Predicting Oxygen Stressed Conditions in Cape Cod Bay

Julia Weppler, Katharine Baker,
Madie Simmons, Rebecca Trailor

01

Background

What is Hypoxia?

- A state of extremely **low dissolved oxygen concentrations in aquatic environments**
- Poses significant threats to marine ecosystems, fisheries, and tourism

Factors of Hypoxia

Leading contributors:

- Nutrient pollution (nitrogen, phosphorous)
- Climate change (increased temperatures)

Occurs following periods of

- Upwelling
- Algal blooms
- High surface temperatures

Existing Approaches

- High risk of bias in current ML models
- Lack of models for oxygen stressed conditions
- No models for New England watersheds

Ensemble Tree Based Methods	• Harmful algal blooms (Ahn et al., 2023) • Hypoxia in the Gulf of Mexico (Li et al., 2023) • Hypoxia in a small lagoon, integrated with logistic regression (Politikos et al., 2021)
Opportunities for Improvement	• Data accessibility for large number of features • Models are highly specialized to a small area with a high frequency of hypoxic events

Our Goal:
A model which can...

1. Classify a station's data as hypoxic/at high-risk for hypoxia
2. Represent the highly complex relationships of environmental data
3. Provide feature importance for decision-support guiding real actions of scientists

Our Approach:

- Implement **Gradient Boosting** methods combined with SMOTE vs **Logistic Regression**
- Incorporate fewer parameters
- Increase area for training data
- Target oxygen-stressed conditions, not just hypoxia

02
Dataset

Data

- Collected from **24** Water Quality Monitoring Stations in Cape Cod Bay
  - *The Center for Coastal Studies*, Provincetown, MA
- Informations about water temperature, salinity, **dissolved oxygen levels**, chlorophyll

Data

Preprocessing

- Construct dataframe using Pandas lib.
  - Gathers all data from each station
- Remove data missing dissolved oxygen information
- Used mean for other missing entries
- Defined pre-hypoxic conditions – under 7 mg/L of dissolved oxygen

Target

Primary indicator of hypoxia is dissolved oxygen levels

Partition

80% of data used for training, 20% for testing

Features

Dissolved nitrogen levels, particulate organic nitrogen levels, total nitrogen levels, total dissolved, dissolved phosphorus levels, and total ammonium levels

03

Model Selection

Logistic Regression vs eXtreme Gradient Boosting

Both

- Work well for small but structured datasets of numerical features
- Have highly interpretable results

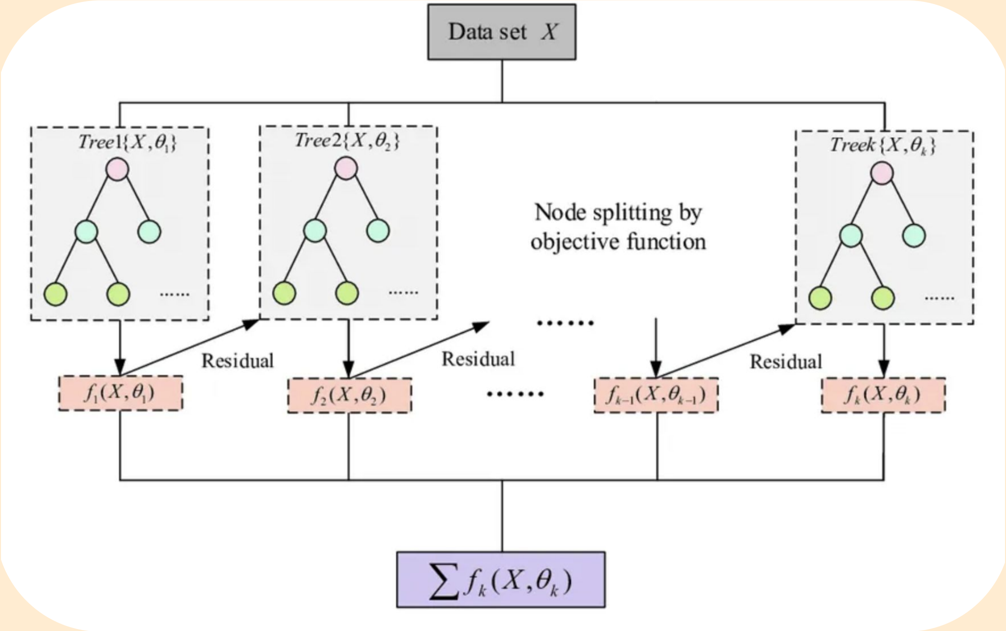
Logistic Regression

- Simple yet reliable/trusted model within the environmental science field

XGBoost

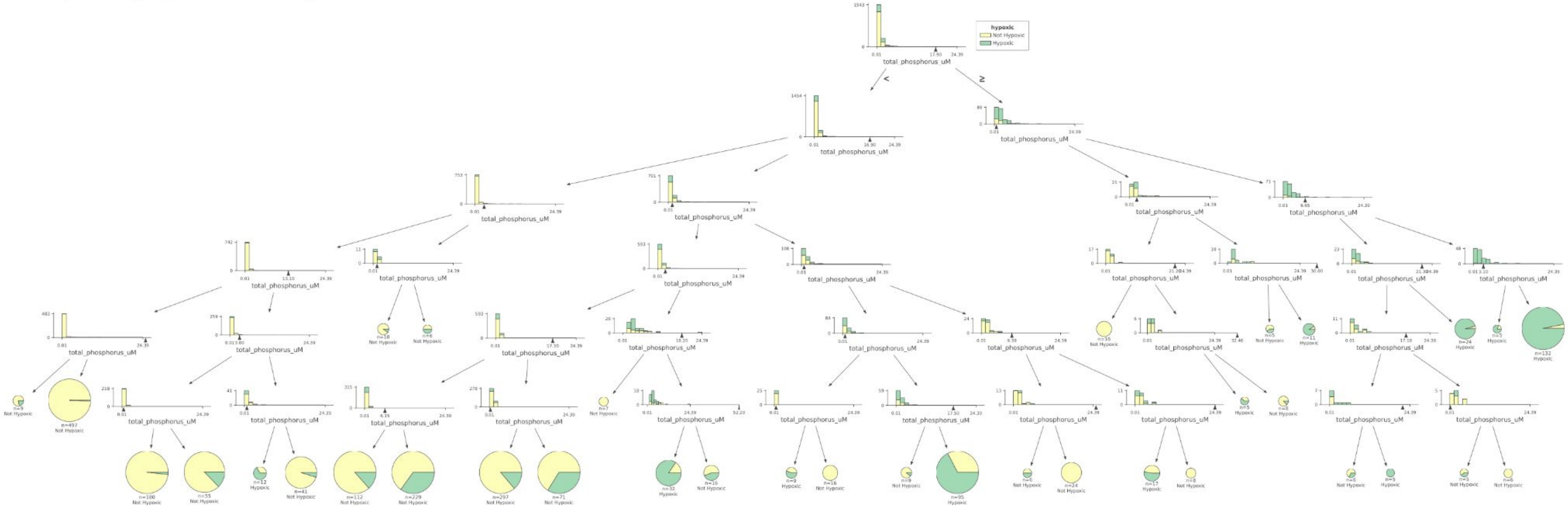
- Can handle complex, non-linear relationships common for environmental data

- Initialization
- Iterative improvement
- Gradient descent step
- Update model with learning rate
- Regularization



What is XGBoost (Classifier)?

XGBoost – Decision Tree Operation

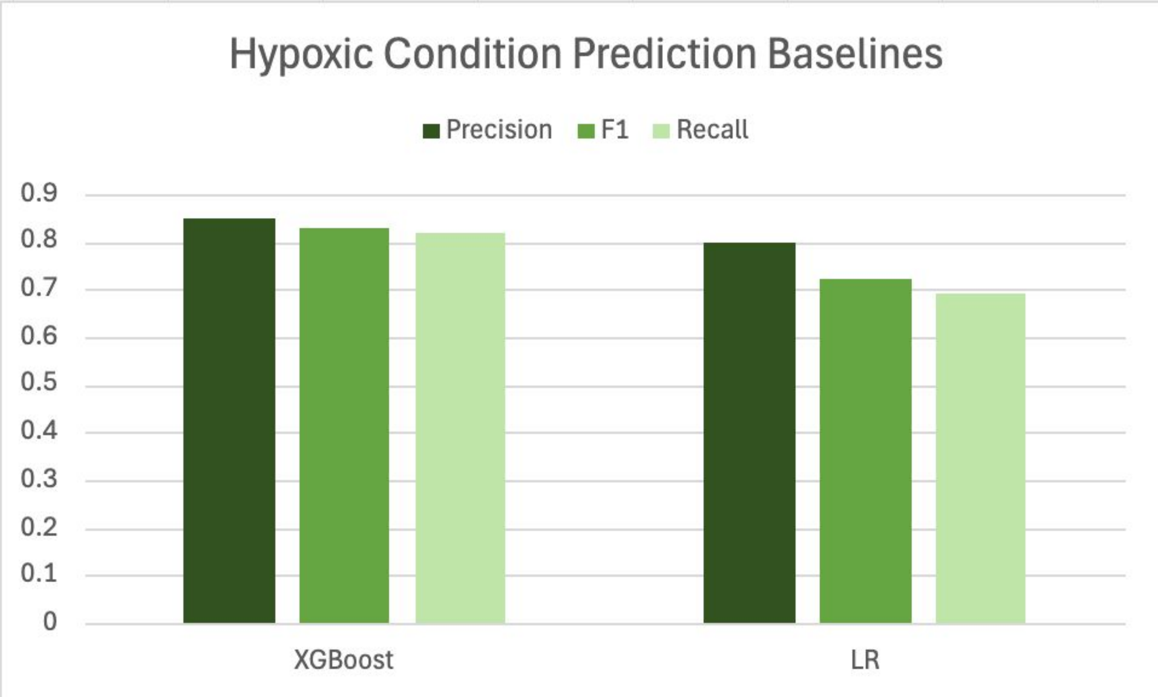


04

Methods

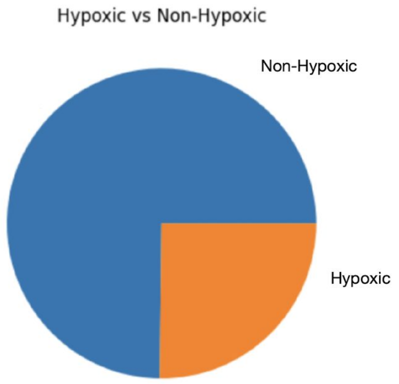
Baseline Models

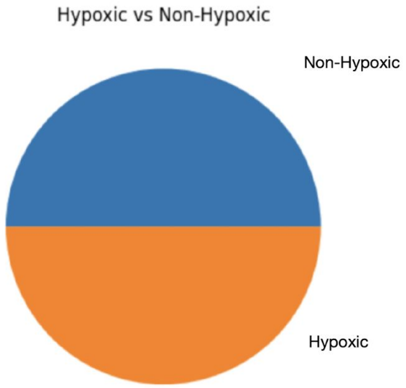
- XGBoost
- Logistic Regression



1. Implementing SMOTE

- Imbalances in data regarding hypoxic conditions
- SMOTE balances out dataset
    - Creates synthetic minority classes, preventing bias in the model's selection
    - imblearn.over_sampling library





2. Hyperparameter Tuning

Used RandomizedSearchCV to

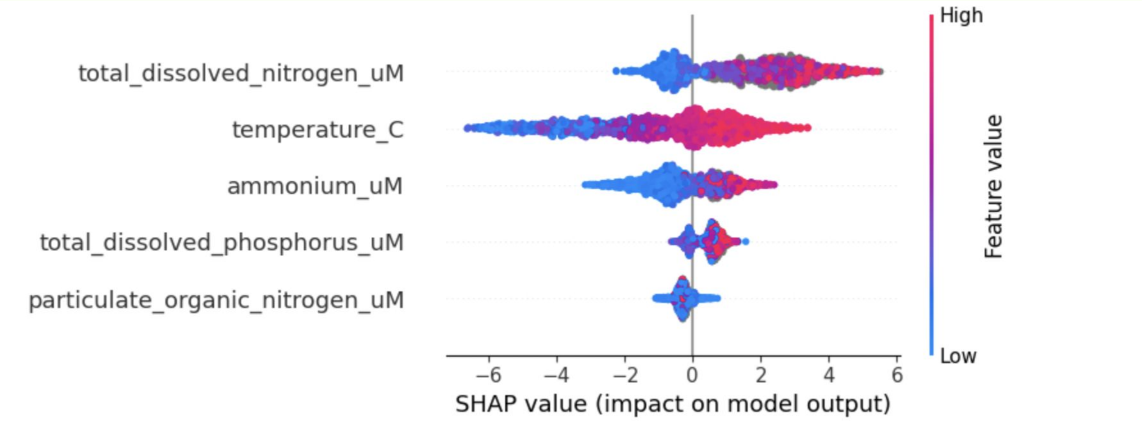
1. Randomly draw one hyperparameter combination from our distributions
2. Train estimator with those hyperparameters on each fold of your cross-validation split
3. Score the held-out fold according to scoring metric of roc_auc
4. Repeat 30 times

```
Best hyperparameters:
{'xgb__subsample': 0.8,
'xgb__scale_pos_weight':
np.float64(2.9728),
'xgb__reg_lambda': 10,
'xgb__reg_alpha': 0.1,
'xgb__n_estimators':
1000, 'xgb__max_depth':
8, 'xgb__learning_rate':
0.2, 'xgb__gamma': 0,
'xgb__colsample_bytree':
1.0}
```

05

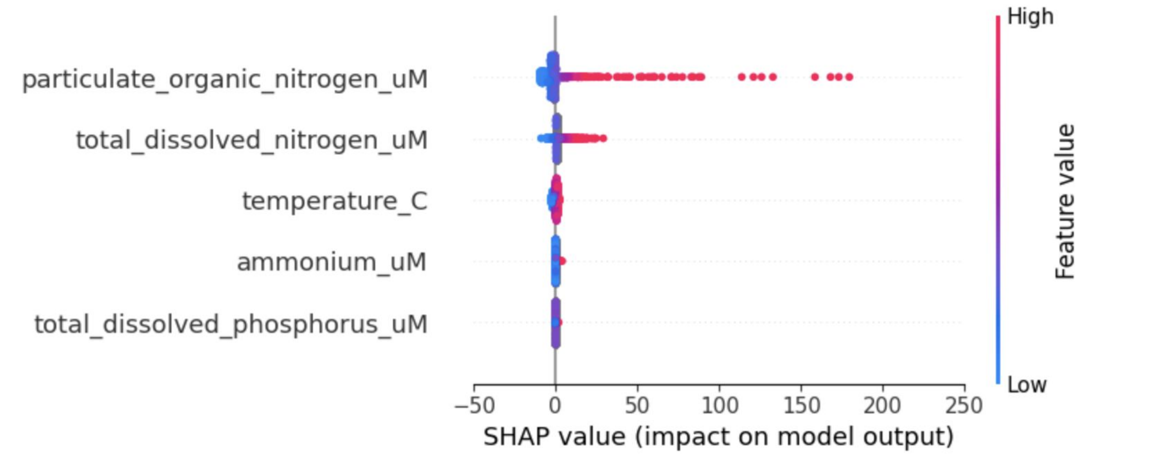
Results

Feature Importance and Impact



XGBoost

Feature Importance and Impact



Logistic Regression

AUROC

XGBoost

- Baseline model: **0.901**
- With SMOTE: **0.906**
- With SMOTE and RandomSearchCV: **0.913**

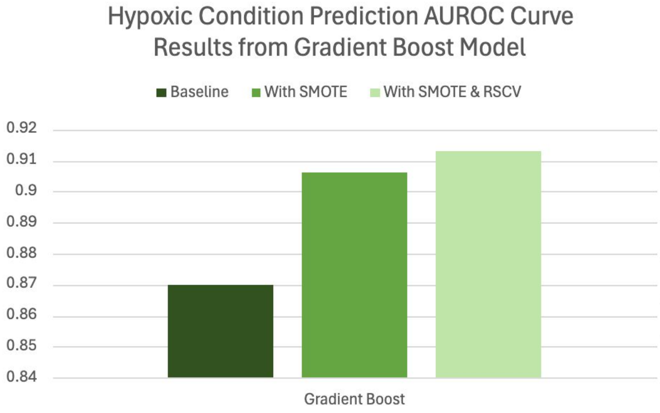
LR

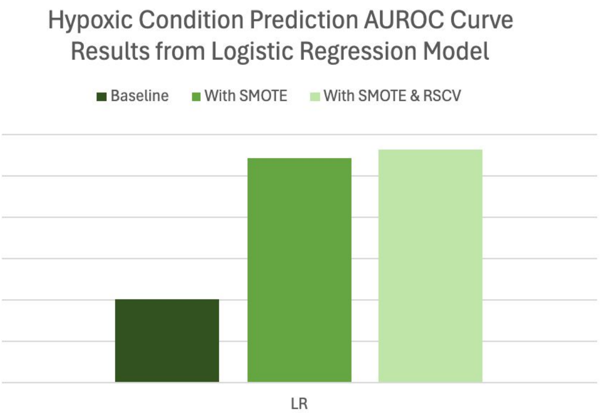
- Baseline model: **0.816**
- With SMOTE: **0.817**
- With SMOTE and RandomSearchCV: **0.818**

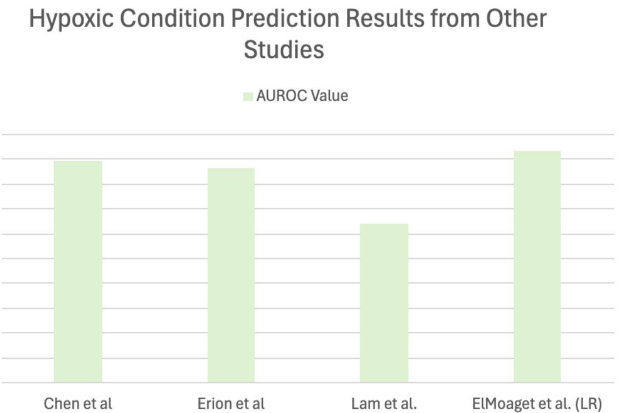
Other Models

- Chen et al., 2021: **0.89**
- Erion et al., 2017: **0.86**
- Lam et al., 2022: **0.64**
- ElMoaqet et al., 2014 (Linear regression): **0.93**

(Pigat et al., 2024)







06
Conclusion

Conclusions

Impact to existing models

Our model yielded a **comparable AUROC value** to other highly-localized studies with a greater number of features (Pigat et al., 2024)

Practical implications

Regions with less resources for data collection can utilize this approach to forecast oxygen-stressed conditions, which could harm fisheries and lead to hypoxia. **Earlier intervention at reduced costs.**

Conclusions

Future work

Further tailoring models to **different watersheds** to continue efforts of prevention and restoration of hypoxic waters.

Summarizing our findings

As expected, XGBoost outperformed our logistic regression model

Questions?

THANK YOU!!

TEAM CONTRIBUTIONS

Rebecca Trailor

Researching datasets and collecting data; initial data preprocessing and XGBoost baseline; organizing materials in Github; writeup work; cross-validation implementation

Madie Simmons

Researching datasets and collecting data; fine-tuning LR model and analyzing/interpreting summary statistic results and their impact on next steps; writeup work; XGBoost visualizations

Katharine Baker

Researching datasets and collecting data, building LR baseline model, implementing SMOTE, organize presentation slides, outline data collection, data preprocessing, XGBoost, writeup work

Julia Weppler

Project conception, background, data identification, RandomizedSearchCV for hyperparameter tuning

https://github.com/trailorr/MLProject

References

Pigat, L., Geisler, B. P., Sheikhalishahi, S., Sander, J., Kaspar, M., Schmutz, M., Rohr, S. O., Wild, C. M., Goss, S., Zaghdoudi, S., & Hinske, L. C. (2024). Predicting Hypoxia Using Machine Learning: Systematic Review. JMIR medical informatics, 12, e50642. https://doi.org/10.2196/50642

Elmoaqet, H., Tilbury, D. M., & Ramachandran, S. K. (2014). Evaluating predictions of critical oxygen desaturation events. *Physiological measurement*, *35*(4), 639–655. https://doi.org/10.1088/0967-3334/35/4/639

Chen, H., Lundberg, S. M., Erion, G., Kim, J. H., & Lee, S. I. (2021). Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *NPJ digital medicine*, *4*(1), 167. https://doi.org/10.1038/s41746-021-00536-y

Lam, C., Thapa, R., Maharjan, J., Rahmani, K., Tso, C. F., Singh, N. P., Casie Chetty, S., & Mao, Q. (2022). Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study. *JMIR medical informatics*, *10*(6), e36202. https://doi.org/10.2196/36202

Erion, G.G., Chen, H., Lundberg, S.M., & Lee, S. (2017). Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning. *ArXiv, abs/1712.00563*.