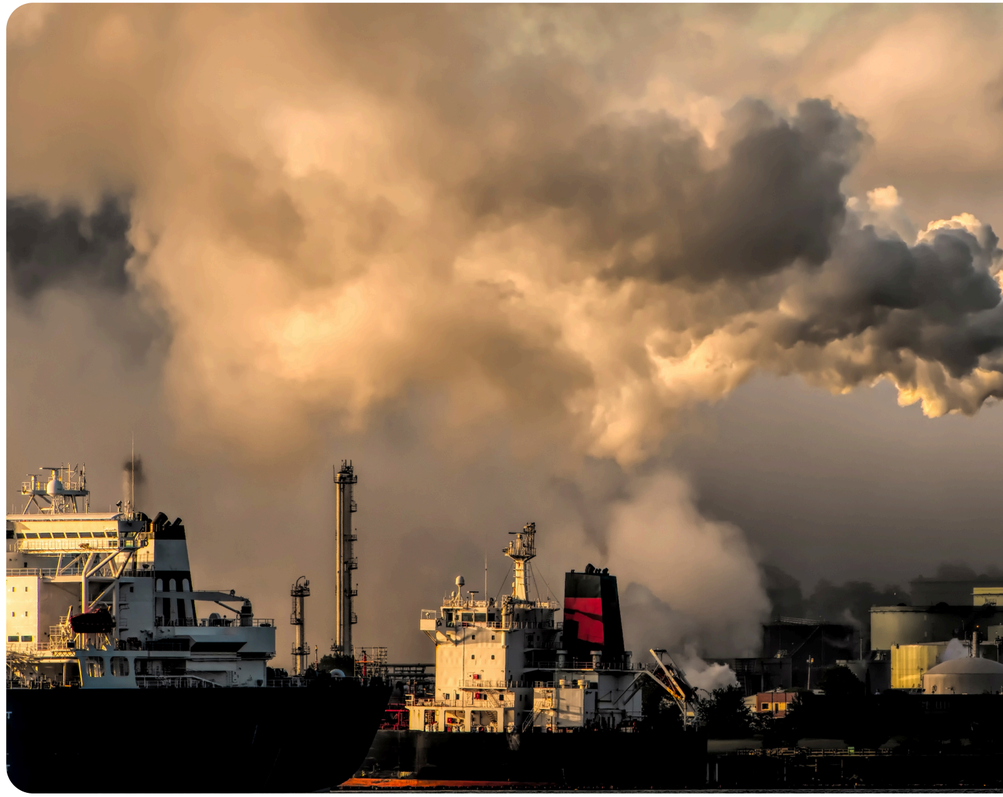
Air Quality Monitoring

By Vasilisa Matafonova, Tymur Marchenko, Sofia Zanti

Baseline models we used

- LinearRegression
- KNeighboursRegressor
- DecisionTreeRegressor
- BaggingRegressor (DecisionTreeRegressor)
- RandomForestRegressor
- DummyRegressor (strategies = mean, median)



Results we got for them

Linear regression

Test metrics:
Linear Regression Mean Squared Error: 12429
Linear Regression R2 Score: 0.35
Linear Regression Mean Absolute Error: 77.9
Linear Regression Root Mean Square Error: 111.4

Cross Validation values:
[-47348.11252565, -63450.6 , -75081, -2710.13, -258416.88]

KNeighboursRegressor

Test metrics:
knn regression model Mean Squared Error: 19279.5
knn regression model R2 Score: -0.006
knn regression model Mean Absolute Error: 79
knn regression model Root Mean Square Error: 138.85

Cross Validation values:
[-22850, -30903.9, -5758.9, -10549.10, -8746.88]

Results we got for them

DecisionTreeRegressor

Test metrics:
Decision Tree Regressor Mean Squared Error: 57.8
Decision Tree Regressor R2 Score: 0.99
Decision Tree Regressor Mean Absolute Error: 2.44
Decision Tree Regressor Root Mean Square Error: 7.6

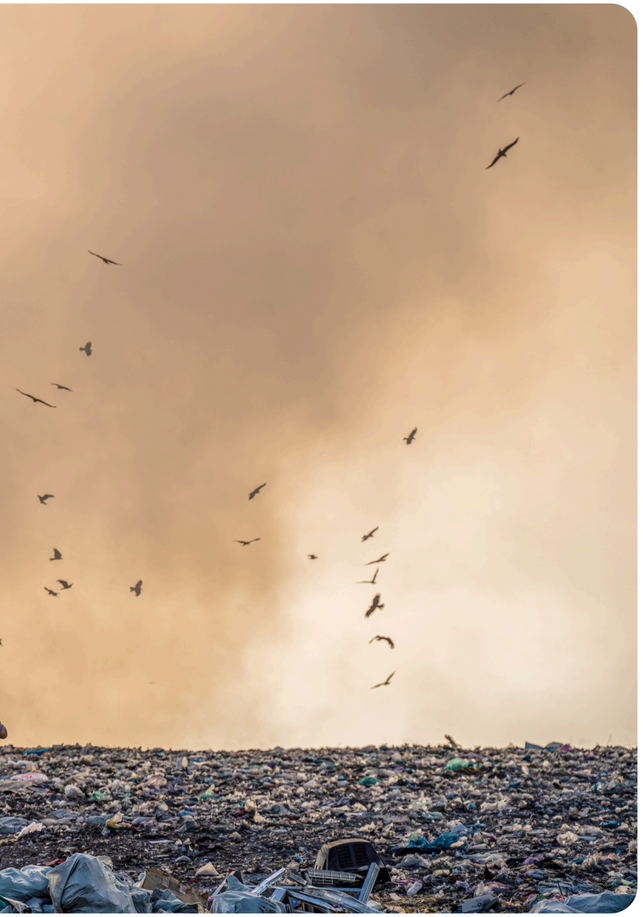
Cross Validation values:
[-22571.5, -31497.53, -5512.65, -26092.68, -61564.21]

BaggingRegressor (DecisionTreeRegressor)

Test metrics:
Bagging Mean Squared Error: 30.36
Bagging R2 Score: 0.99
Bagging Mean Absolute Error: 1.89
Bagging Root Mean Square Error: 5.51

Cross Validation values:
[-22738.95, -18886.88, -5919.73, -18601.85, -61297.69]

Results we got for them



RandomForestRegressor

Test metrics:
Random Forest Regressor Mean Squared Error: 25.57
Random Forest Regressor R2 Score: 0.99
Random Forest Regressor Mean Absolute Error: 1.78
Random Forest Regressor Root Mean Square Error: 5.05

Cross Validation values:
[-22716.59, -18374.44, -5798.02, -17882.38, -61232.59]

Results we got for them

DummyRegressor (strategy = mean)

Test metrics:
Mean model Mean Squared Error: 19168.77
Mean model R2 Score: -0.00099
Mean model Mean Absolute Error: 82.86
Mean model Root Mean Square Error: 138.45

Cross Validation values:
[-8656.9, -34862.41, -65789.09, -97.75, -97.29]

DummyRegressor (strategy = median)

Test metrics:
Median model Mean Squared Error: 19279.58
Median model R2 Score: -0.006
Median model Mean Absolute Error: 79.04
Median model Root Mean Square Error: 138.85

Cross Validation values:
[-6466.41, -29067.97, -54703.56, -0., -0.]

Insight and surprises

Our test metrics are not performing too badly, in fact we can even see some r2 scores at 0.99, however when we look at the cross validation scores, we can notice extreme values.

We re-looked at our dataset and saw that in fact, seemingly our main issue was the distribution of y - our response variable, which spiked at 400 and then mainly consisted of widely spread outliers, and the correlation between y and X, our response variable and our predictors, which if we look at the heatmap we produced for the dataset, had basically no correlation

In our attempts to fix this using preprocessing methods figuring out if that would work, we eventually noticed that not only do they not fix the issue, they very much magnify it, meaning that the problem lies within our dataset.

Our Next steps

1

See what we can do to this dataset, if things like limiting it would work, or try any other scaling to verify that absolutely nothing can be done

2

If there is nothing, go on a search for another dataset. As our dataset is composed of 2 separate datasets, we can at first try looking for a separate dataset that would record the particulate matter in the UK during a prolonged period of time (our y) and then see if the new y will have any correlation to our X. If that does not work then we can redirect completely and try to find two new datasets for x and y.

3

After having made sure that our datasets have correlation, so our predictors would not be useless, we will then re-clean, preprocess if necessary and attempt the models again.

Thank You