Zindi
Urban Air Pollution Challenge

Machine Learning Project

NeueFische

14.02.2023 – 17.02.2023

Clara Bredow

Isabelle Flaig

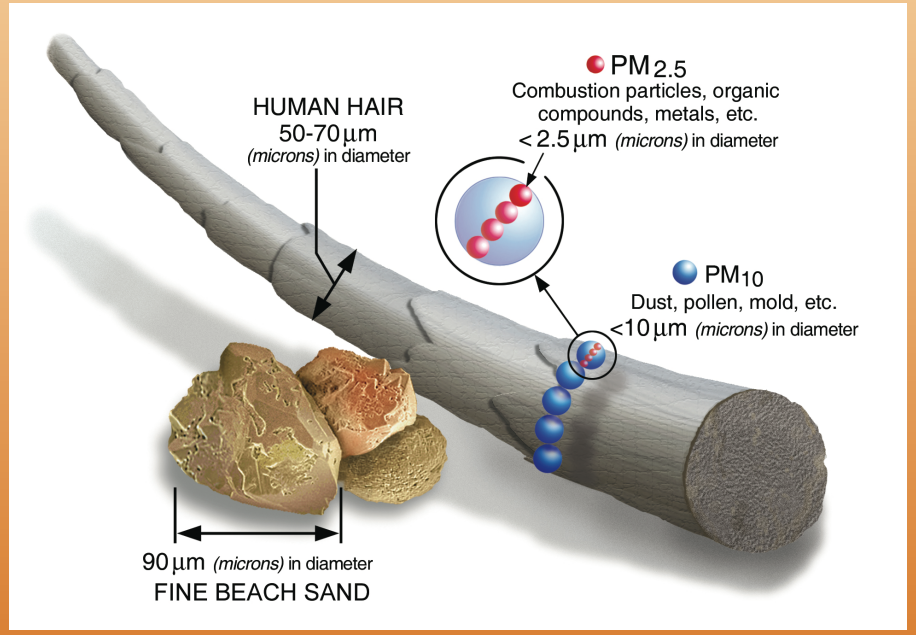
Gunnar Oehmichen

About the challenge

- World Air Quality Index
  - Non-profit
  - Promote air pollution awareness
  - Provide world-wide air quality information
- PM2.5 particulate matter
  - Atmospheric particulate matter < 2.5μm diameter
  - Common measure of air quality

Challenge

- Objective
  - Predict PM2.5 concentration based on
    - Global Forecast System (GFS) for weather
      - Humidity, temperature, wind speed
    - Sentinel 5P satellite monitors
      - pollutants in the atmosphere
        - NO_2
        - CO
        - SO_2
        - CH_4
        - O_3
        - HCHO
      - Cloud coverage
      - Pollutant-absorbing aerosol index
- Metric: RMSE



United States Environmental Protection Agency Website, 16.02.2023
https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

PM2.5

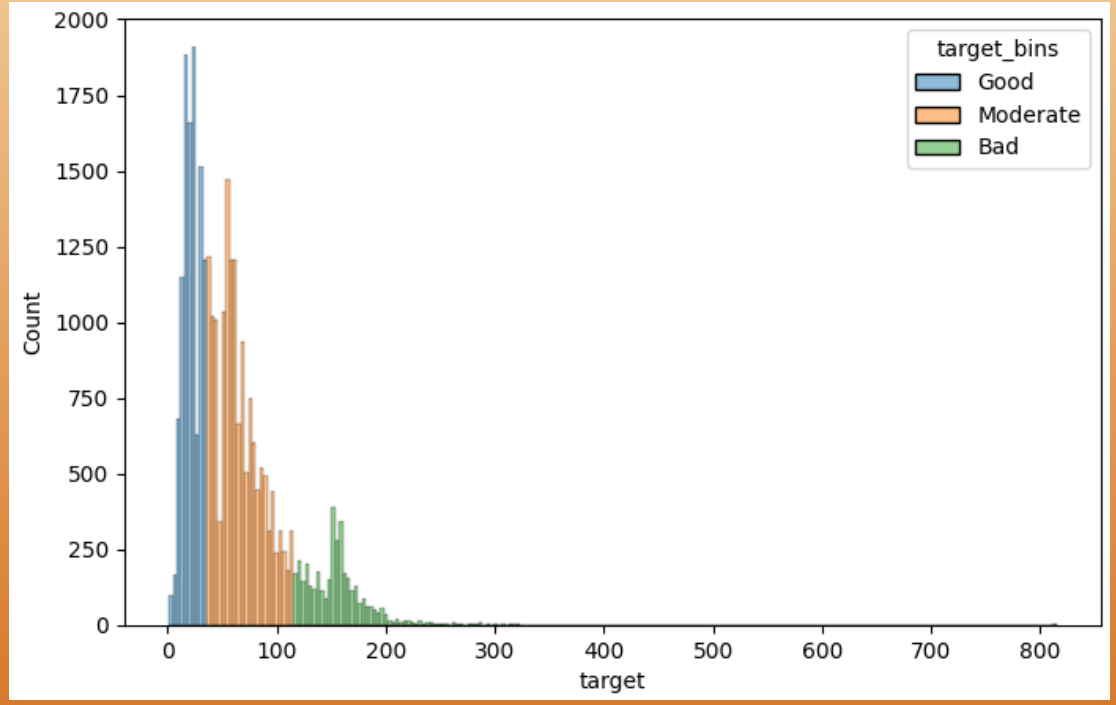
Levels of Health Concern	PM2.5(μg/m³) over 24h	Health Implications
Good	0 – 35	Air pollution poses little to no risk
Moderate	35 – 75	Moderate health concern for a very small number of people
Unhealthy for Sensitive Groups	75 – 115	People with heart and lung disease, older adults, and children are at greater risk
Unhealthy	115 – 150	Everyone may begin to experience adverse health effects (for sensitive groups: serious health effects)
Very Unhealty	150 – 250	Health alert (serious health effects for everyone)
Hazardous	> 250	Health warning of emergency conditions (entire population is likely to be affected)

Particulate Matter Matters
Dominici et al., Apr 2014
Science, DOI: 10.1126/science.1247348

Population Exposure to PM2.5 in the Urban Area of Beijing
Zhang et al., May 2013
PLoS ONE 8(5):e63486, DOI: 10.1371/journal.pone.0063486

PM2.5 values

Levels of Health Concern	PM2.5(μg/m³) over 24h	Health Implications
Good	0 – 35	Air pollution poses little to no risk
Moderate	35 – 75	Moderate health concern for a very small number of people
Unhealthy for Sensitive Groups	75 – 115	People with heart and lung disease, older adults, and children are at greater risk
Unhealthy	115 – 150	Everyone may begin to experience adverse health effects (for sensitive groups: serious health effects)
Very Unhealty	150 – 250	Health alert (serious health effects for everyone)
Hazardous	> 250	Health warning of emergency conditions (entire population is likely to be affected)



Population Exposure to PM2.5 in the Urban Area of Beijing
Zhang et al., May 2013
PLoS ONE 8(5):e63486, DOI: 10.1371/journal.pone.0063486

Cleaning – features kept

- Weather station measurement
  - Windspeed, temperature, humidity, precipitation- Satellite
- Weekday
- Measurements:
  - Aerosol-index
  - Cloud properties
  - Molecule concentrations (NO_2, SO_2, O_3, HCHO, CO)

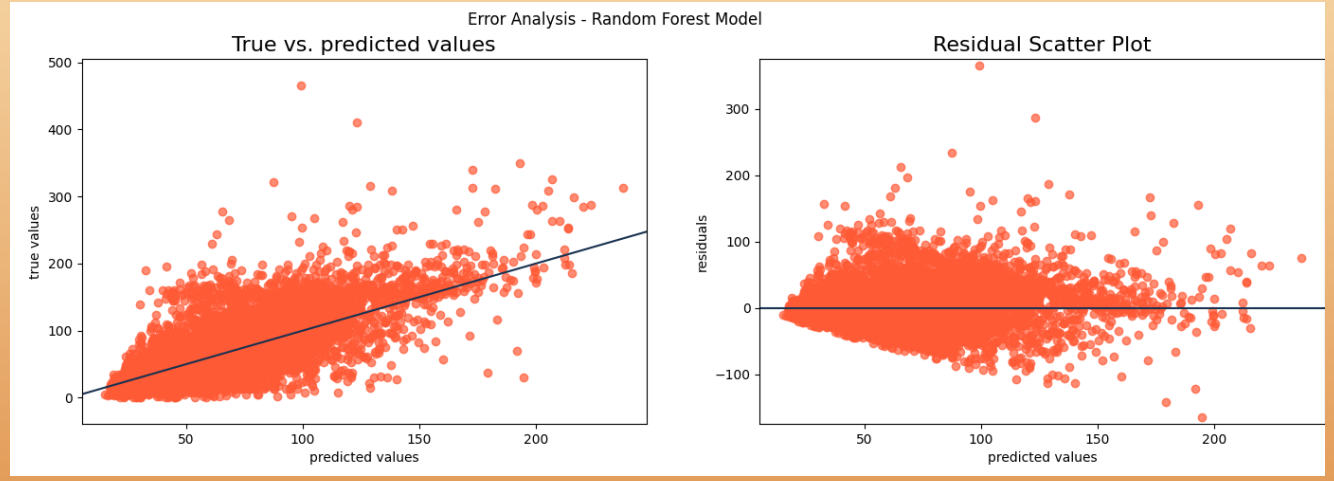
  →Scaling if necessary for the model (KNN, LogReg)
- 82 features → 25 features

Models

Type	Classification			Regression	
Metrics (Test)	KNN	LogReg	Random Forest Classifier	KNN Reggressor	Random Forest Regressor
RMSE				27.94	30.72
F1-Score	0.67	0.49	0.62		
Cohen‘s Kappa Score	0.44	0.15	0.34		

Error Analysis

Random Forest



KNN Regressor

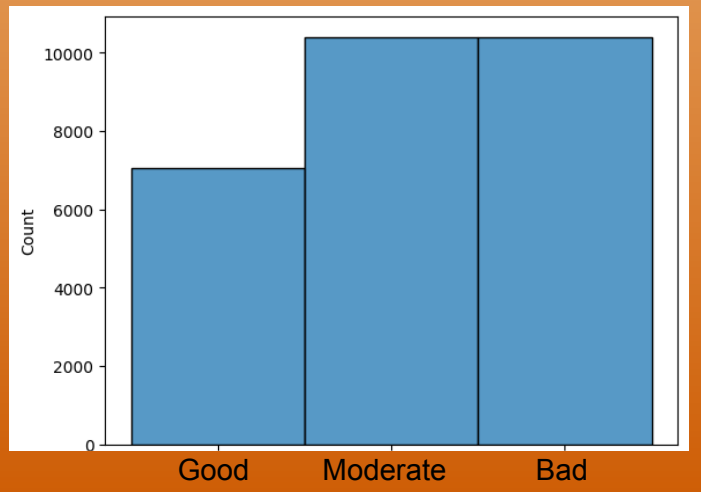
Refinement

- Oversampling
  - Randomly introducing duplicate examples in the minority class

```
# define oversampling strategy
        oversample = RandomOverSampler(sampling_strategy='minority’)
```

Before oversampling

After oversampling



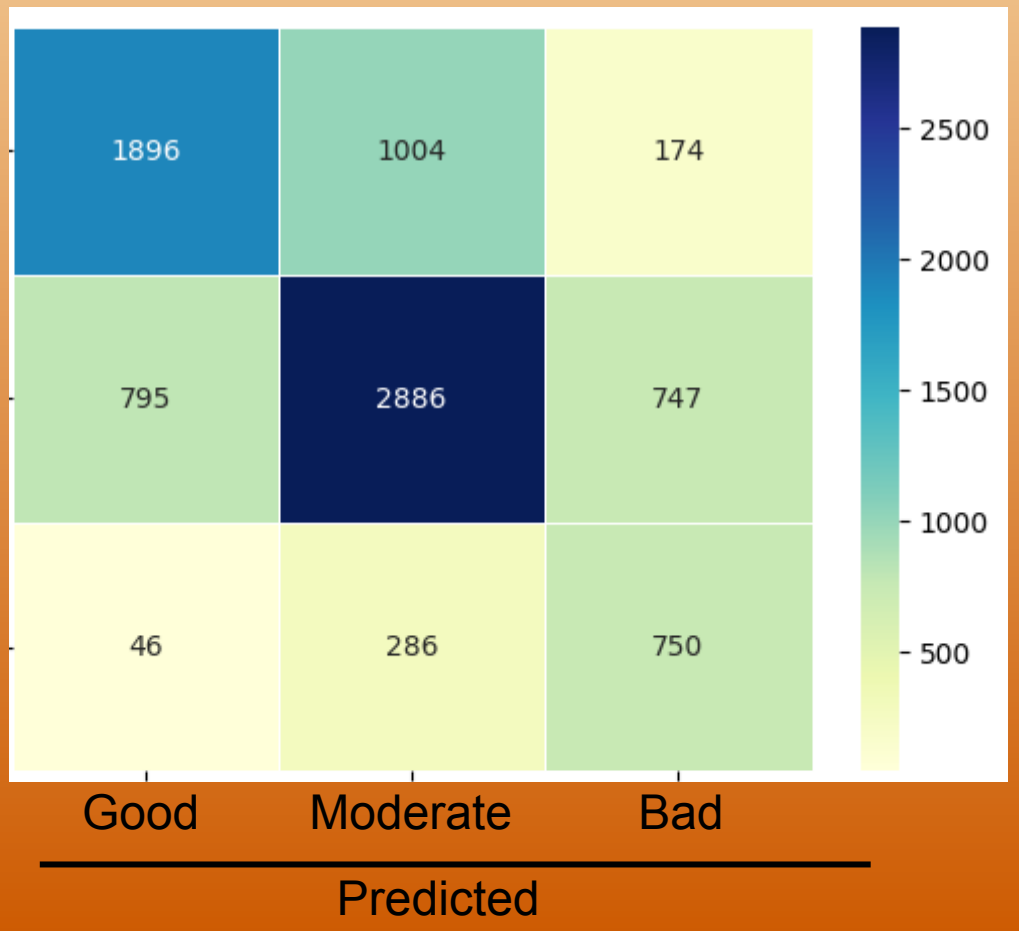
Error Analysis

KNN without oversampling

F1 score: 0.67 Kappa score: 0.44

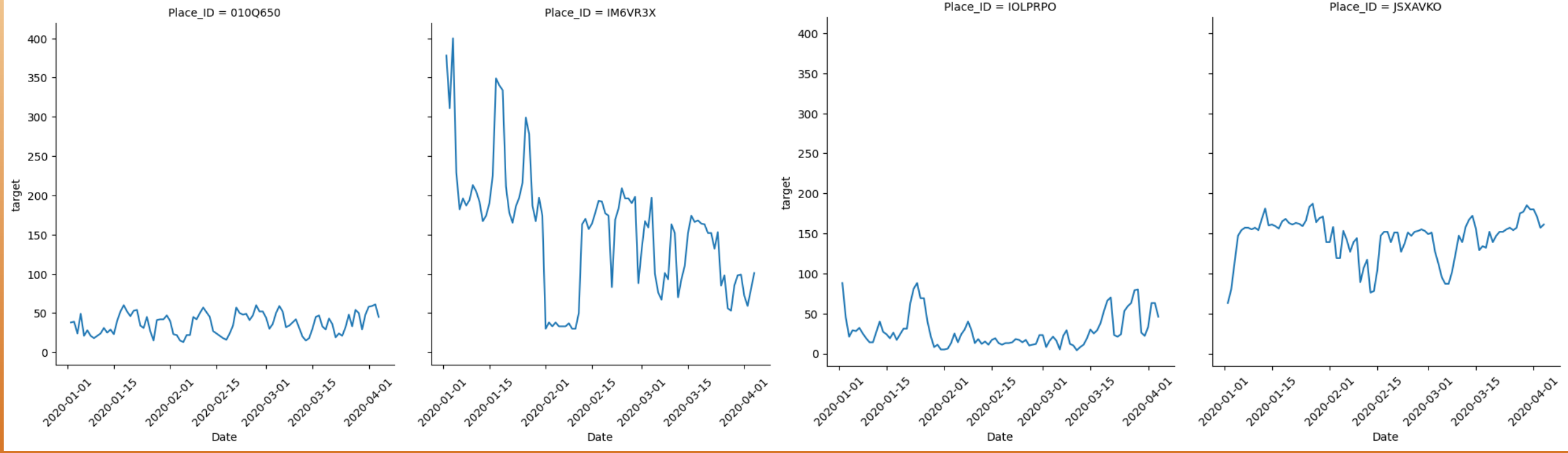
KNN with oversampling

F1 score: 0.64 Kappa score: 0.42



Error Analysis

Time and place are important explanatory features



Summary & Outlook

- Recommendation: KNN Classifier
  - F1 score: 0.67
  - Kappa score: 0.44 (Moderate agreement)
- Re-analyse with a time-series

  “We will learn that next week”

  - Lina Willing, 17.02.2023
- Original challenge: predict precise PM2.5 concentrations
  - Use RMSE as metric for Regression analysis

Competition Leaderboard

This is the final leaderboard. The competition is officially closed and will not accept any more submissions. Congratulations to all that participated, points were distributed on 20 April 2020.

RANK	USER	PUBLIC SCORE	PRIVATE SCORE
1	devnikhilmishra	29.13316471	26.09969719
2	CoviData Team	29.22652306	26.27988143
3	Klai	29.51718161	26.61821965

Thank you for your attention

Cohen‘s Kappa Score

$$k = \frac{(p0 - pe)}{(1 - pe)}$$

p_0: relative observed agreement among raters

p_e: hypothetical probability of chance agreement

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

https://www.statology.org/cohens-kappa-statistic/
18.02.2023