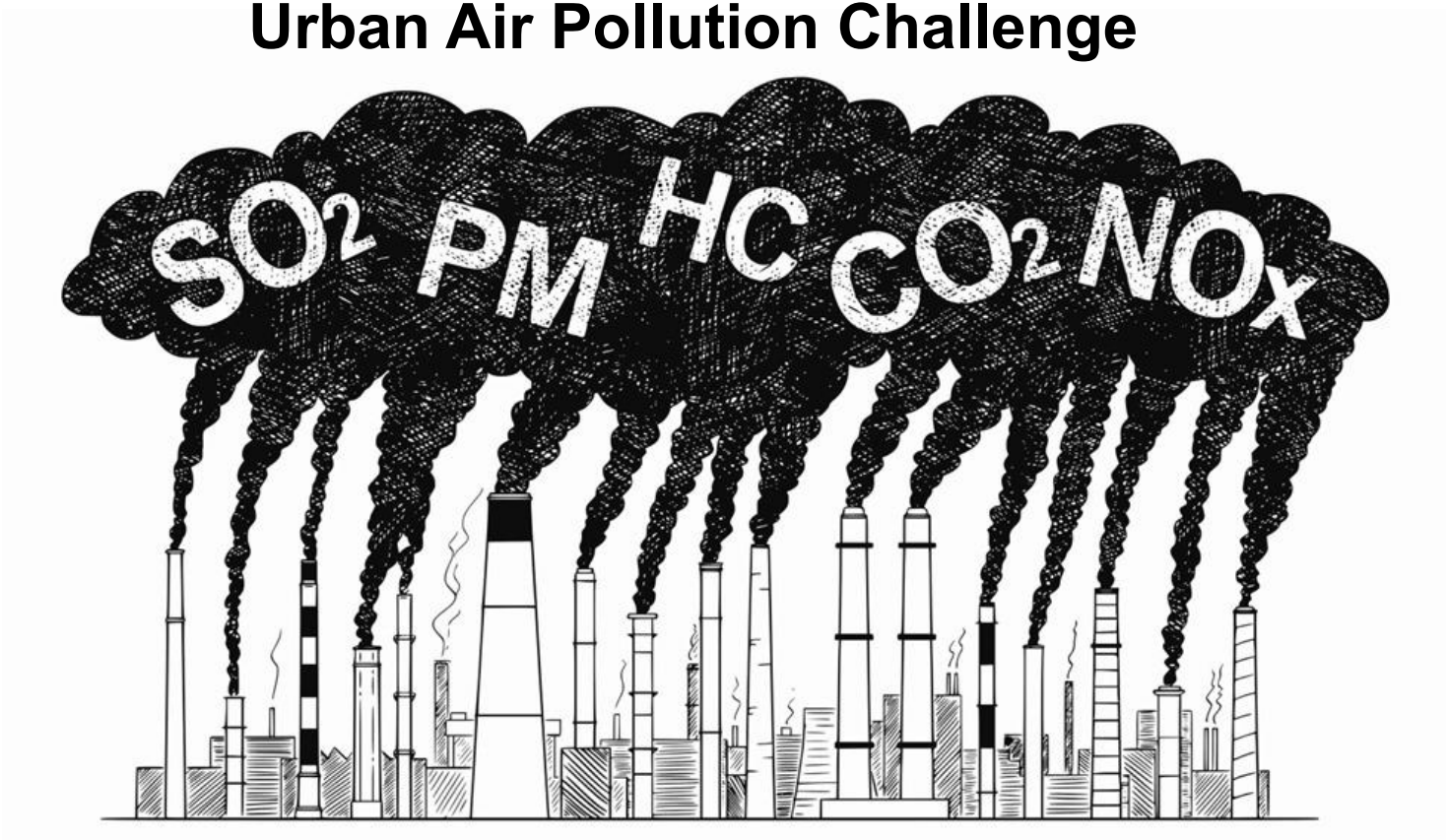
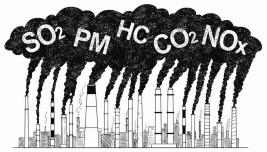
Urban Air Pollution Challenge



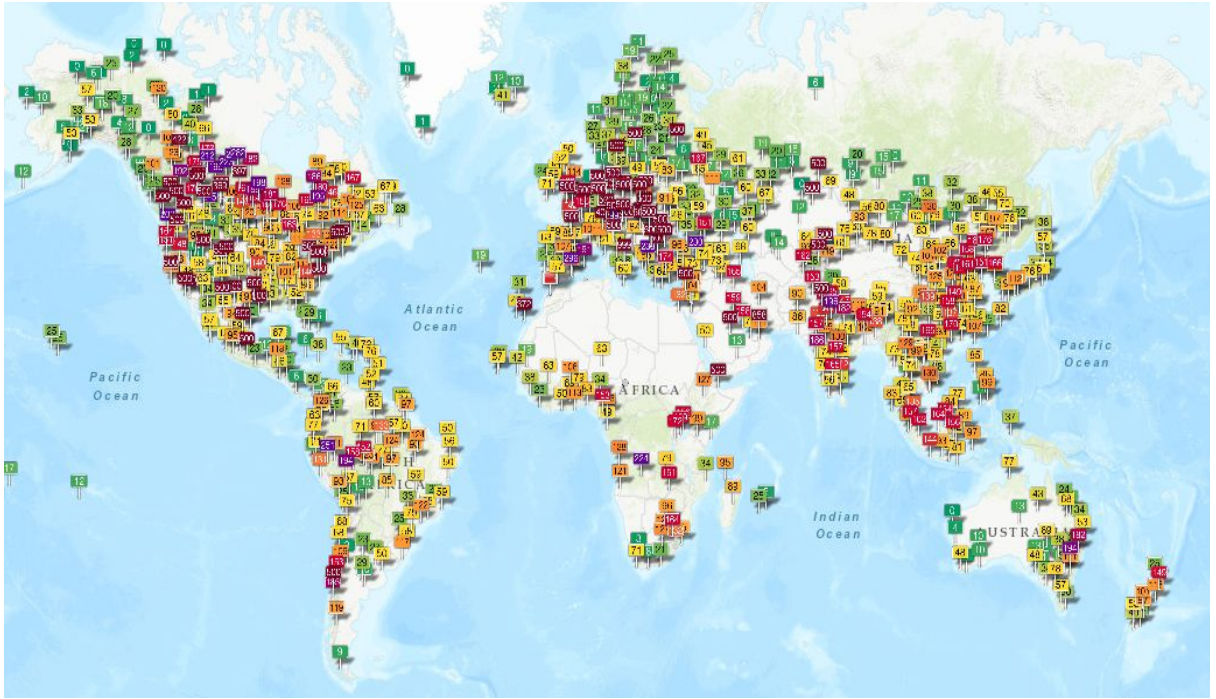
Andreas Kunzendorf, Selchuk Hadzhaahmed, Stefan Berkenhoff, Till Meineke, 04.09.2023



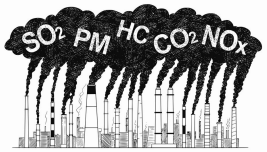
Overview

Introduction	
Baseline model	*Dummy Regressor*
1st model	*Random Forest Regressor*
2nd model	*Random Forest Regressor w/ engineered features*
3rd model	*Lightgmb w/ engineered features*
Summary	

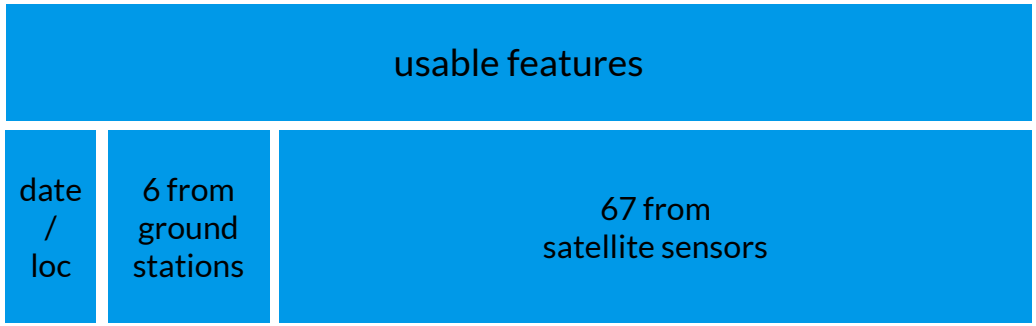
Introduction: Urban Air Pollution Challenge - Air pollution by PM2.5



https://aqicn.org/map/world/



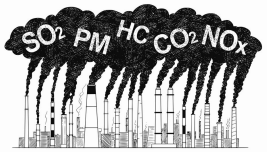
The Data - Columns



e.g.
temperature, windspeed

& the target:
PM2.5 particle concentration

e.g.
sensor readings for: UV Aerosol Index, Cloud, Carbon Monoxide, Formaldehyde, Nitrogen Dioxide, Ozone, …
Angles for each sensor

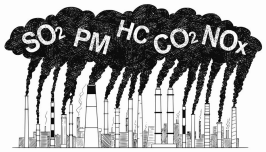


The Data - Rows

1. For 35k entries for ~350 locations and ~3 months

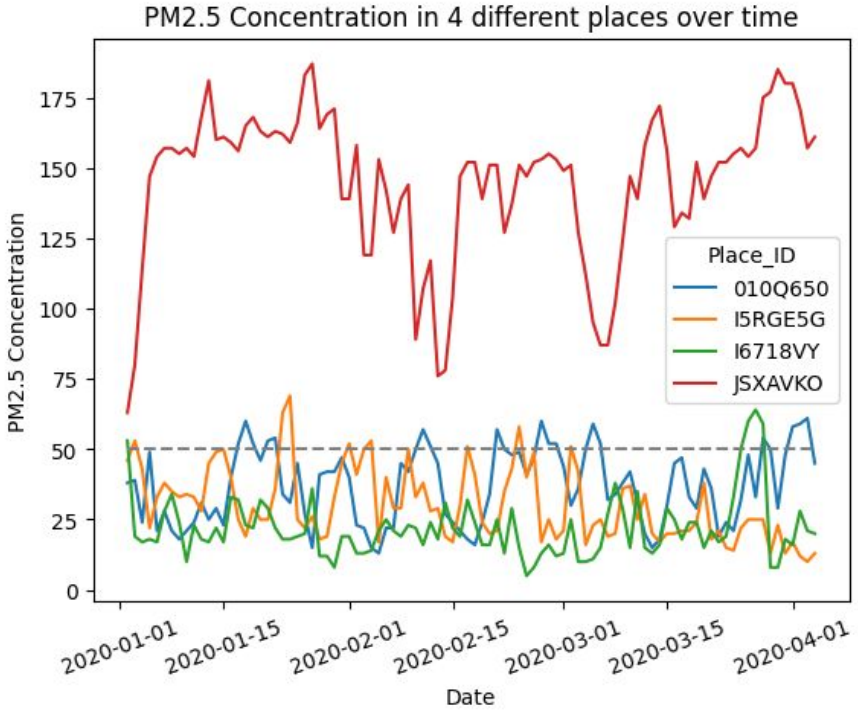
2. High variety in quality, completeness and size of records per location

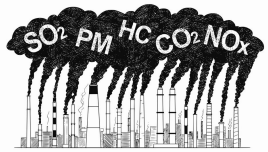
3. Hyperlocality vs “spaceview”: Smallest dimension of a satellite is 1 x 1 km, whereas ground sensors measure for a very particular ground position of several meters



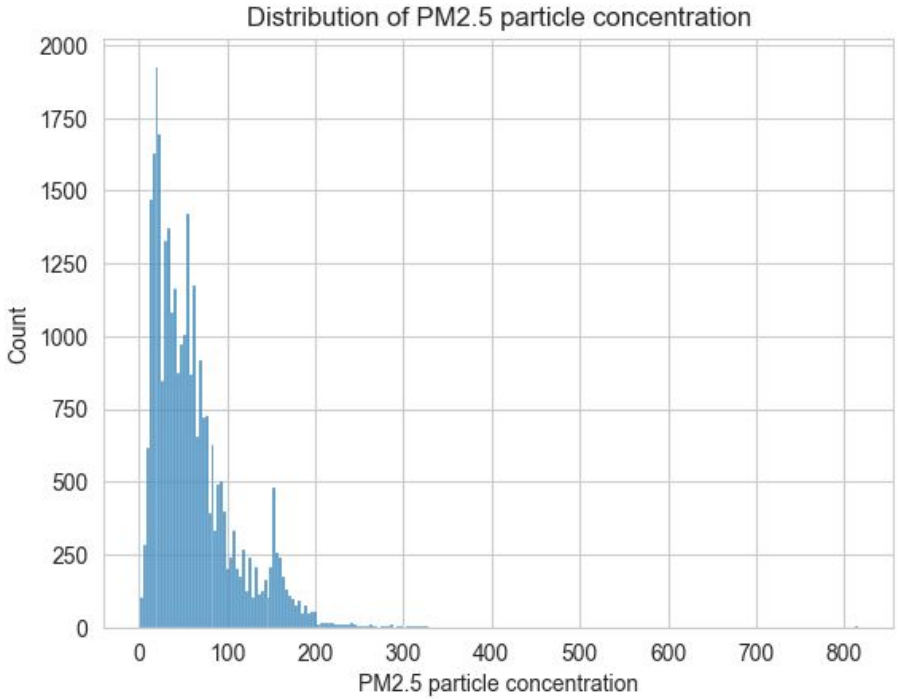
Baseline model: A starting point but not more

- Median value of PM2.5 particulate matter concentration over all cities and dates.
- $PM_{2.5}$ = 50 mg/m^3
- RMSE: 50.67





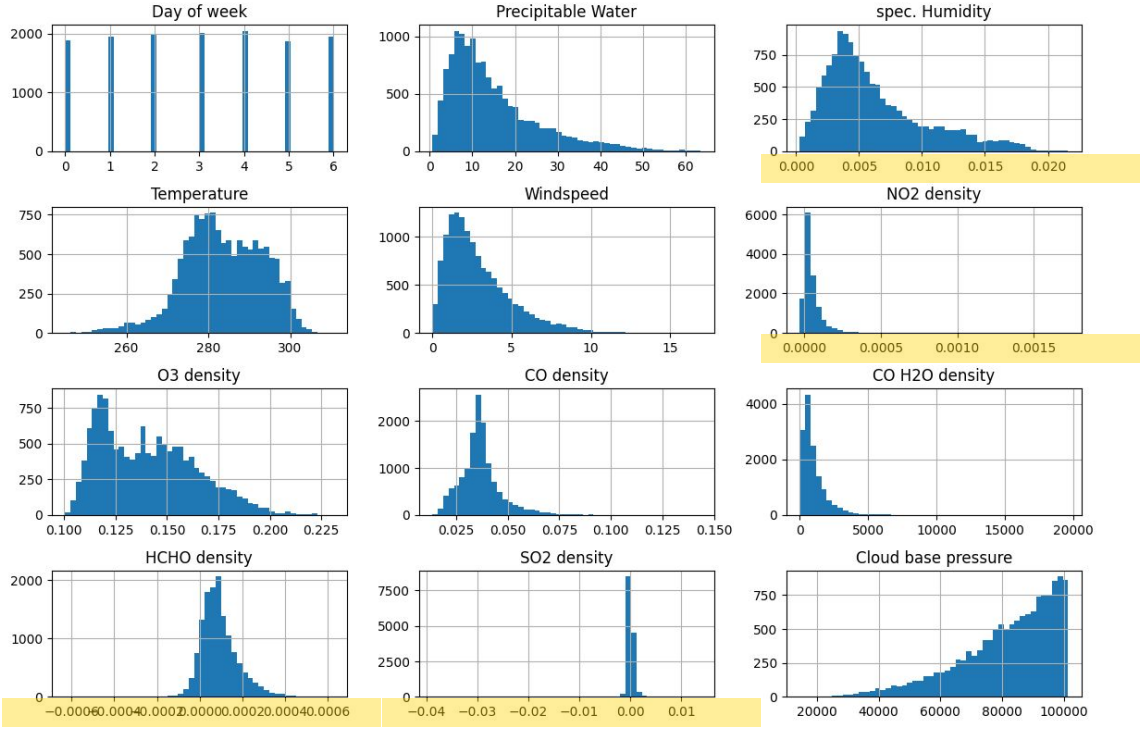
The Target: Higher PM2.5 particle concentration is bad for the health

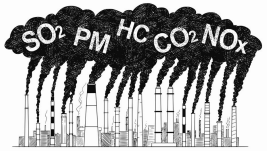


0 - 50	Good
51 -100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
300+	Hazardous

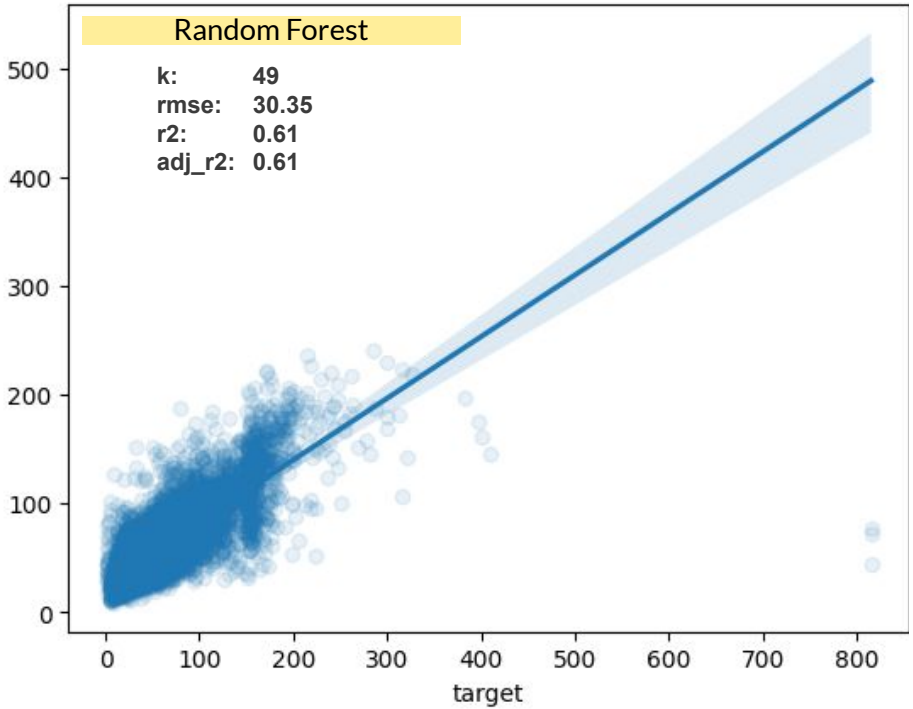
https://aqicn.org/scale/

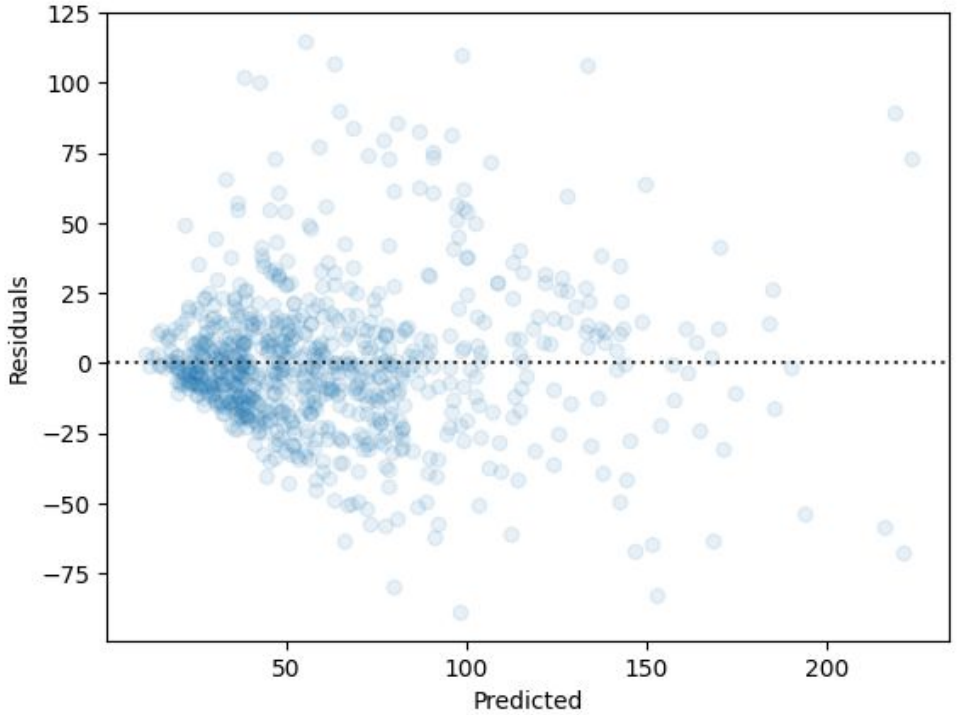
Baseline model: Data is skewed and resistant to scaling

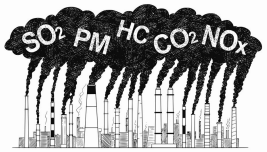




The initial ML model on the given data performed surprisingly well







Feature engineering: Calculating Trends adds more information based on the given data

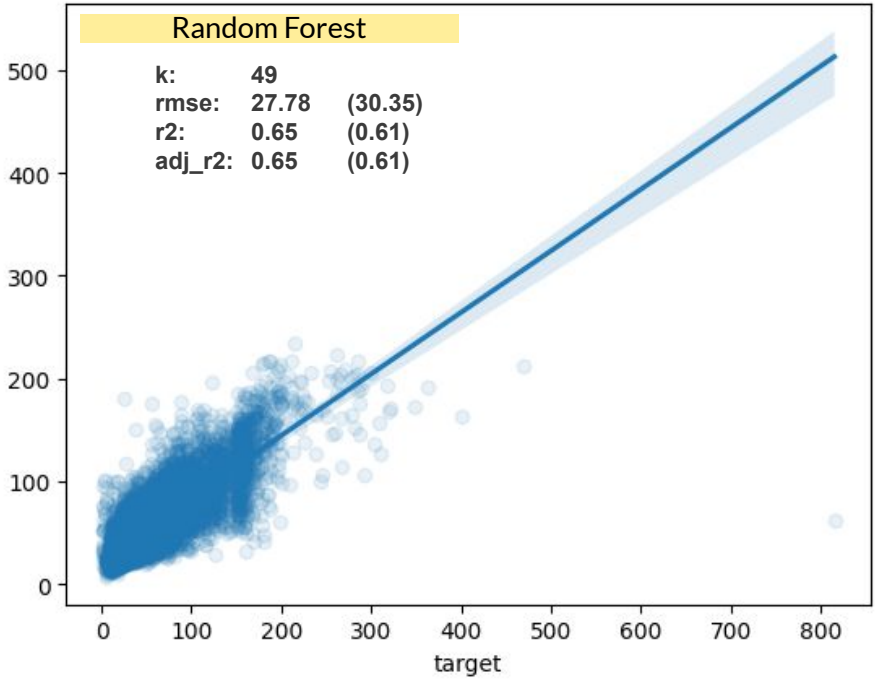
Trend based on time series

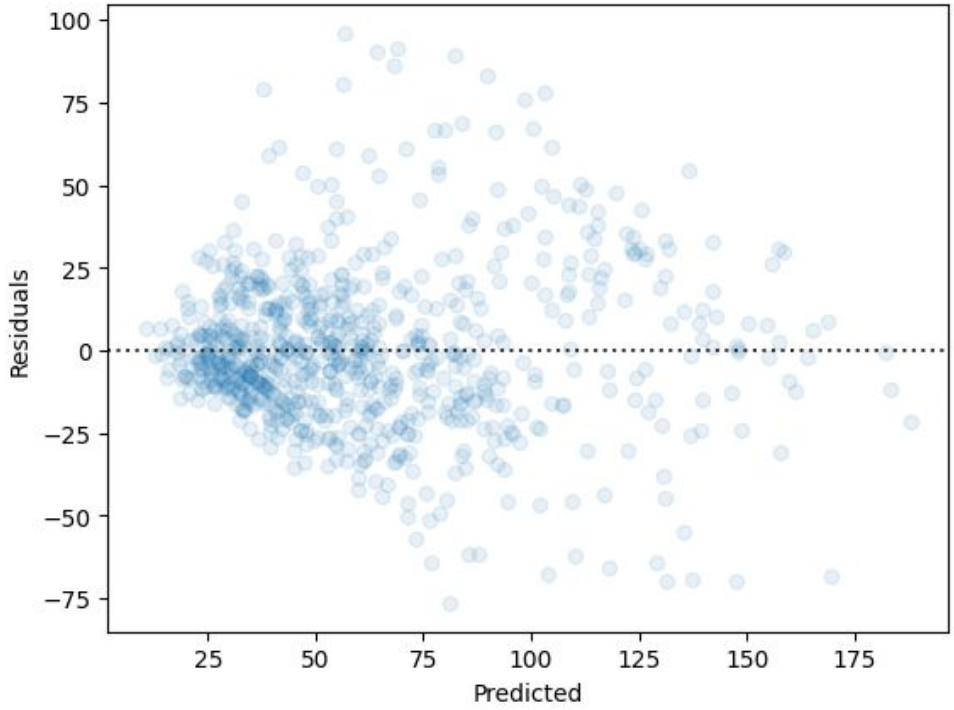
Location	Date	Temperature	Trend
LG56B	01.02.2017	20°	0
LG56B	02.02.2017	22°	+2
LG56B	03.02.2017	21°	- 1

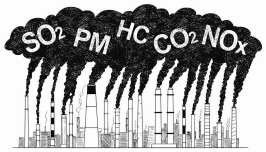
Trend per location

Encoding changes over time for the model

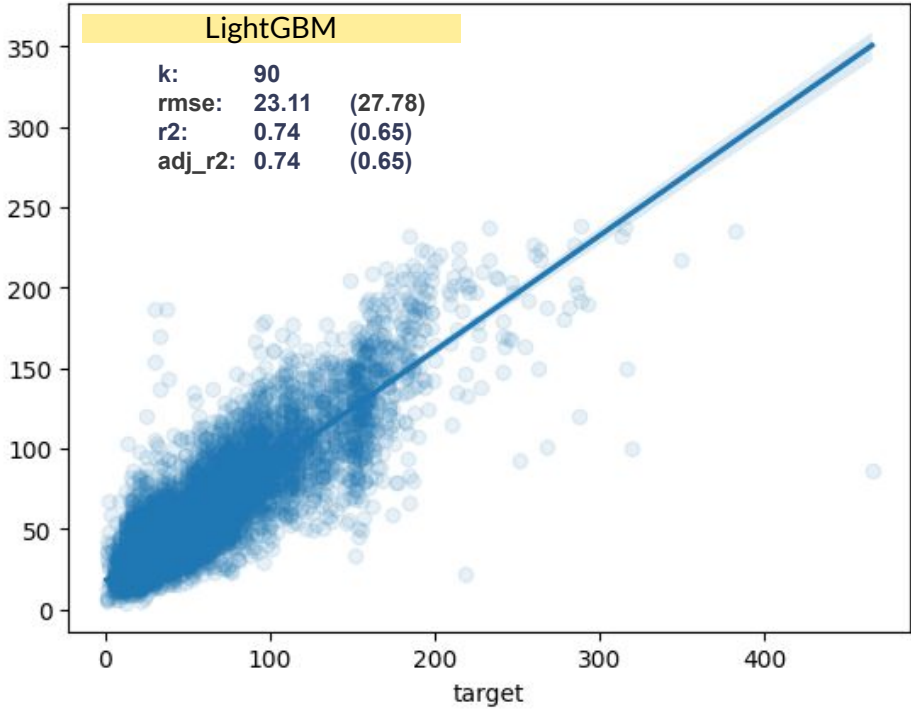
Introducing Trend data improved the second model tremendously

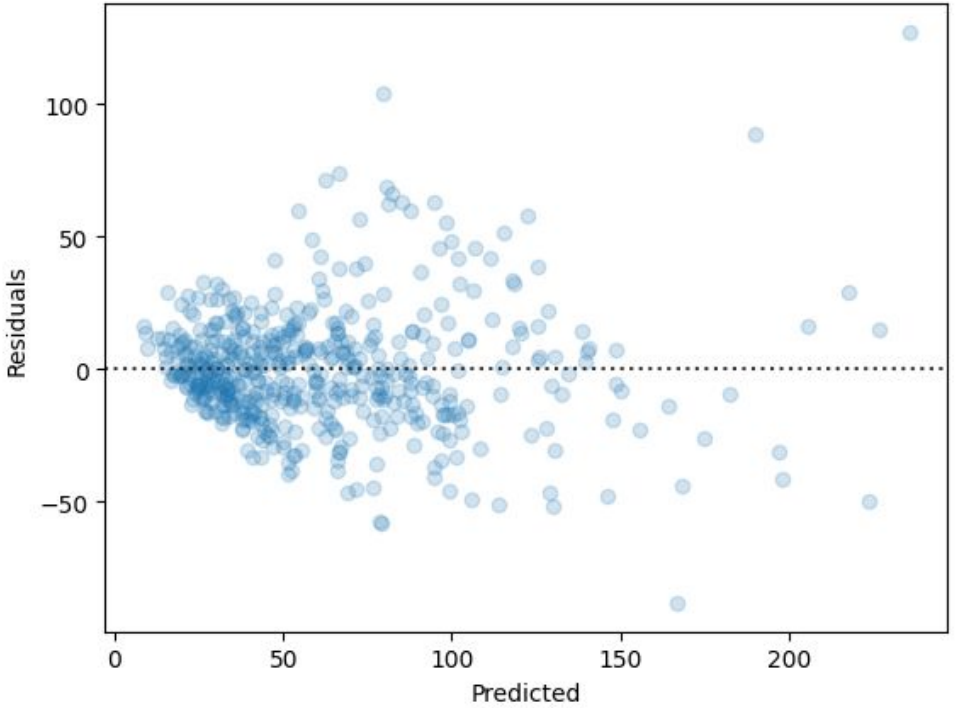


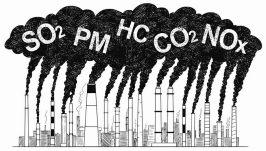




Switching from Random Forest to LightGBM gives another boost

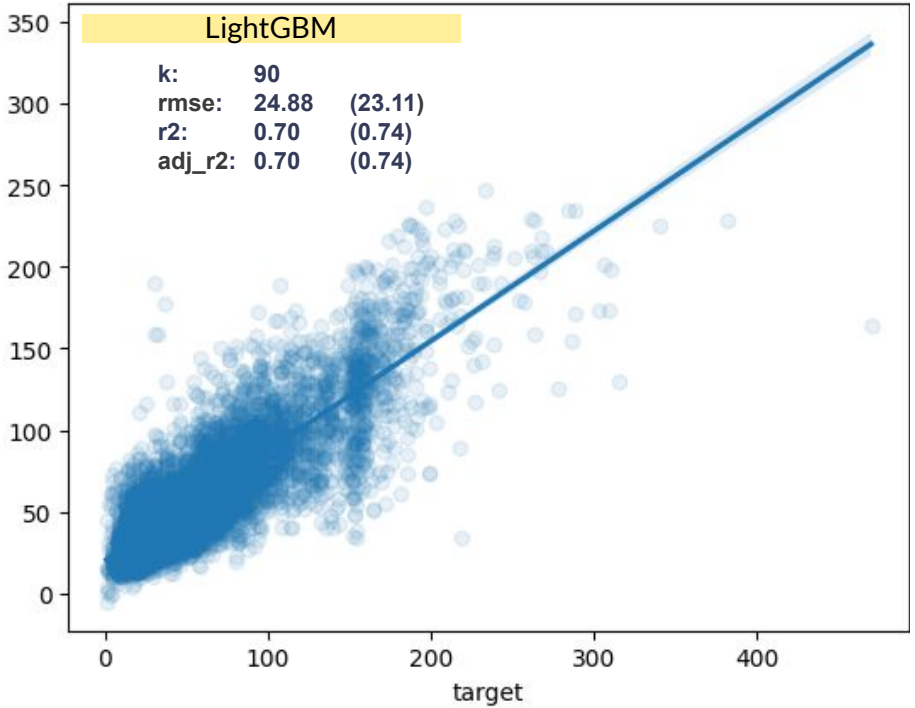


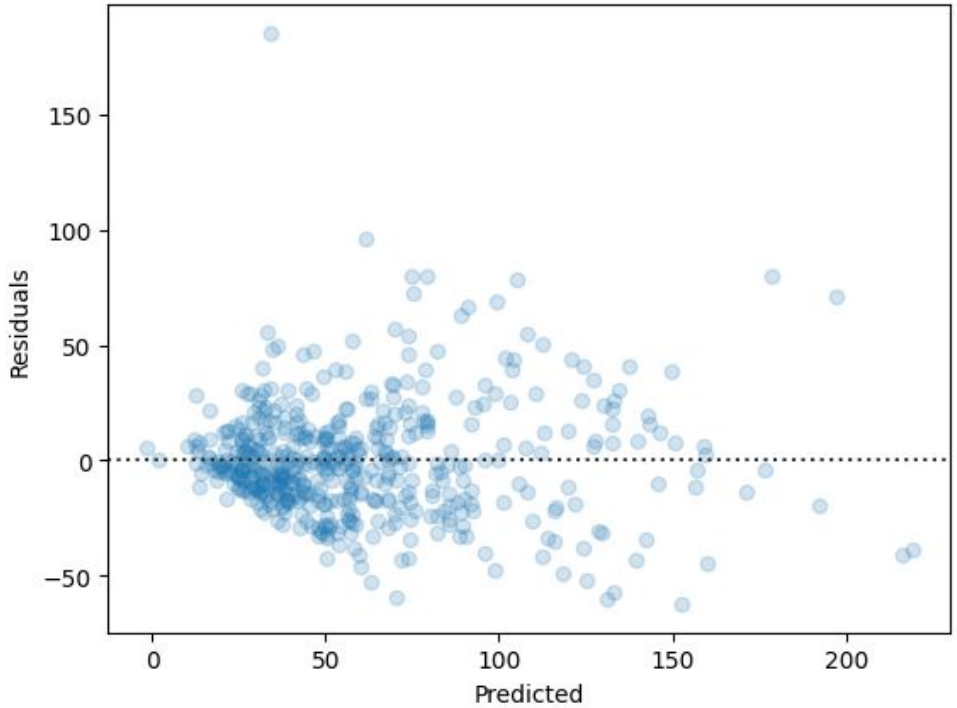


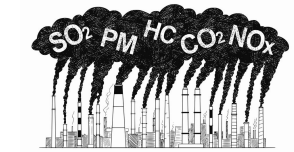


Even excluding ground sensor data performed better than the 2nd model

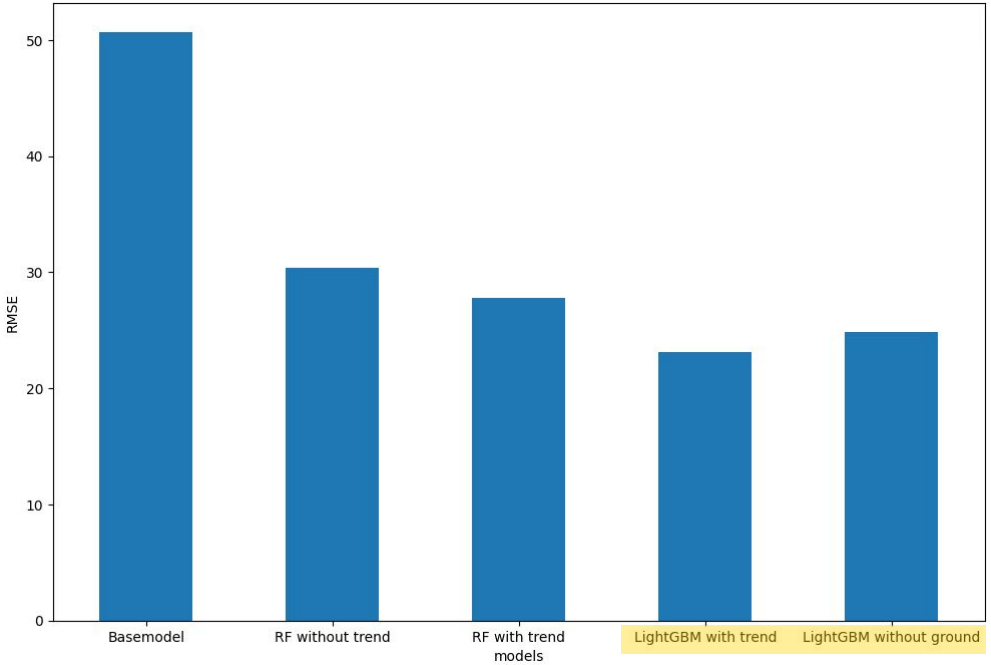
(No windspeed, temperature etc.)





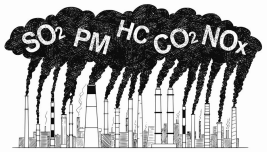


Summary



Key findings:

- Approximate prediction of air quality with satellite data is possible, but only rough estimations
- Feature engineering with time-based changes (trends) can have a positive effect on prediction



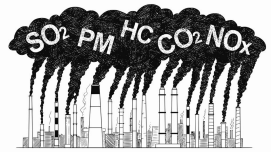
Potential Data product - how can the model be used?

1. Combine satellite data with available local weather data to predict PM2.5 particles in Africa

2. Early warning system for the population based on this predictions of the pollution levels (e.g. via app).

3. For a productification, use a rough classification (air harmful yes / no meaning concentration > 100)

0 - 50	Good
51 -100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
300+	Hazardous



Future work

- Test model predictions for unknown places
- Predict mean target over longer times (week, month)
- Build classification model (Good - Hazardous)
- Hyperlocality vs “spaceview”:
  a. Smallest dimension of a satellite is 1 x 1 km
  b. Ground sensors measures within meters
- Improve data basis
  a. use more records with high air pollution
  b. encoding geo locations in model