Bias and Fairness

In predicting severity of cyanobacteria blooms

by Fiona Chow and Jennah Gosciak

Data: Inputs

Metadata

23,570 rows of training and test data with latitude/longitude and date

Elevation Data

Copernicus Digital Elevation Model (DEM) with 30-meter resolution with data such as maximum elevation and difference in elevation

Satellite Data

Sentinel-2 Level-2A satellite imagery from the European Commission in partnership with the European Space Agency (ESA) with data such as spectral bands – red, blue and green

Auxiliary Data: US Census

American Community Survey (ACS) 2015-2019 5-year estimates at the census tract level with data such as race and ethnicity, median household income, and poverty rate

Process of Sampling Data

Because test data do not have labels, we conduct our audit using a sample of training data.

1. Filtered for training Data from 2017
2. Sample so as to ensure proportional levels in each region to the original data

   Created 72 / 28% split of training and test data
3. Split the new training data (the 72%) further into a training and validation set

Outputs

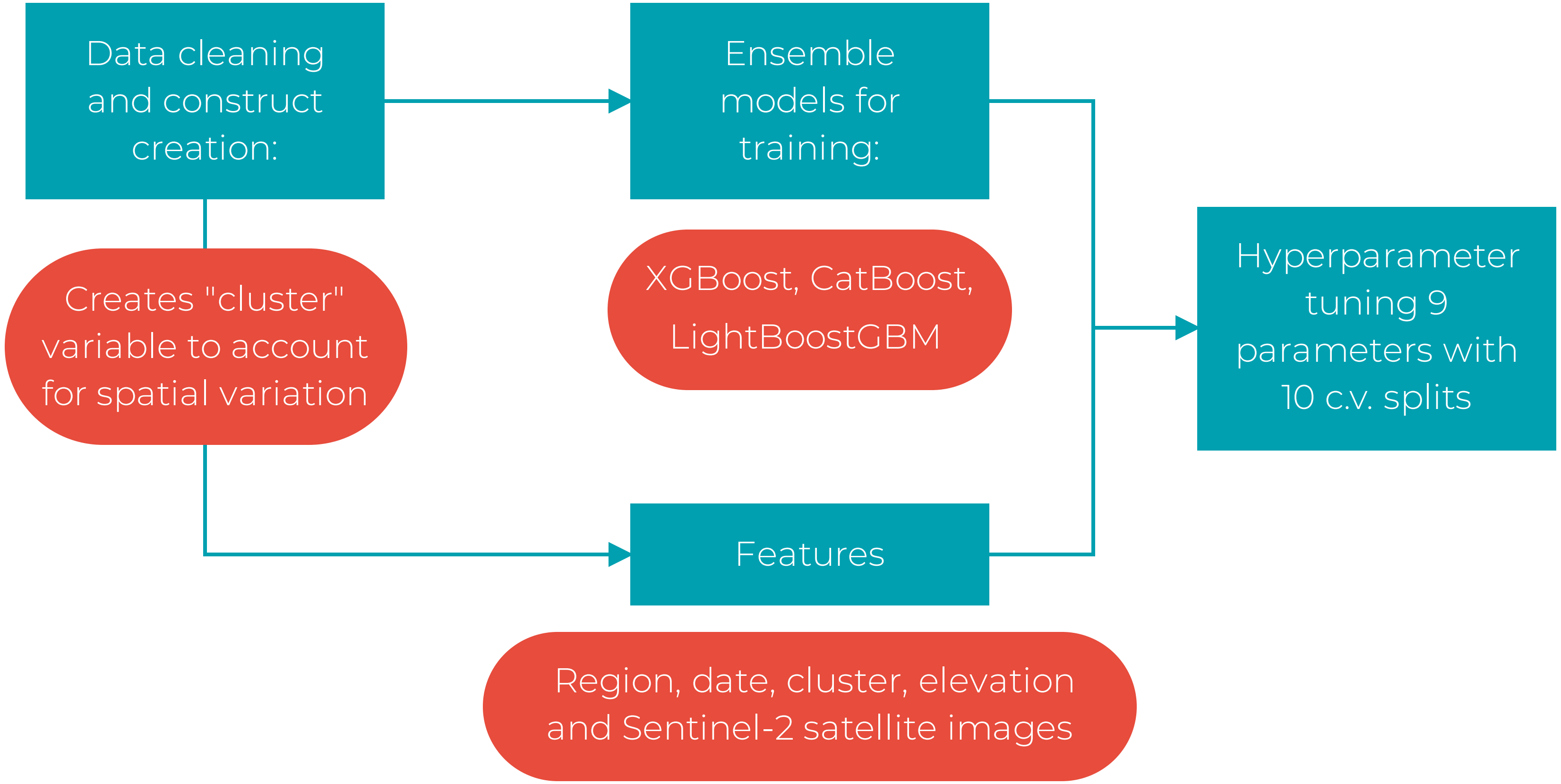
Severity	Density range (cells/mL)
1	< 20,000
2	20,000 - < 100,000
3	100,000 - < 1,000,000
4	1,00,000 - < 10,000,000
5	> 10,000,000

Binary classification

Low risk

High risk

Implementation



Validation: Original Competition

Region-averaged root mean squared error (RMSE) using the estimated and observed severity values for each region

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(Y_i - \hat{Y}_i)^2}$$

$$\Rightarrow \frac{RMSE_{Midwest} + RMSE_{West} + RMSE_{South} + RMSE_{Northeast}}{4}$$

- Y_i = true severity level for the $|i^{\text{th}}|$ sample
- $\hat{Y}_i$ = predicted severity level for the $|i^{\text{th}}|$ sample
- N = total number of samples

DrivenData leaderboard performance: 0.76

Performance on sampled data: 0.73

Subpopulations

Income characteristics

High poverty and low income areas

Above the statewide average

Below the statewide average

Population characteristics

Percent non-white and Hispanic or Latino

Performance

With only low and high severity labels, we use common metrics to evaluate the ADS like accuracy, precision, recall, and false negative and false positive rates

	Baseline	Overall
Accuracy	0.69	0.81
Precision	0.69	0.82
Recall	1	0.93
False Negative Rate	0	0.07
False Positive Rate	1	0.48

Evaluation of Performance Overall in Comparison to the Baseline Classifier

- Define a baseline classifier that always predicts the majority class
- Evaluate the model on an unseen test set for a range of metrics
- Identify high-priority metrics and compare performance to baseline

Aequitas Fairness Metrics

1 Why these fairness metrics?

FNR Disparity: ensure ADS does not fail to provide assistance to protected subgroups

Recall Disparity: ensure the results of the ADS is distributed in a representative way

	Poverty Rate	Low Income	Non-White	Hispanic/Latino
FNR Disparity	1.12	0.38	1.83	0.26
Recall Disparity	0.99	1.07	0.95	1.07
Predicted Prevalence Disparity	1.09	1.18	0.93	1.19

2 Results:

Recall / Predicted Prevalence parity across all subgroups

FNR disparity for the non-White subgroup is 1.83 and falls outside of the rule-of-thumb for fairness of $0.8 \le$ disparity ≤ 1.25

Interpretability: Feature Importance

1. Longitude
2. Cluster
3. Elevation
4. Month

Insight:

Primary focus of the ADS solution is on fitting a curve based on temporal and spatial variations; May not generalize well if underlying temporal and spatial patterns change

Summary

Data

Standard and **appropriate dataset** used on other contexts (e.g., computer vision for algae bloom prediction in oceans)

Performance

High recall and **precision**

Few disparities across subgroups

Deployment

Deployment in the **public sector** with **human oversight**

Recommendations

Collect more data from the **Northeast**

Better account for spatial variation