Booz | Allen | Hamilton

CMC
Chesapeake Monitoring
Cooperative

MODELING WATER POLLUTION WITH ML

HACK THE BAY

OVERVIEW

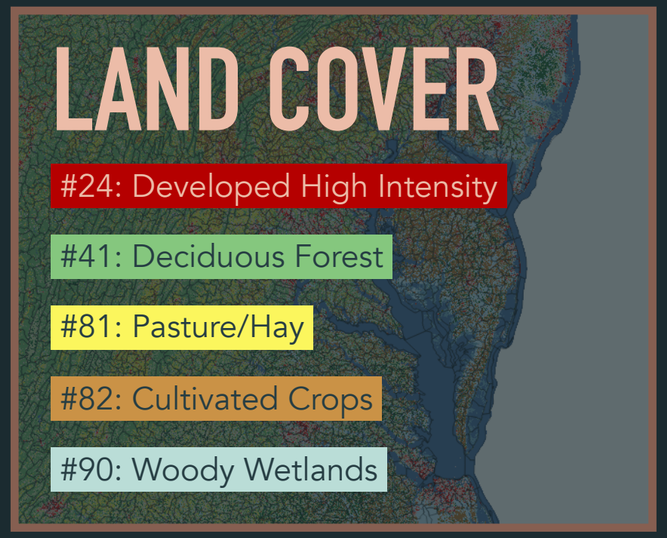
- Predicting Total Nitrogen in the Chesapeake Bay
- Feature variables
  - Datasets: Weather / land cover / air pollution
  - Created: outflow distance, point source NO_2
- Modeling and Selection
- Final model feature importance, and next steps

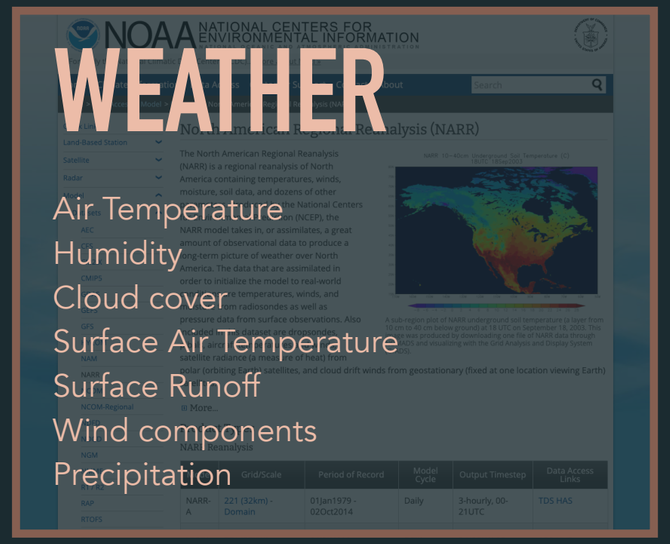
DATA EXPLORATION

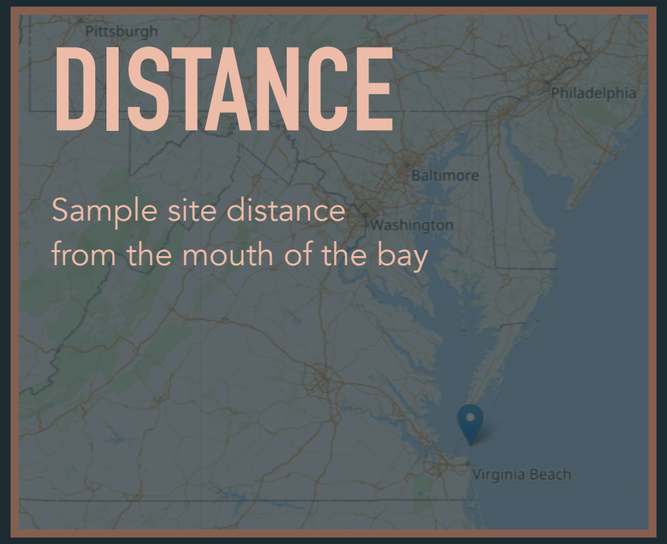
Target: Total Nitrogen (CMC/CBP)

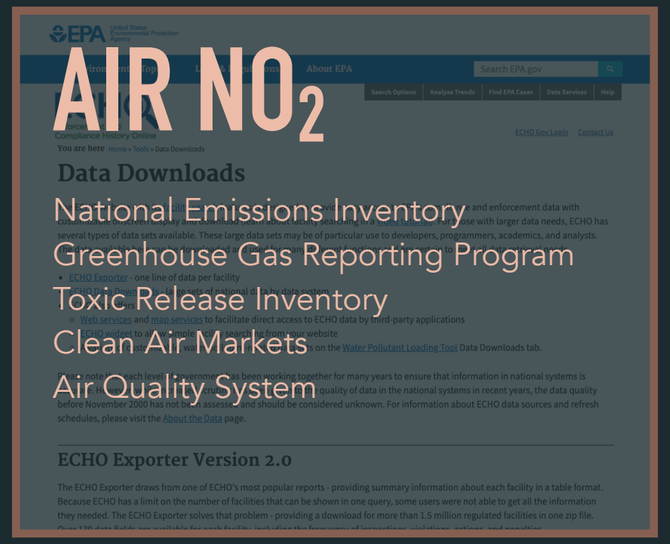
Features:

1. Land cover by use type (MLRC)
2. Weather (NARR)
3. Created: Distance from outflow
4. Created: NO_2 from point sources and air monitoring stations









MODELS

- XGBoost
  - KNN imputer, robust scaler
  - Excluded highly correlated features
- CatBoost
  - Transformed time feature
  - Consolidated land cover types

XGBOOST

RMSE	R^2	Explained Variance
0.90	0.83	0.83

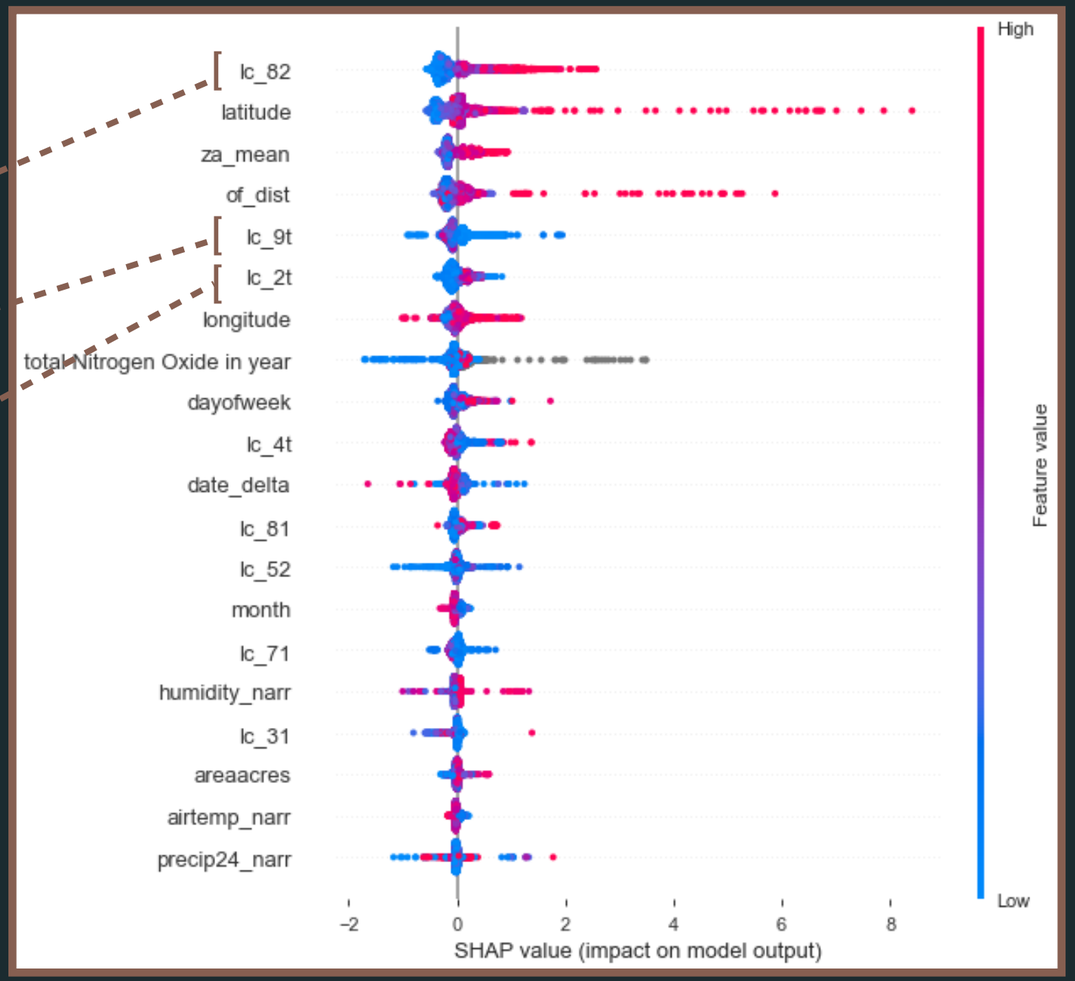
CATBOOST

RMSE	R^2	Explained Variance
0.92	0.84	0.84

FEATURE IMPORTANCE

- Feature importance highlights
  - Land cover types
    - lc_82: Cultivated Crops
    - lc_9t: Wetlands
    - lc_2t: Developed Land
  - Location (lat/long, outflow distance)

SHAP PLOT



teamHackTheBay / hackTheBay

Watch 0 | Star 0 | Fork 0

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

master | hackTheBay / models / catboost / HackTheBay Catboost water_final dataset-no huc all features .ipynb | Go to file

berenice-d Add files via upload — Latest commit 02fe418 3 days ago | History

1 contributor

1.76 MB | Download

HackTheBay: predictive analysis

1. Feature selection

In this notebook, we use the following variables for prediction: 'latitude', 'longitude', 'areaacres', 'za_mean', ('lc_21', 'lc_22', 'lc_23', 'lc_24') combined as lc_2t, 'lc_31', ('lc_41', 'lc_42', 'lc_43') combined as lc_4t, 'lc_52', 'lc_71', 'lc_81', 'lc_82', ('lc_90', 'lc_95') combined as lc_9t, month', 'year', 'week', 'dayofweek', 'hour', 'min', 'quarter', 'airtemp_narr', 'precip3_narr', 'humidity_narr', 'cl_cover_narr', 'sfc_runoff', 'windspeed_narr', 'wdirection_narr', 'precip24_narr', 'precip48_narr', 'of_dist', 'total Nitrogen Oxide in year', and 'date_delta'.

Date_delta is a numeric variable which capture the time in seconds from the latest record. We could not keep 'new_date' in a datetime format (not supported by Catboost). The reasoning behind creating date_delta is that other time variables (month, year, week, day of week and quarter) are categorical. They can capture a seasonal phenomenon (pollution from industry on weekdays for example) but not a trend over time.

We removed the following variables: 'new_date' (replaced by date*delta which is numeric), 'huc12*', and 'huc12_enc'.

The dependant variable (target) is the total nitrogen ('tn') in mg/L.

2. Catboost

Catboost can deal with missing values internally by giving them the minimal value for that feature (which translates into the guarantee to have a split that separates missing values from all other values). We want to test its capabilities on the non-imputed dataset.

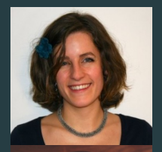
Furthermore, categorical variables in the dataset don't need to be dummified (that's where the name Cat-boost comes from, it's good with categorical variables).

Because it's an ensemble method, feature scaling is not necessary.

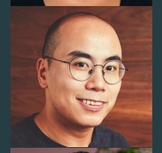
In [1]:
```
# Load packages
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
import sklearn
from datetime import date
from sklearn import metrics
from pandas_profiling import ProfileReport
import shap
import plotly

#metrics
from sklearn.metrics import r2_score
from sklearn.metrics import explained_variance_score
from sklearn.metrics import mean_squared_error
```

In [2]:
```
# Import catboost package
import catboost
```

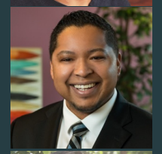
In [3]:
```
# Load dataset
df = pd.read_csv('data/final_water.csv', index_col=0)
df.head()
```



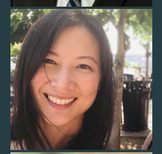
Berenice Dethier



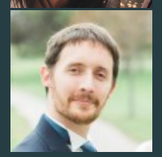
Justin Huang



Bryan Dickinson



Jen Wu



Tim Osburg

TEAM GITHUB

HTTPS://GITHUB.COM/TEAMHACKTHEBAY