Meteorological, oceanographic, and water-quality data correlations with chlorophyll content @ Chesapeake Bay

Pushkar Sathe and Jun Bo Lau

Github page

Introduction

Chesapeake Bay spans an area with rich biodiversity and supports jobs in commercial and recreational fishing. The Bay is sensitive to climate change which can have drastic effects on the ecosystem. The goal of this project is to study the interactions between biogeochemical indicators and chlorophyll levels in the Bay. Datasets are obtained from Chesapeake Bay Interpretive Buoy System and Chesapeake Bay Program Data Hub.

Problem and Metrics

- Boosting models are very suitable for tabular data and sparse data.
- We also compare Boosting models with MLPRegressor and RNNs with a large number of hyperparameters.
- SHAP Analysis was done to understand the importance of features.
- Mean square error loss function was used for all models as our objective is regression.

Data and Cleanup

- The two datasets are large, ~ 1GB, ~ 277.8 million entries where ~31% of all entries are NaN, and spans across 12 files from various sources with 62 (possibly same) features that are a mix of numerical, date and categorical values.
- Focused on numerical features, allowing us to severely reduce the size of the data set (at the cost of discarding a lot of information).
- Removed features/columns with extremal/irrelevant values and merged the datasets.
- Performed data imputation using mean, normal distribution, least squares and stochastic regressors for recurrent neural networks. All models used grid search with 5-fold CV for hyperparameter tuning.

Models

- XGBoost Regressor
- LightBGM Regressor
- Multi-layer Perceptron Regressor
- Recurrent Neural Network with dense layer for regression

Results

- We found that there was a large positive correlation of years and months with chlorophyll over the decades - indicating that it is increasing with time. Other important factors were Salinity, Dissolved Oxygen etc.
- We confirmed boosting models perform well in comparison to neural networks(NNs). However, RNNs performed much better than NNs with this data. RMSprop optimizer gave the best results within RNNs.

Summary and Future Work

Boosting models performed significantly better than MLP (1-2 hidden layers) and comparable performance with RNN. RNNs can also be used for forecasting. Obtaining high quality datasets is the most challenging step in this project. It can be refined further to improve the performance of all models. Furthermore, additional hyperparameter tuning can improve performance measures and additional computational resources would allow exploration of other models such as LSTM.