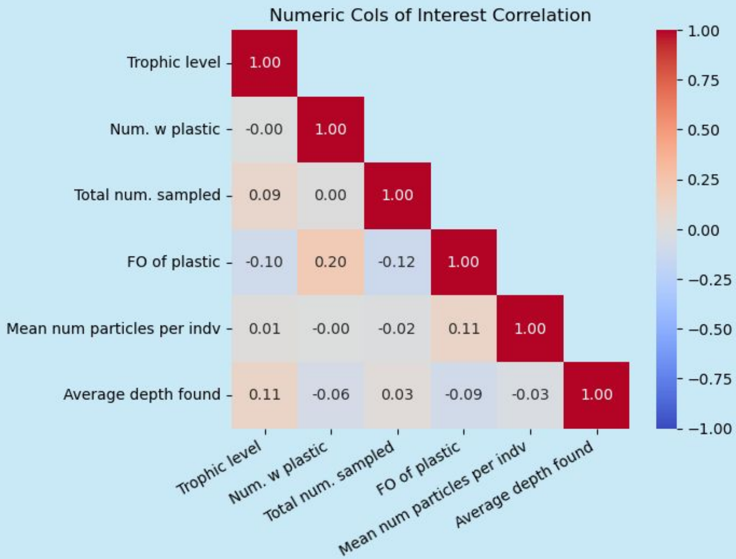
Data Pipeline



Baseline Model Evolution

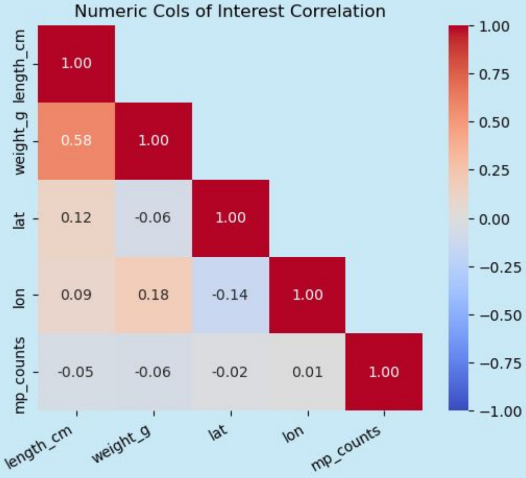
- Linear Regression (correlation matrix)

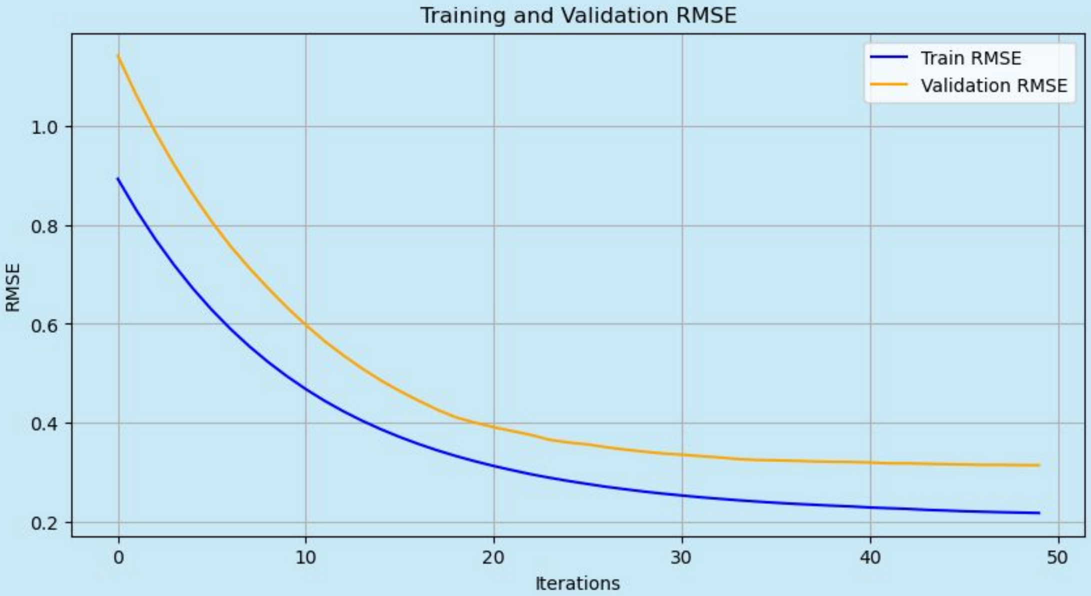
- XGBoost Regression as baseline ML model
  - Initial Inputs
    - Fish Species Name
    - Geographic Coordinate Location
    - Microplastic Counts per fish
  - Input Changes
    - Species Name ➔ Fish Characteristics
    - Geographic ➔ Oceanographic Region
    - Counts per fish ➔ Features by Region



Previous Baseline Model Design

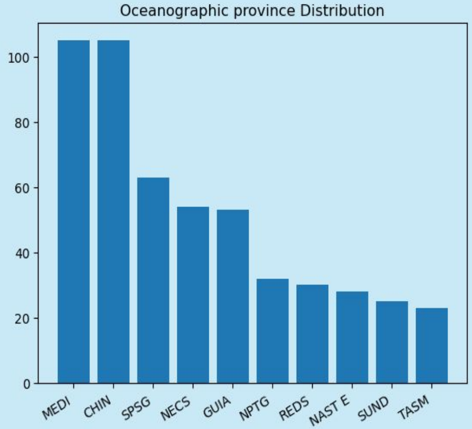
- XGBoost Regression
  - Proven Bioaccumulation modeling
  - Fish Size, Geographic Location, MP Counts
  - Gridsearch optimized
    - LR = 0.1
    - Max Depth = 3

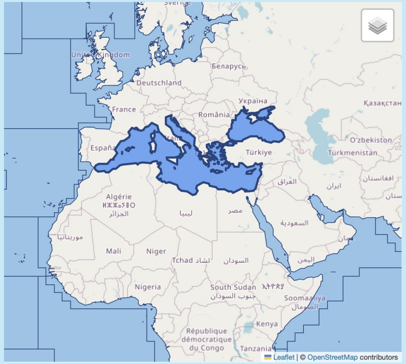


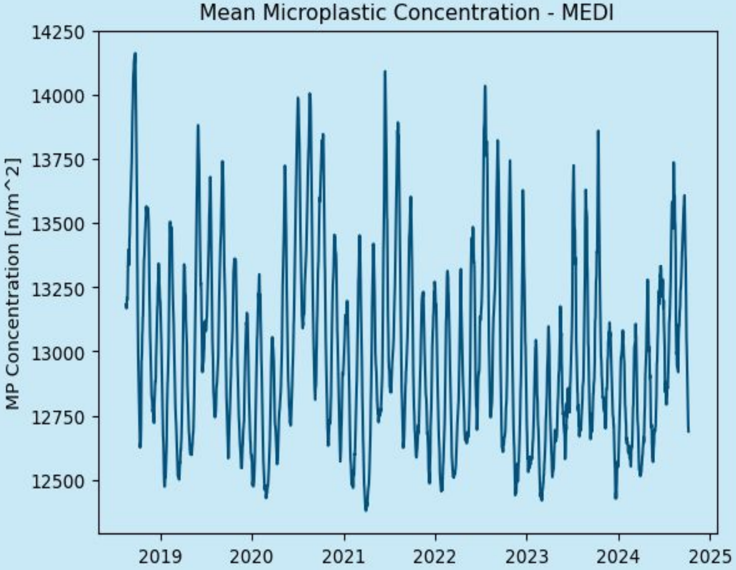


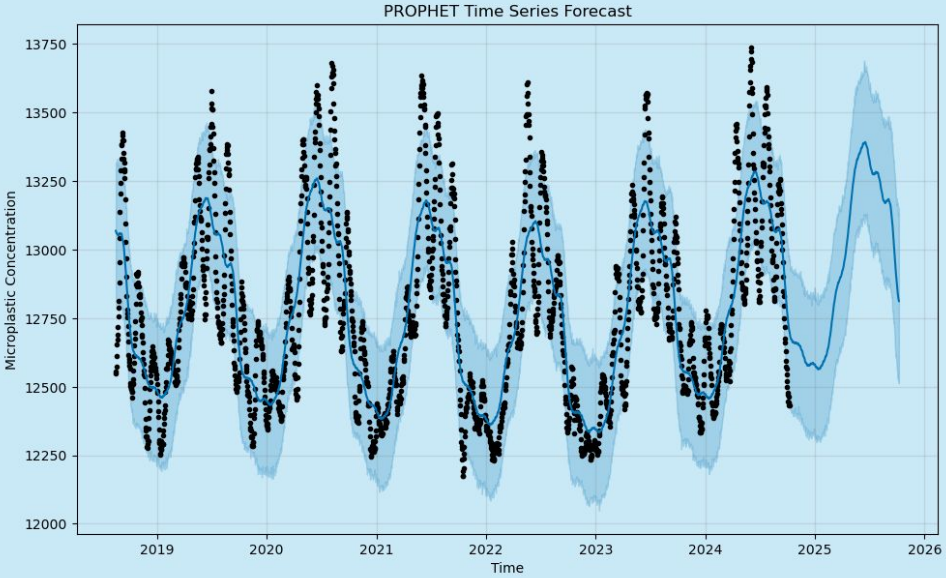
Secondary Modeling

- MP Time-series modeling by province
  - Facebook PROPHET Model
- Optimize for backcasting
  - RMSE, MAE



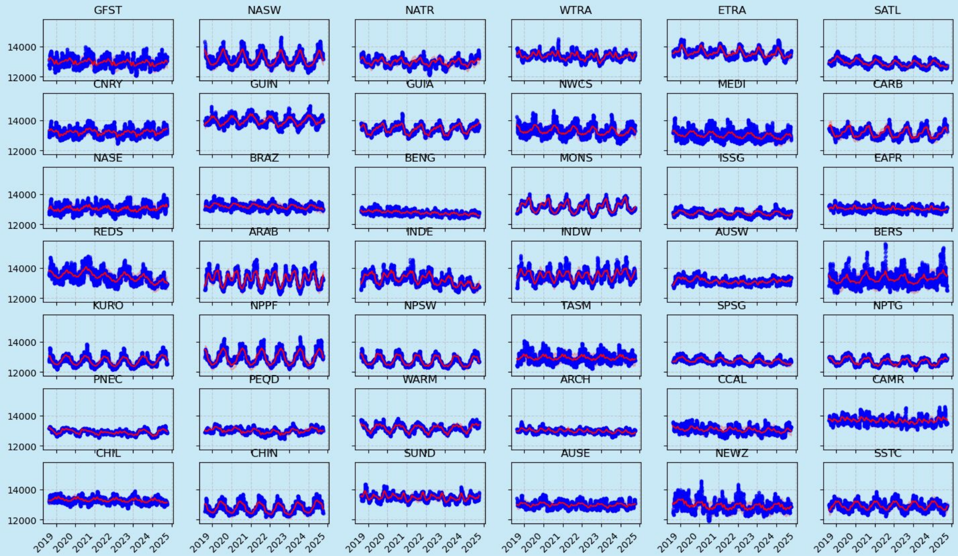




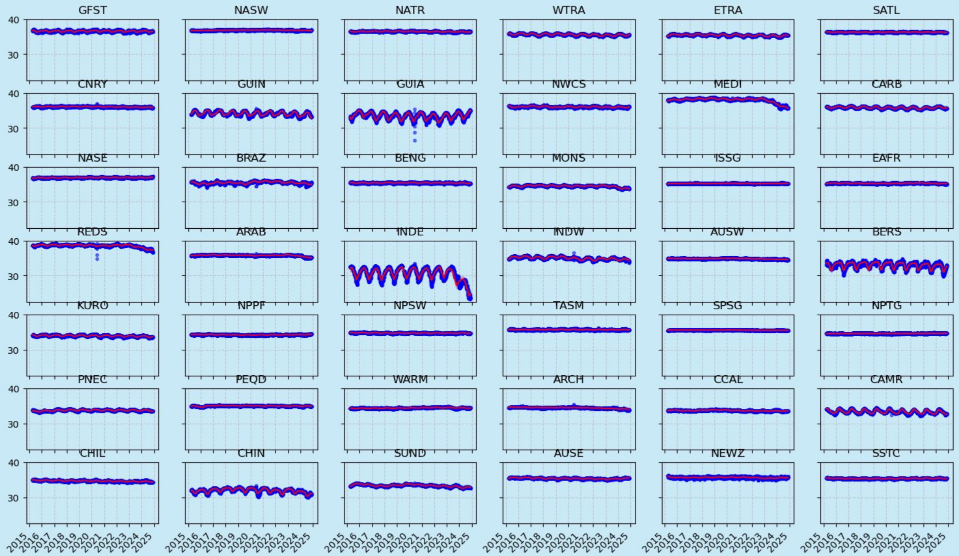


Secondary Modeling Results

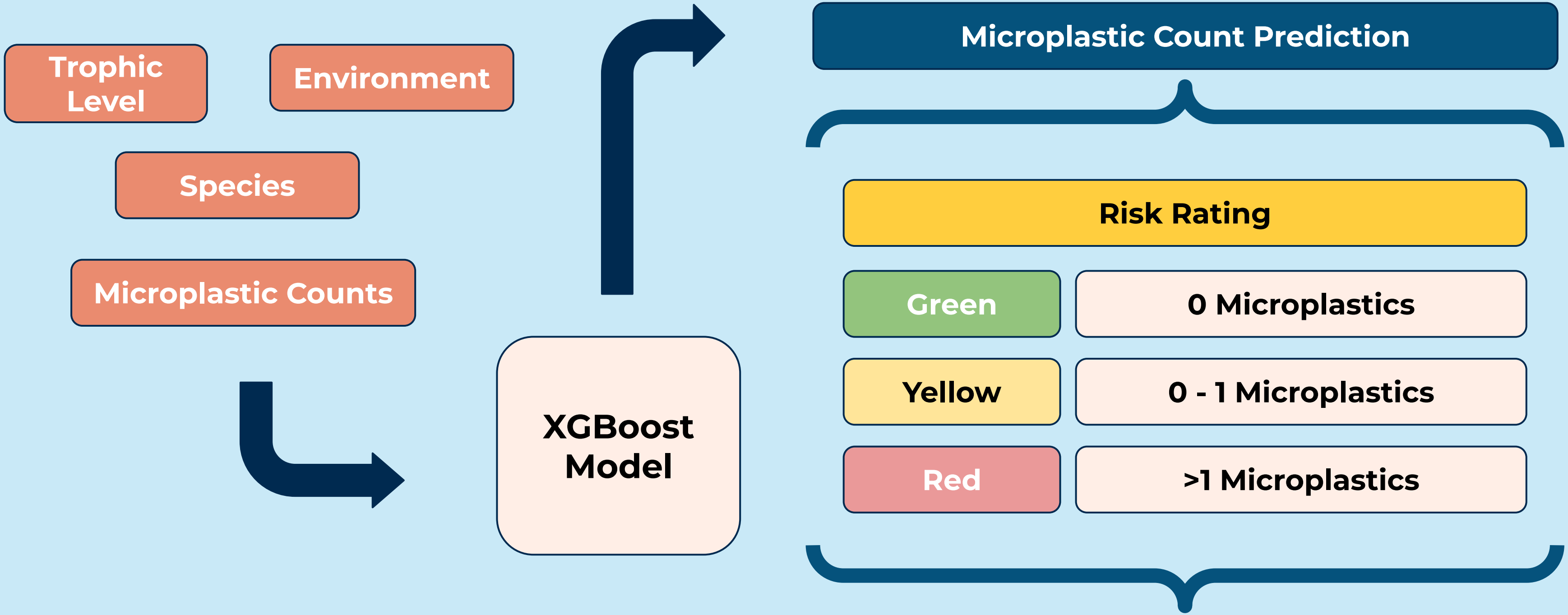
Mean MP Concentration By Province



Mean Sea Surface Salinity By Province



Final Model Architecture



Model Features

Model Inputs	
Fish Common Name	**Province**

- Lookup in Stored Data
  - Trophic Level, Integrated MP Counts, Feeding Pattern, Status as prey, Microplastic Concentration by Location

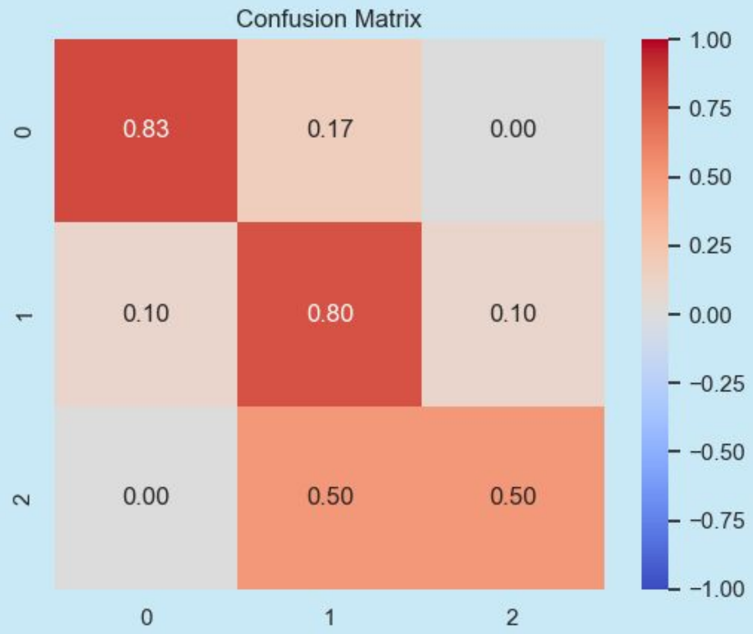
Model Training Variables

	Feature	Importance
1	Integrated Salinity 5yr	0.088932
2	Integrated MPs 2yr	0.087386
3	Integrated Salinity 2yr	0.086357
4	Integrated Salinity 1yr	0.078104
5	Integrated MPs 1yr	0.06587
6	Aquaculture status	0.065556
7	Order	0.052001
8	Integrated MPs 5yr	0.050751
9	Habitat	0.050277
10	Scientific name	0.047132

Technical Model Evaluation

Final Optimized Model Parameters			
gamma	0.1	*estimators*	10
learning rate	0.3	*reg. alpha*	0.1
max depth	7	*reg. lambda*	0.1

Class	Precision	Recall	F1-Score	Support
0	0.71	0.83	0.77	6
1	0.80	0.80	0.80	20
2	0.60	0.50	0.55	6
Accuracy			0.75	32
Macro Avg	0.70	0.71	0.70	32
Weighted Avg	0.75	0.75	0.75	32



Project Takeaways

Challenges	Approach and Takeaways
Microplastic Data Availability	• Utilized a more central study that included data from hundreds of other studies. • The tool will need further data in the future to be as accurate as possible.
Limited Location Data	• The team observed microplastic counts to be a major predictive factor in our model. • Utilizing microplastic concentrations from defined zones as a model input helps us to predict in locations where we won't have training data.
Lack of Commercial Seafood Transparency	• The team made commercial viability a stretch goal that may be possible with stronger commercial data around fishing. • Building a framework to build interest for research/data collection.

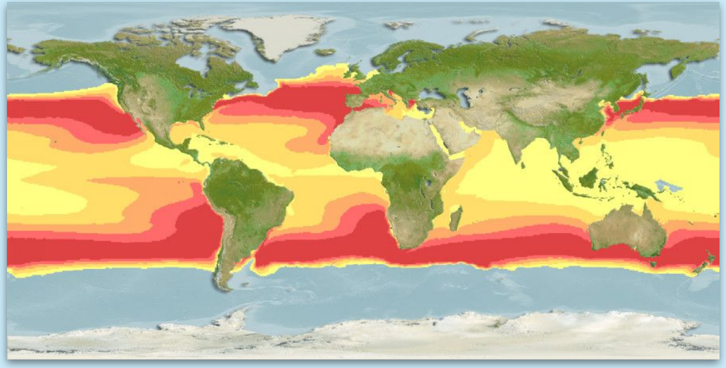
Future Goals

An exploration of potential opportunities
for the team and tool.

Stretch Goals and Roadmap

The following items are reserved for potential future work on the project:

- Implementing the use of AquaMaps data to more accurately pinpoint where fish are to ultimately overlay with satellite data.
- Gather data regarding fish species lifespan and incorporate into calculation of relevant satellite data features such as sea temperature and microplastic exposure.
- Incorporate multi-modal modeling practices. Utilize other models such as a KNN model.
- Incorporate barcodes data for consumer ease of use.



Albacore tuna - Aquamaps

CONCLUSION

As society becomes aware of new challenges it faces regarding our food and health, it is important for the average person to have the needed tools to be informed. Through our project, we hope to provide everyday people with the opportunity to assess their risk with microplastics in the seafood they consume.

Thank You!

Any questions?

Acknowledgements and References

- PoDAAC CYGNSS Microplastics Tracker: https://podaac.jpl.nasa.gov/dataset/CYGNSS_L3_MICROPLASTIC_V1.0
- Matthew Savoca, Researcher Stanford University
- Alan R. Longhurst - Ecological Geography of the Sea
- Todd Holloway and Zona Kostic - W210 Professors
- Thomas Dolan, Marine Life Expert
- W210 Classmates, Section 002