A breath of fresh air

ML Project

neuefische Data Science Bootcamp 1/2023

Christopher Hedemann, Stephen Kelly, Sarah Wiesner

The problem: PM2.5

- PM2.5 particulate matter concentration in the air
- diameter of less than 2.5 µm
- one of the most harmful air pollutants

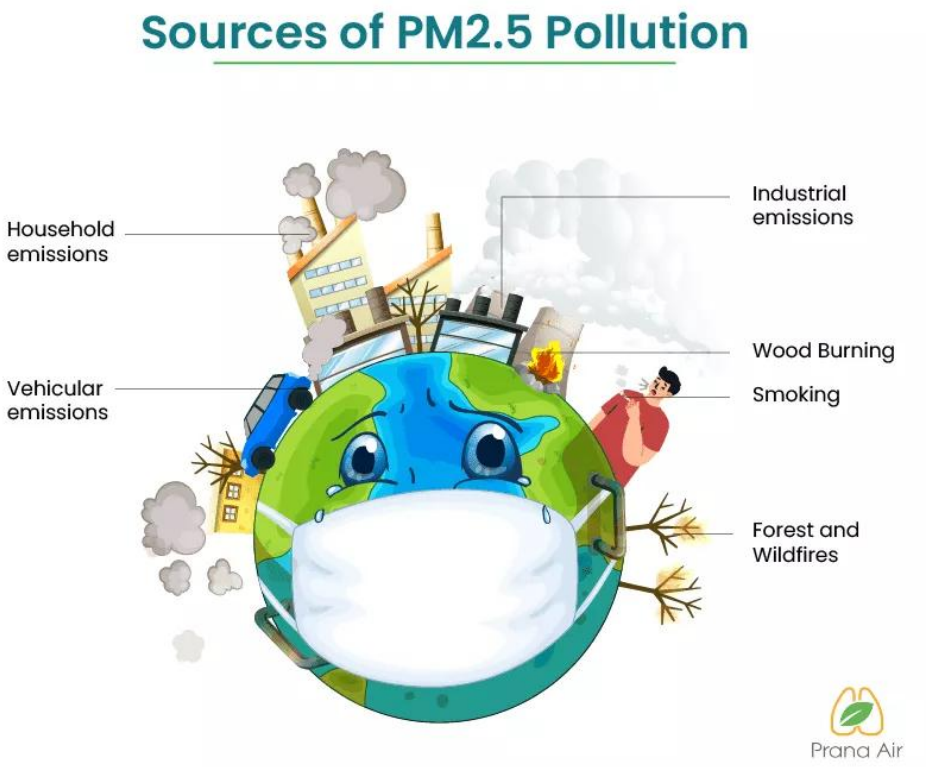


image source: www.pranaair.com

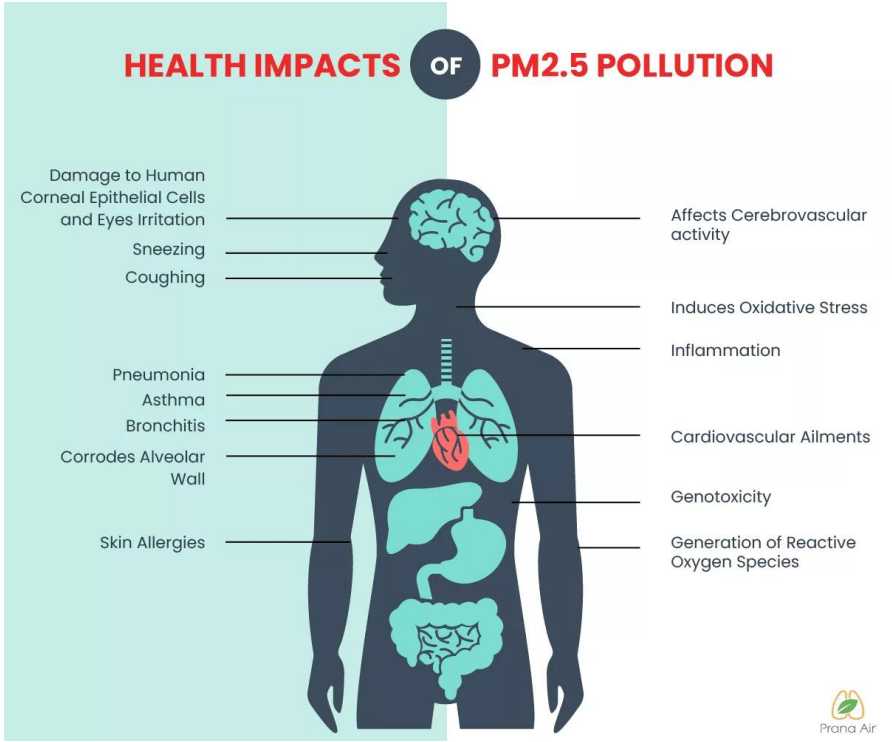


image source: www.pranaair.com

Our solution: a PM2.5 Warning System

- A **cheap estimate** of PM2.5 for global cities without PM2.5 sensors
- Based on meteorological parameters and atmospheric values derived from basic weather data and satellite data
- Use 'traffic light' warning system for PM2.5: Will it be harmful for me to leave the house today, for up to an hour?

Framed as a classification problem

Air quality category	$PM_{2.5}$ µg/m^3 averaged over 1 hour
Good	Less than 25
Fair	25–50
Poor	50–100
Very poor	100–300
Extremely poor	More than 300

image source: www.epa.vic.gov.au

PM2.5

Stay at home! > 50 µm/m^3

Be cautious / wear a mask. > 25 µm/m^3

Go out and have fun! <= 25 µm/m^3

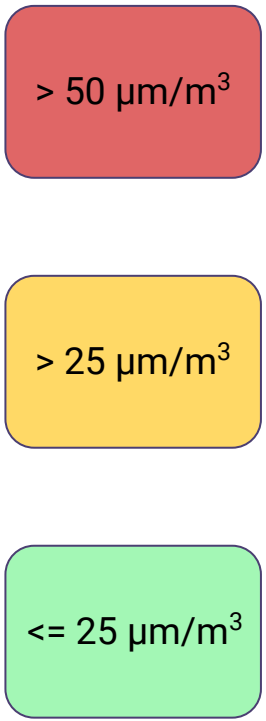
Data

- Dataset: Zindi “Urban air pollution challenge”
- Jan - Mar 2020
- Over 300 cities around the globe
- In-situ weather and Sentinel 5P satellite data
- We created new features: meteorological condition of yesterday, wind speed and direction, and day of the week

image source: www.esa.int

Evaluation metric and predictions with our baseline

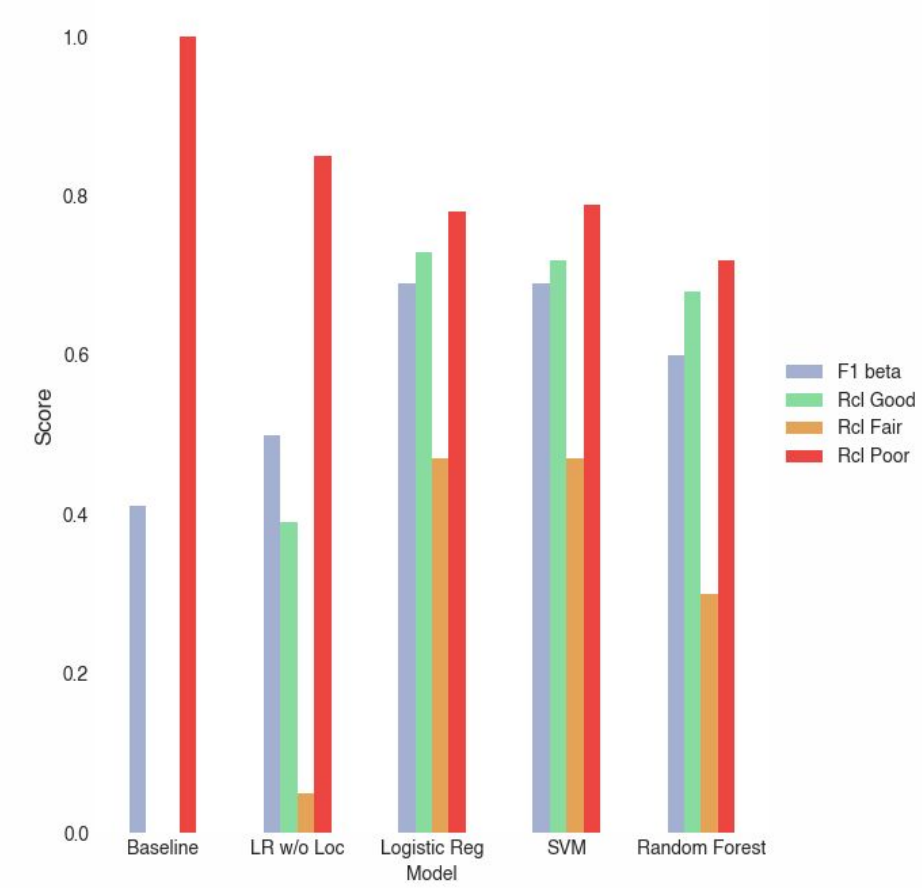
- Evaluation metric, **weighted F1** (beta=2): rewards correct prediction of as many **Poor** cases as possible (because this is the most numerous class), but partly penalises misclassification.
- Second metric, **recall**: How much of each class do I correctly predict?
- Both metrics ranges from 0 to 1, where 1 is perfect.
- Our Baseline ignores the data and always predicts **Poor**
  - F1-beta: **0.41**
  - Recall: Good: **0**, Fair: **0**, Poor: **1**.



Making predictions with better models

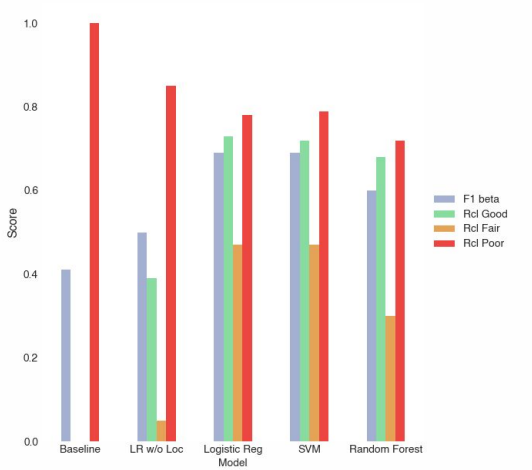
We used 3 machine learning models to check which performs best

- *Random Forest*: some improvement but poor prediction of **Fair**
- *Logistic Regression, SVC*: scored F1 of 0.69 (compared to 0.41 baseline) and better recall performance across all classes
- *Logistic regression* was faster, for predicting on the fly. Hyperparameter tuning with SVM can require several hours.



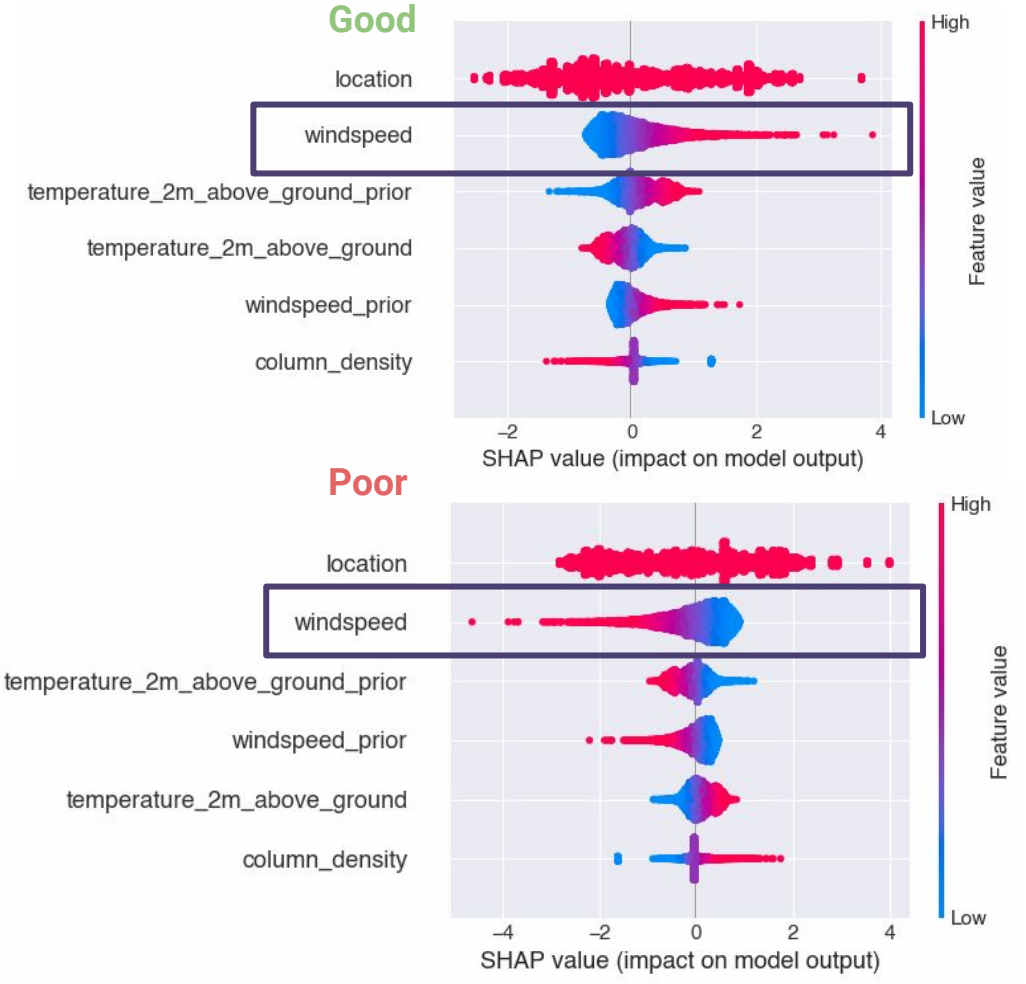
Findings

- *Location* of each city is an important predictor of PM2.5 levels, performance dropped considerably when we ignore location
- Historical weather & satellite data improves predictions of current PM2.5 Concentrations
- Similar predictions from Logistic Regression & Support Vector Machines
- Logistic Reg Model considerably faster processing speed
- Prediction of "**Good**" and "**Poor**" is easier than "**Fair**"



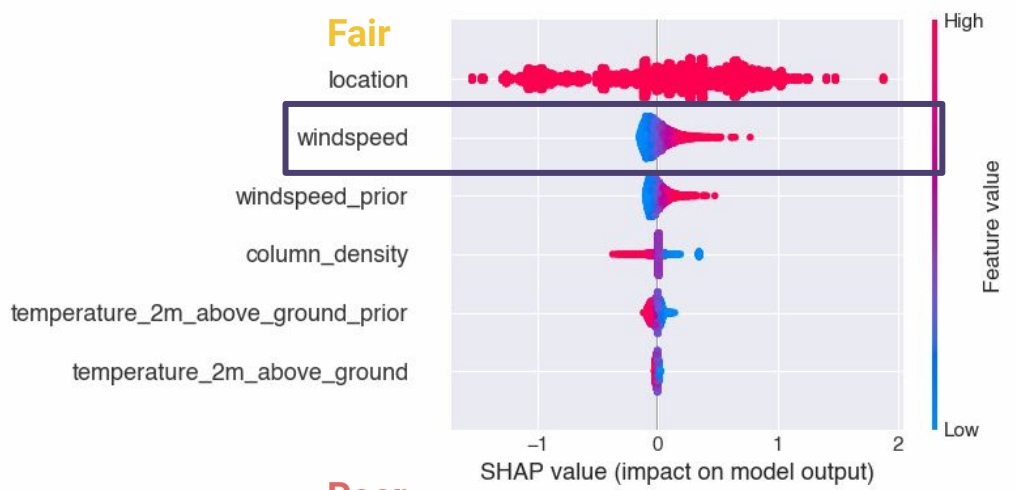
Why is Fair our poor performer?

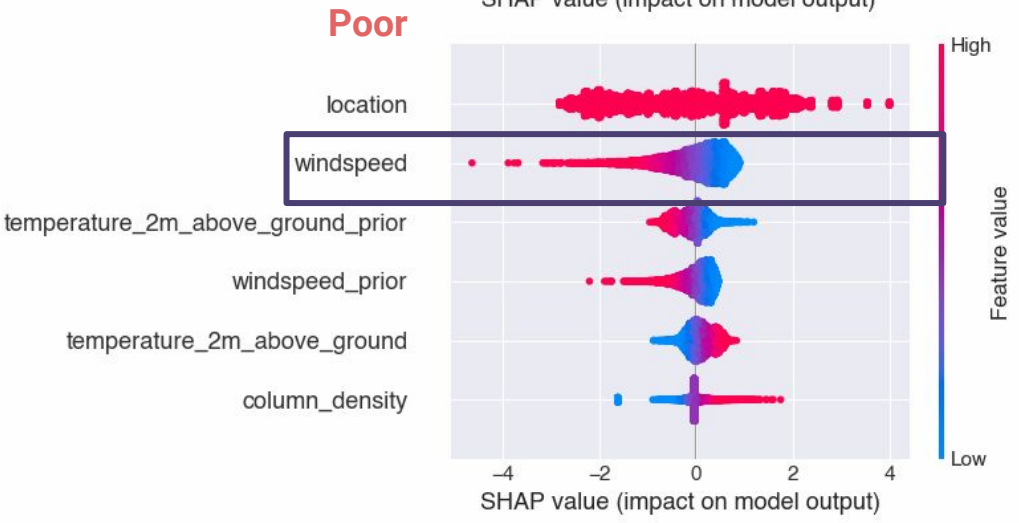
- Retrained our model using agglomerations of certain features, same F1 score
- A so-called SHAP analysis estimates the **importance of features** for predicting each class
- We see strong influencing effects for **wind speed**, **temperature above ground**, and **column density** of pollutants



Why is Fair our poor performer?

- Fair has weaker effects in all of these variables, it's torn between the opposing values that determine Good and Poor
- Fair doesn't really know who it wants to be!





Future Work

- Combine with other harmful pollutants, e.g. NOx
- Include PM2.5 actual values from the prior day(s) in predictions
- Include meteorological data with longer time frames of previous days
- Make it a dual-class problem, replacing Fair, with a “certainty” metric: how certain is the model prediction
- Take part in neuefische UX/UI bootcamp to develop a warning app