Predicting Blue-Green Algae Quantity in a Semi-Urban Reservoir

Nathan Jahns, M.Sc.

Data Science Intensive Capstone Project, Sept 23 Cohort

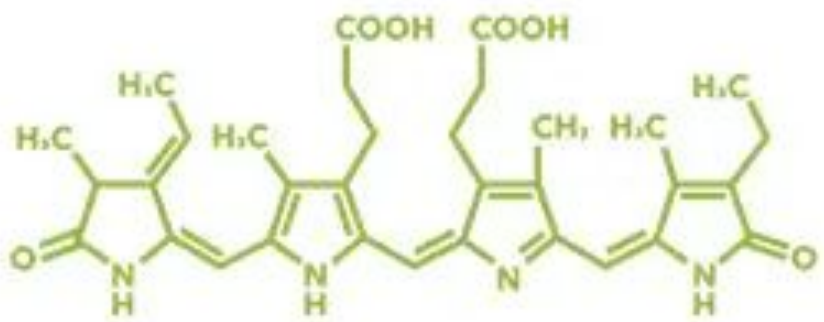
Springboard

Background: Blue-Green Algae

- Harmful algae blooms (HABs), mostly blue-green algae, in lakes across USA every summer and can grow over a few days
- Can release toxins which kill fish, waterfowl, livestock, and pets and make humans very ill
- Measure phycocyanin (pigment in blue-green algae) with *in situ* probe
- Lake operators monitor readings and shut down recreation
    - Immediate, no warning to the public or staff, management actions are reactionary
    - Method to predict blue-green algae bloom before a shutdown would allow for planning and allocation of resources

Objective

- Develop machine learning models to predict phycocyanin using water quality measurements at a semi-urban reservoir I named Lake Fictitious
- Establish models capable of forecasting phycocyanin multiple days in advance using the explanatory variables collected at the reservoir
- With this model and data the magnitude of a HAB could be predicted before it occurs which would avoid immediate shutdowns



Phycocyanin

Data Acquisition

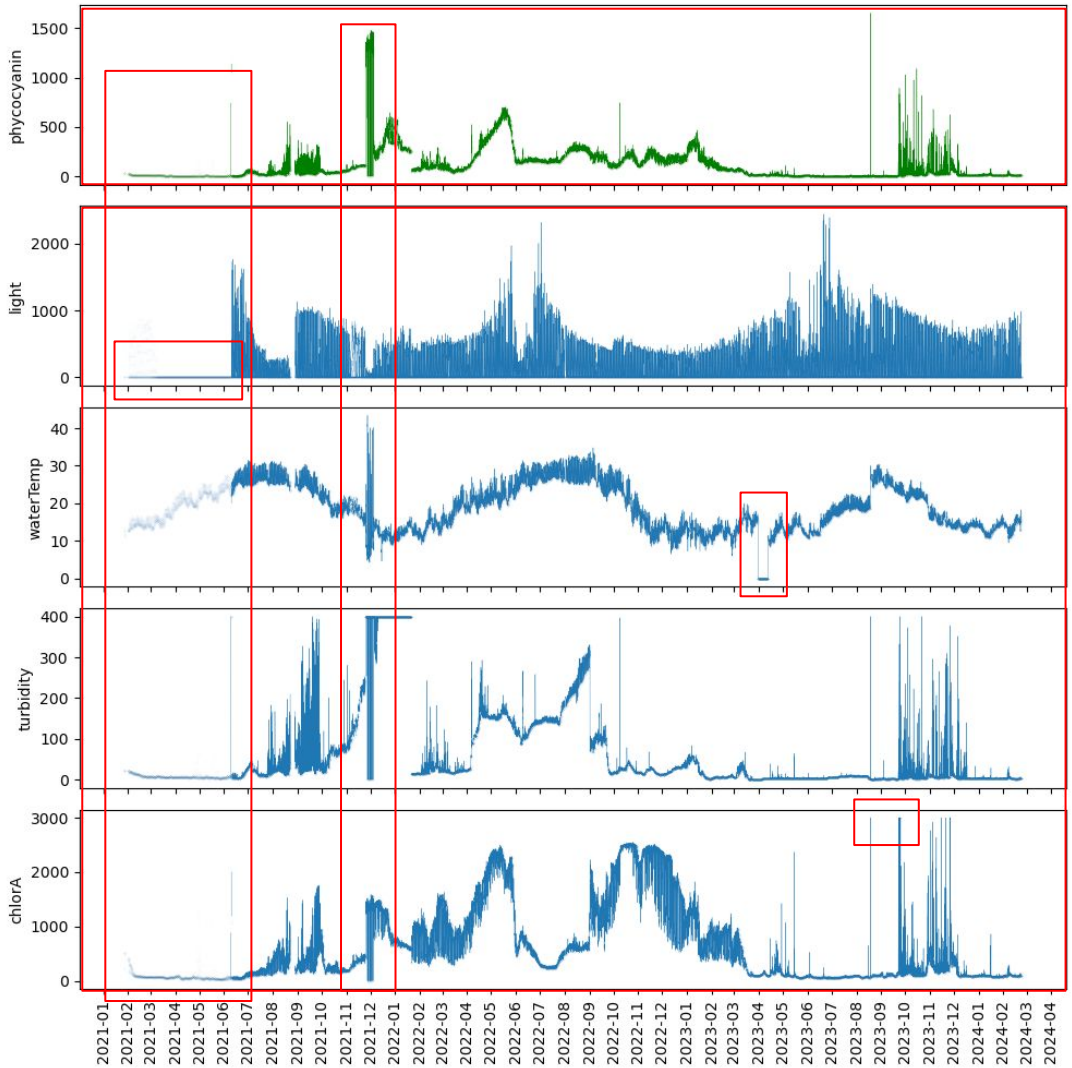
- Data collected at Lake Fictitious every 30 minutes for three years
- *In situ* AlgaeTracker manufactured by AquaRealTime
- 50,071 rows and 20 columns

Data Cleaning

- Removed unneeded columns
  - Phycocyanin, sunlight, water temperature, water turbidity, and chlorophyll-a
- Removed duplicate date and time
  - Calculated mean because column values for pairs were often different
  - Both forward and backward directions from each NaN value
- Inserted missing half hour rows
  - Largest gaps was 6.25 days, not too large to impute

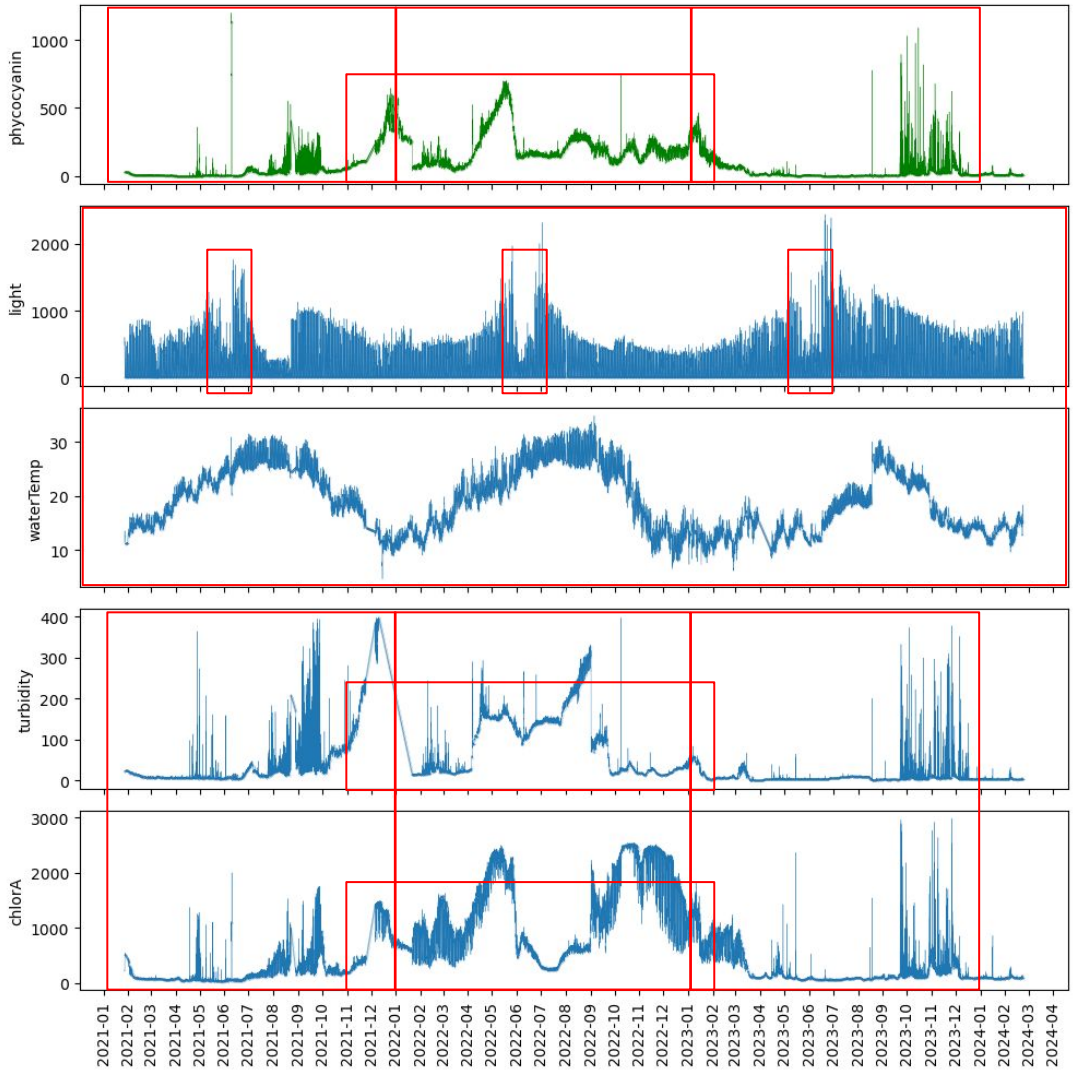
Data Cleaning

- Unreliable data
  - Readings were collected every hour
  - Low sunlight readings
  - Unit flipped upside
  - Water temperature unlikely to be 0 ℃.
  - Maximums also unlikely.
- Replaced with NaNs
- Sunlight - replaced NaNs with mean 2022 and 2023 value
- Spline linear interpolation to impute elsewhere



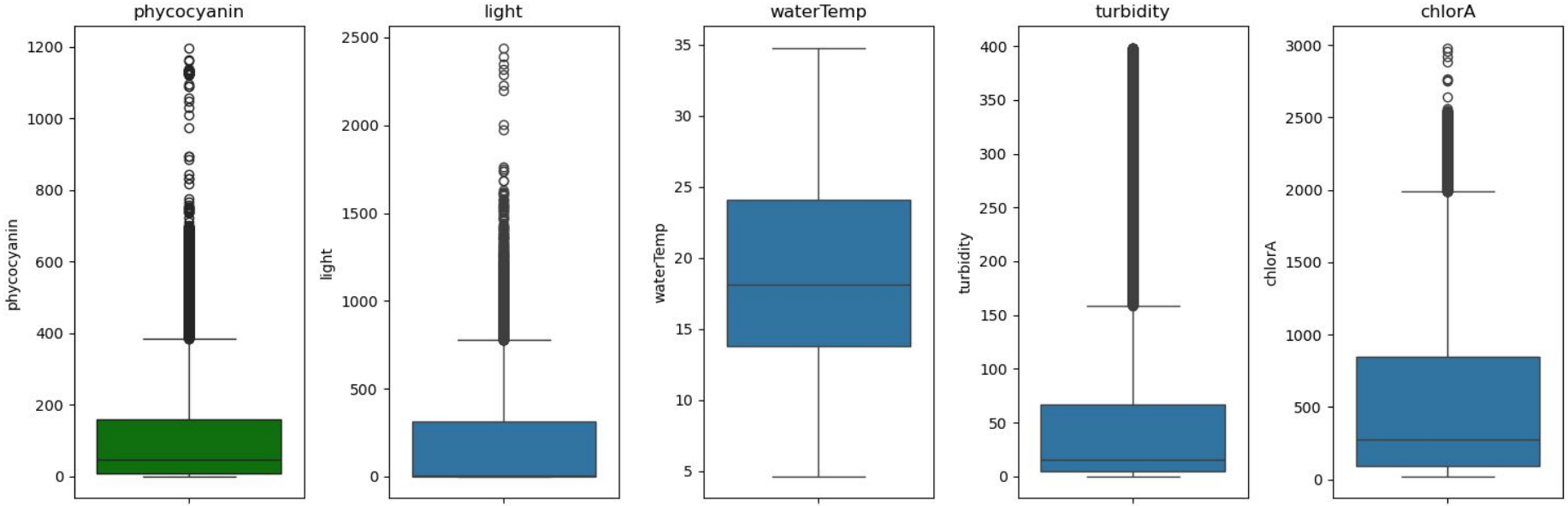
EDA: Trends & Patterns

- Sunlight and water temperature were seasonal
- June was cloudy each year and 2021 was also cloudy in July, Aug, and Dec
- Phycocyanin, turbidity, and chlorophyll-a not seasonal and three years were different
- Baseline was higher and readings often did not return to baseline for all metrics from 11/21-2/23



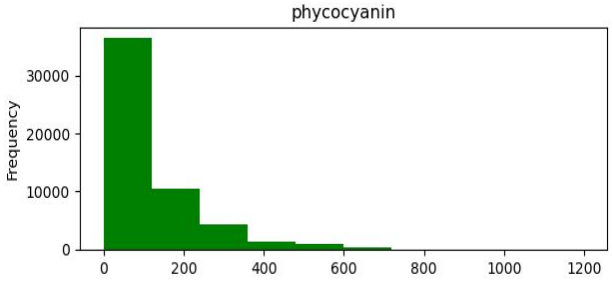
EDA: Distribution

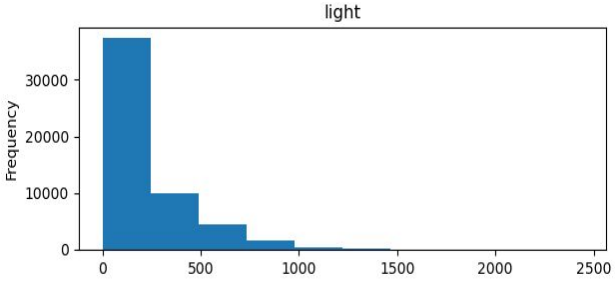
- Water temperature fell within the whiskers (1.5 times the IQR)
- Other values greater than 3 std from the median were not removed because no reason to believe these are erroneous.

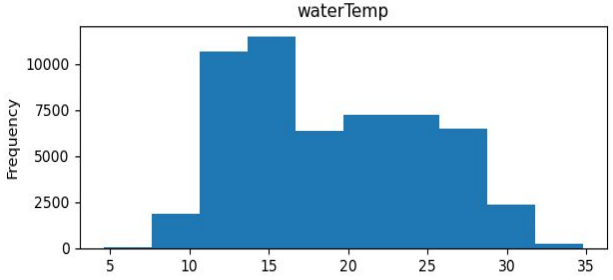


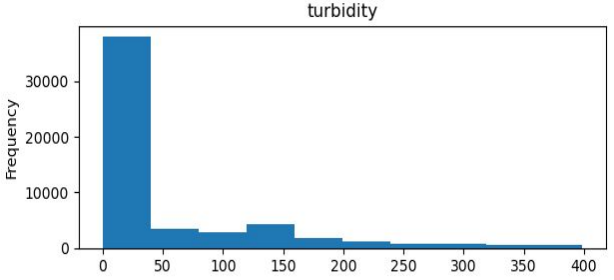
EDA: Distribution

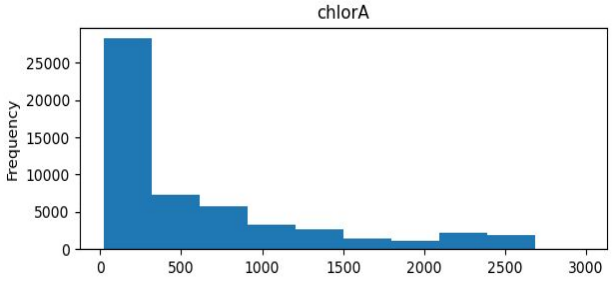
- Reflec boxplots
- Not normally distributed
- Right-skewed





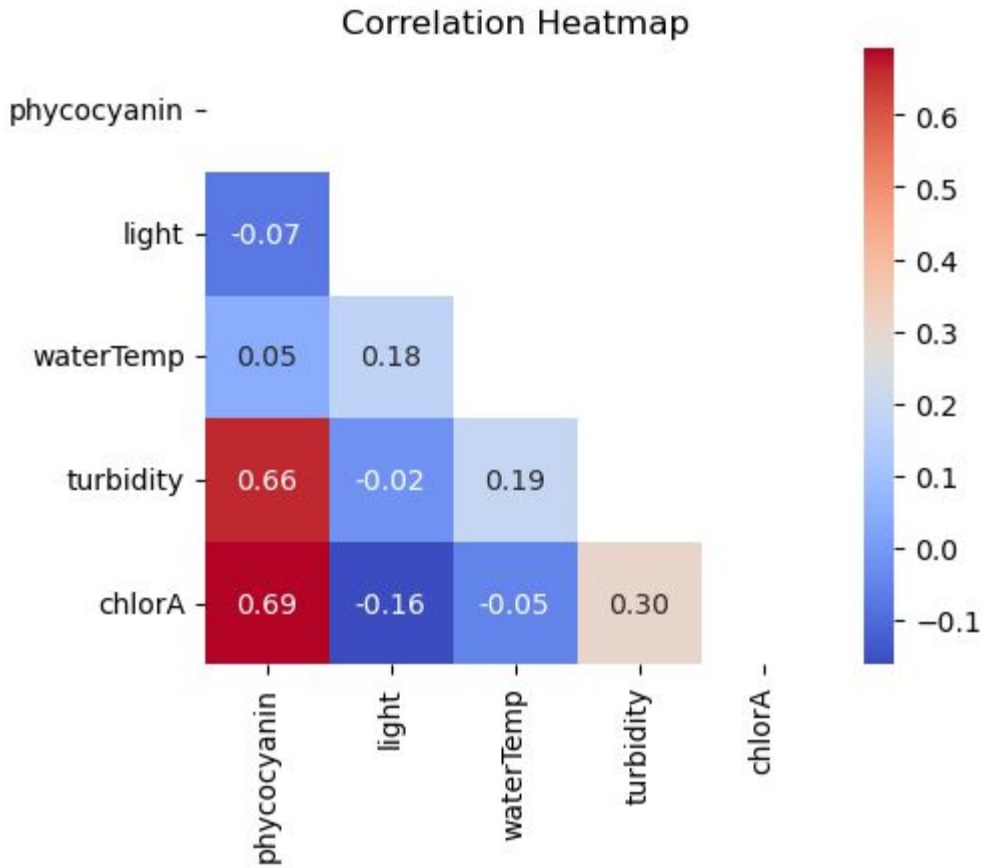






EDA: Correlations

- Moderately strong correlations between phycocyanin and turbidity and between phycocyanin and chlorophyll-a.
- Moderately weak correlations between turbidity and chlorophyll-a.
- Metrics correlated with phycocyanin likely to be good predictors
- Metrics with a low correlation to phycocyanin likely to be less relevant in the model



EDA: Stationarity

- KPSS and ADF tests for each metric
- Temperature, turbidity, and chlorophyll-a metrics are non-stationary
- Phycocyanin and sunlight
  - KPSS test indicated non-stationary
  - ADF test indicated stationary
  - “Difference stationary”
- All metrics need to be detrended

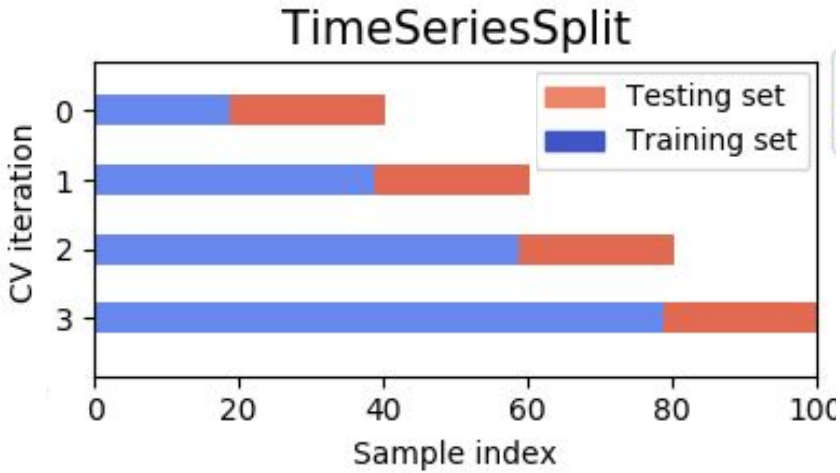
Engineering

- One day mean of metrics so same as predictions
  - 1,123 rows
- Reran Shapiro-Wilks tests
  - Not normal distributions
- Calculated natural logarithms
- Reran Shapiro-Wilks tests
  - Normal distributions
- Reran KPSS and ADF tests
  - Results similar to before
- Detrenended with differencing to make stationary
- Reran KPSS and ADF tests
  - All metrics stationary
- Created 1 through 20 day lagged versions of the original features and the logged and differenced features

Pre-processing

- Defined two different sets of X and y
  - Original data and lagged features of the original data
  - Transformed features and lagged features of the transformed data
- Allowed comparison of the performance of the models using unaltered and transformed data

Modeling



- Regression, random forest, XGBoost
- Cross-validation using TimeSeriesSplit
- Root mean squared error (RMSE)
- Models were run multiple times starting with including non lagged feature data and adding a level of lagged feature data with each iteration up through the 20 lag features
- Performed for both sets of X and y
- RMSEs for each iteration were graphed and model with lowest RMSE was identified
- Lowest RMSE from models using transformed data were un-transformed before comparison.

Results

- 0.7-3.3% of the phycocyanin range (1654.64) which is considered low error.
- Median (55) error would not constitute a blue green bloom at a dangerous level
- Hypertuning for best RMSE of 12.4

	Original Data		Transformed Data	
Model	Lagged Features	RMSE	Lagged Features	RMSE
Linear Regression	0-3	29.5	0	14.4
Random Forest	0-13	43.7	0-8	12.8
XGBoost	0-14	55.3	0-10	12.5

Conclusion

- XGBoost model using transformed data lagged at zero to 11 days produced the highest performing model to predict phycocyanin in Lake Fictitious
- Error is small in the grand scheme of HABs where phycocyanin values are, at a minimum, in the several hundreds
- Model could be very helpful for predicting blue-green algae in a lake before implementing a shutdown, allowing for planning and allocation of resources