**Capstone Project Proposal: Predictive Modeling of Urban Air Quality**

1. **Problem**

Air pollution is a major cause of respiratory illness, lost productivity and premature death in both developed and developing countries. Globally, air pollution causes more than 5.5 million premature deaths each year[[1]](#footnote-1). Although China and India are responsible for the majority of these deaths, Europe and North America still struggle with localized smog and its associated health impacts[[2]](#footnote-2).

In developed countries, many local jurisdictions have monitoring programs in place to measure ambient concentrations of pollutants, identify when air quality degrades to a degree that may impact public health, and issue warnings to residents.

These warnings, such as “Ozone Action Days” or “Smog Alert”, are reasonably effective at modifying behavior and shielding vulnerable populations from the worst impacts, but typically can only be announced at most 24 hours in advance[[3]](#footnote-3).

With better predictive capabilities, local or regional authorities could issue warns with greater advance notice, thus allowing for more effective mitigation measures and protections for sensitive populations.

1. **Client**

The most likely client for this type of analysis would be a public sector agency with a mandate to regulate air quality and issue warnings, such as an air resources control board, air quality district or department of natural resources[[4]](#footnote-4).

In addition, non-governmental environmental groups lobbying for stricter air pollution regulations may have an interest in these findings. Finally, industrial firms and other operators of polluting equipment or facilities may seek to use predictive air quality modeling to forecast when short-term restrictions might go into effect that impact their operations.

1. **Data**

The University of California, Irvine (UCI) Machine Learning Repository contains a high quality dataset of 9,357 observations of ambient air quality in Milan, Italy. Although these data were originally collected for purposes of testing and calibrating a new sensor technology, they serve as an excellent longitudinal dataset of urban air quality.

Please note the data and resulting model for this project will be specific to conditions in Milan, Italy. However, the hope is to develop an approach that will be reproducible for other locations.

The observations in the dataset were taken hourly between 10 March 2004 and 4 April 2005.

The air pollutants measured are carbon monoxide, tin oxide, non-methane hydrocarbons, benzene, nitrogen oxides (NOx), tungsten oxide, nitrogen dioxide (NO2) and indium oxide. These are common categories of air pollutants found in urban areas, originating from fuel combustion and industrial processes, and contributing to ground-level ozone and visible “smog.”

In addition, the researchers’ sensors measured temperature, relative humidity and absolute humidity. These atmospheric conditions can play an important role in the formation of smog, as well.

1. **Approach**

After undertaking an exploratory analysis of the data, I intend to pursue two main lines of inquiry.

My primary goal is to develop an improved tool for long-term forecasting of air quality based on risk factors that can be known with certainty (e.g. time of day, day of the week, season of the year) or derived from outside the tool with reasonable accuracy (e.g. week-ahead weather forecast).

In this part of the project, I intend to test a number of methods, such as linear regression, decision trees and random forest, to determine the influence of these factors on air quality and produce the most accurate predictive model. I will need to train and test the model on distinct sub-sets of the data.

Also, I will need to use my domain knowledge and outside research to determine how each variable may need to be normalized, scaled or weighted. For example, each pollutant has its degree of toxicity, so straight comparisons may not be appropriate[[5]](#footnote-5) [[6]](#footnote-6).

The second line of inquiry will be focused on short-term forecasting, namely identifying leading indicators that suggest a particularly bad air pollution event is forming. Unlike long-term forecasting indicators, these data only emerge through real-time monitoring in the hours ahead of a peak air pollution event, but they can still be valuable from a policy and management perspective.

To search for leading indicators, I will isolate those days in the worst quintile for overall air quality, assess which traits they share, and identify at what time and for which pollutant(s) they diverge from other days in the dataset.

1. **Deliverables**

I intend to produce the following deliverables for this project.

1. A GitHub repository containing all code used to clean and analyze the dataset, as well as to perform the machine learning described above. Using a data visualization package (likely Seaborn), this repository will also contain the code to generate a number of plots.
2. I will describe the process and findings of this project in a paper. The paper will be detailed and well-documented, but not necessarily academic in style or format. An approximate length of 10 pages seems to be typical for a capstone project in this course.
3. Finally, I will produce a PowerPoint presentation to accompany a 15-20 minute oral presentation to be given via video conferencing tools to an audience of my peers and classmates. Again, the intent is to be detailed and thorough, without being overly technical in aspects not relevant to the data science course.

1. www.bbc.com/news/science-environment-35568249 [↑](#footnote-ref-1)
2. http://www.bbc.com/news/world-europe-38078488 [↑](#footnote-ref-2)
3. https://airnow.gov/index.cfm?action=airnow.actiondays [↑](#footnote-ref-3)
4. https://cfpub.epa.gov/rblc/index.cfm?action=Agency.AgencyLinks [↑](#footnote-ref-4)
5. https://www.epa.gov/haps [↑](#footnote-ref-5)
6. https://web.stanford.edu/class/e297c/trade\_environment/energy/hsmog.html [↑](#footnote-ref-6)