Urban Air Pollution Machine Learning Project

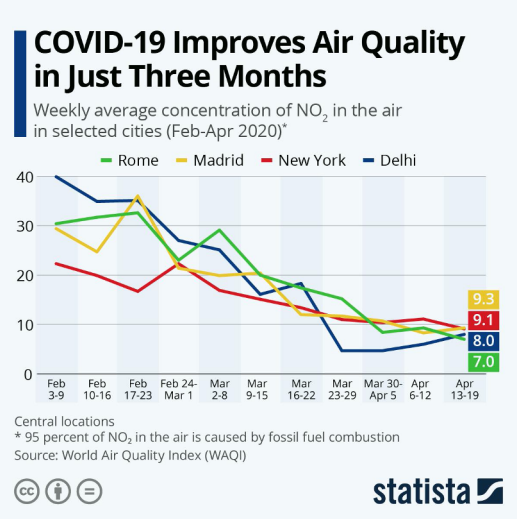
Rachel Kwon, Winston Tang

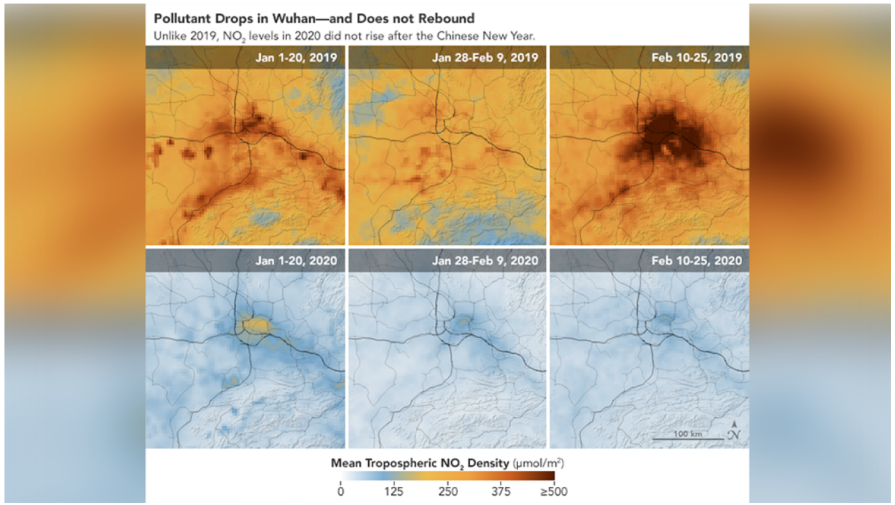
COVID-19 and the Environment

In the midst of a devastating global pandemic people have been searching for glimpses of positivity

One such idea is that nature is “healing” due to decreased human activity

Particularly that COVID-19 has led to cleaner air





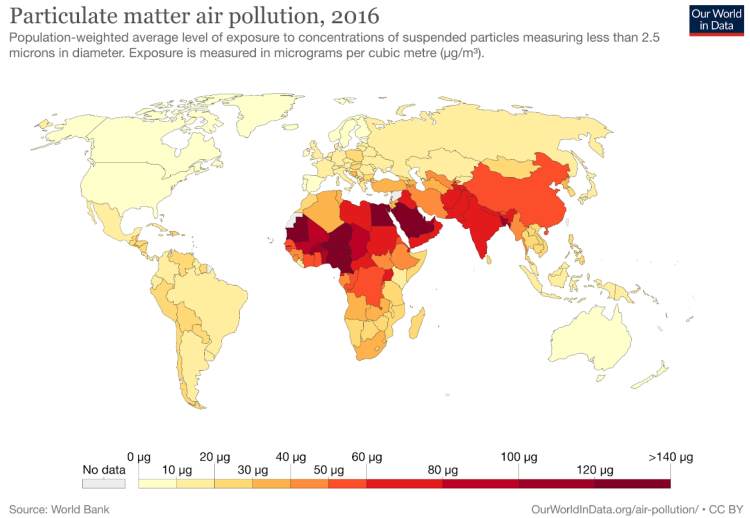
Air Pollution in Africa

Most statistics speak of NO2 concentrations, however, the most harmful air pollutants are particulates known as PM2.5 which can enter the lungs and bloodstream

We can see that Africa has had alarmingly high rates of PM2.5 concentration

During COVID-19, air pollution in African cities is worsening as people stay home

This is a serious problem in the long run and in the present as poor air quality can make respiratory illnesses more severe



Problem & Task

Problem: PM2.5 concentration is measured by ground-based sensors but not all cities in Africa have these sensors

Task: Use weather data and daily observations from the Sentinel 5P satellite tracking various pollutants in the atmosphere to predict PM2.5 every day for each city

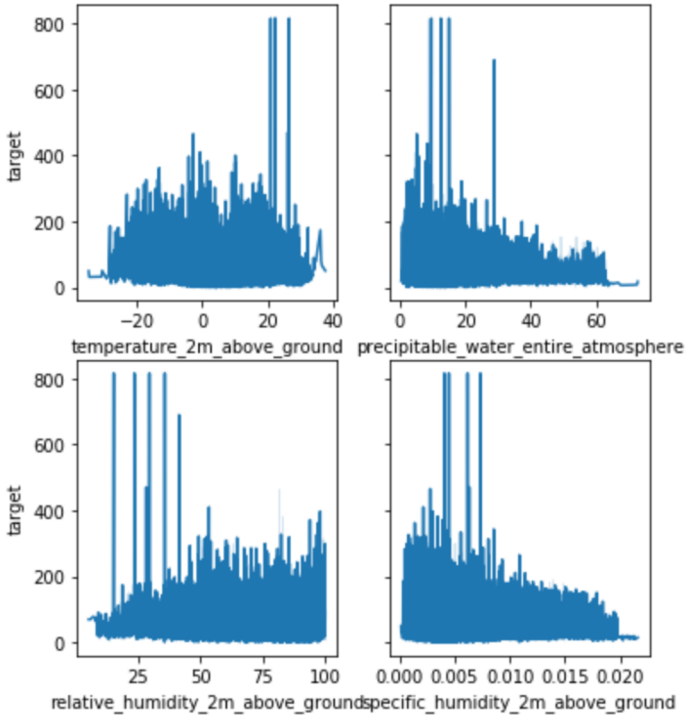
Data Collection

- Weather data provided by Global Forecast System: humidity, temperature, precipitation, etc
- Sentinel 5P satellite measures various pollutants in the atmosphere: NO2, SO2, CH4, etc

Data Preparation

- Dealing with missing values
- Creation of datetime columns: months, weeks, days

Insight with Outliers



The outliers with very high PM2.5 concentrations (~800) seem to be in areas of high temperature, low precipitation and humidity.

- The data roughly follows this trend as well

Possibly the most dry and arid regions of Africa

Feature Selection

Initially 82 columns

Correlation / Multicollinearity Analysis between ‘Target’ (PM2.5) and other variables

Also ran RandomForestRegressor to get top 20 most important features

Combined features from those 2 analyses together to reduce to 20 columns

	target
target	1.000000
L3_NO2_NO2_column_number_density	0.295235
L3_NO2_NO2_slant_column_number_density	0.303845
L3_NO2_tropospheric_NO2_column_number_density	0.252196
L3_CO_CO_column_number_density	0.341727
L3_HCHO_HCHO_slant_column_number_density	0.285927
L3_HCHO_tropospheric_HCHO_column_number_density	0.309343
L3_CLOUD_cloud_top_pressure	0.122033
L3_SO2_absorbing_aerosol_index	0.136017

Prediction Models

We used cross validation to evaluate different models and the RMSEs of their errors in predictions — the lower the score, the better the prediction is.

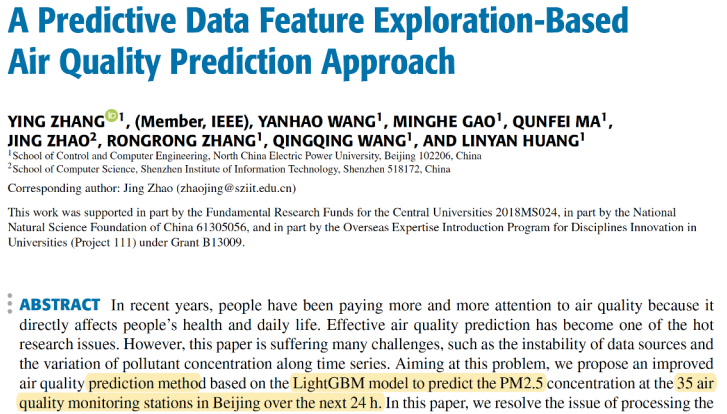
	fit_time	score_time	test_score	train_score
linear reg	0.36	0.06	39.09	38.27
tree	1.31	0.08	50.62	0.00
elasticnet	0.24	0.05	39.72	39.30
lasso	0.28	0.05	39.13	38.50
ridge	0.24	0.05	39.09	38.27
bagging_10	3.51	0.23	35.25	13.18
bagging_100	11.61	0.76	33.89	11.04
randomforest	8.53	0.48	33.89	11.02
xgboost_500	8.33	0.54	34.62	7.93
lightgbm_500	2.27	0.65	33.03	16.95

According to the scores, the xgboost and lightgbm models yield the best results. So we picked these two models for our prediction.

Prediction Result

With a few exceptions, both models produced predictions that are within close range with the real values.

We ultimately picked Lightgbm model due to current research which indicates that such a model leads to more accuracy.



A Predictive Data Feature Exploration-Based Air Quality Prediction Approach

YING ZHANG[1], (Member, IEEE), YANHAO WANG[1], MINGHE GAO[1], QUNFEI MA[1], JING ZHAO[2], RONGRONG ZHANG[1], QINGQING WANG[1], AND LINYAN HUANG[1]
[1]School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[2]School of Computer Science, Shenzhen Institute of Information Technology, Shenzhen 518172, China
Corresponding author: Jing Zhao (zhaojing@sziit.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities 2018MS024, in part by the National Natural Science Foundation of China 61305056, and in part by the Overseas Expertise Introduction Program for Disciplines Innovation in Universities (Project 111) under Grant B13009.

ABSTRACT In recent years, people have been paying more and more attention to air quality because it directly affects people's health and daily life. Effective air quality prediction has become one of the hot research issues. However, this paper is suffering many challenges, such as the instability of data sources and the variation of pollutant concentration along time series. Aiming at this problem, we propose an improved air quality prediction method based on the LightGBM model to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing over the next 24 h. In this paper, we resolve the issue of processing the

LightGBM

target	prediction
17.0	19.223499
25.0	22.809158
63.0	35.780773
34.0	33.774975
44.0	43.973038

XGBoost

target	prediction
17.0	15.180000
25.0	24.650000
63.0	40.020000
34.0	47.230000
44.0	41.900002

Recommendations

1. Local and international environmental agencies should pay closer attention to the air quality in major urban areas — people staying at home and less traffic do not equate to less air pollution
2. Researchers can look into the reasons behind changes in air pollution during the COVID-era and leave the implications for further studies in environmental protection and public health policies in the future
3. Confirmed: greenhouse gases densities are directly correlated to PM2.5 values

Q&A Time

Thank you!