**API-222B Machine Learning Final Project**

Group #5: Jonathan Hudgins, Marina Kusumawardhani, Januardy Djong, Moustapha Kasse

**Issue**

Can Machine Learning be used to better Predict PM2.5 levels of Air Pollution?

**Introduction**

Air pollution is a leading environmental threat to human health, and in 2015, it was estimated to be responsible for 6.4 million deaths worldwide.[[1]](#footnote-1) Particulate Matter 2.5 (PM2.5) can play a role in causing severe illnesses because they are small enough to be inhaled deep into the lungs.[[2]](#footnote-2) The Center for Disease Control (CDC) claims that lowering PM levels would prevent deaths, mostly from heart attacks and heart disease.[[3]](#footnote-3) This project intends to utilize machine learning techniques to better predict PM2.5 levels by analyzing atmospheric data gathered over four years from an observation point in Beijing, China.

**Motivation and Research**

Unlike meteorological data observed over the last hundred years, air quality measurements and air pollution has only recently been observed and measured regularly. Modeling and prediction of air pollution have only recently been utilized effectively through government meteorological institutions and companies like AirVisual. Given that PM2.5 is classified as the most dangerous form of particulate matter, we set out to find a robust dataset that could be utilized with machine learning techniques to better understand the correlation between meteorological conditions and air quality. Specifically, we needed a dataset that included both meteorological conditions and the six primary air pollutants that make up an Air Quality Index calculation: Particulate Matter 2.5, Particulate Matter 10, Sulfur Dioxide, Nitrogen Dioxide, Carbon Monoxide, and Ozone. We also needed a dataset from a country/city where air pollution is a continuing problem. Thankfully the UC Irvine Machine Learning Repository recently published multiple datasets from observations taken over four years in Beijing, China.[[4]](#footnote-4) Just as predicting the weather is critical for business, travel, recreation, comfort, and safety, we believe that precise PM2.5 predictions can help save lives by providing accurate air quality information that people can plan their lives around. Pollution predictions can give useful information for daily decisions such as: how best to commute to work, or work from home, buy and wear a mask, exercise indoors, have kids play indoors versus outdoors, vacation planning, etc.

In researching this issue, we found that U.S. Embassies around the world are gathering and reporting air pollution levels for public use, particularly for the health of U.S. citizens living overseas. However, the U.S. government reporting on air quality issues within other countries has been a politically contentious issue in places like India and China.[[5]](#footnote-5) Additionally, it appears that U.S. Embassies are using real-time observation data gathered from sensors located on their buildings. Based on the politically sensitive nature of this issue, we believe the State Department could use a machine learning air pollution model to provide more accurate readings of PM2.5 levels in addition to real-time observations.

**The Data**

The UC Irvine Machine Learning Repository dataset includes atmospheric observations taken over four years beginning in March 2013. Atmospheric observations were made at 1-hour intervals providing over 35,000 total observations. The dataset contains 16 useful variables, including six pollution and six meteorological variables (2 other variables are row number and station - we only use data from one station). The raw dataset is relatively clean, with 5% or less missing observations. The observations were recorded from a station in Beijing, China called Aotizhongxin. **Appendix 1** shows details on the 16 variables.

**Data Cleaning**

As mentioned above, the raw dataset is relatively clean and complete. Hence, only minimum data engineering is needed. To predict PM2.5, we used all available predictors including 4 date variables, 5 variables on other pollutants, and 6 meteorological variables. Additionally, we performed two methods of feature engineering on the data.

First, we impute missing values for the predictors using a Random Forest method. The assumption here is that the data is missing at random. However, we omitted the data for which PM2.5 value is missing. Since the number of missing PM2.5 data is low and appears to be spread out evenly over time, it might not affect the overall result significantly.

Second, we created a categorical variable to classify PM2.5 based on the standard Air Quality Index (AQI) categories shown in **Appendix 2**. We first treat PM2.5 prediction as a regression problem and measure the test MSE of the model. Then, we convert the true and predicted PM2.5 into AQI categories and re-evaluate the model based on the classification accuracy. This is because AQI categories is more interpretable for people. We convert the PM2.5 data directly to AQI categories instead of the index number, but the AQI number can also be obtained if required during the implementation.

**Machine Learning Methods**

We tried various machine learning methods with the data, namely Linear Regression, KNN Regression, Lasso, Random Forest and Boosting. However, linear regression gave poor performance since the relationships are not linear, judging from some of the scatterplot results (between PM2.5 and Temperature, Pressure and Dewp). Lasso regression performs poorly as well because of the same reason (the relationships are not linear). KNN Regression (with k=5 from Cross Validation) performs reasonably well, but cross validation requires computing effort. Random Forest performs better than KNN but with even slower computing. Boosting gives the best performance and also with much faster computing, therefore we chose Boosting as our method (we use specific implementation of Boosting dubbed “extreme gradient boosting” or xgboost).

Boosting is one of the ensemble learning methods to improve the predictive power of trees. The main idea is that trees are formed in a sequential way (in contrast with Bagging where trees are built without considering other trees) in which each tree is grown using information from previously grown trees. In other words, it produces a prediction model in the form of an ensemble of weak prediction models (or trees), and it generalizes them by allowing optimization of differentiable loss function. What makes extreme gradient boosting unique to general boosting is that it uses a more regularized model formalization, which is used to control/reduce overfitting and gives better performance. We found that using xgboost implementation improves model prediction compared to random forest and regular gradient boosting implementation (gbm). It also performs much faster because it allows parallel computation on a single machine.

Another method providing similar returns to xgboosting (but with a significantly slower process) is Random Forest. Random Forest is another Ensemble Learning method such as Boosting and Bagging; however, in Random Forest, we decorrelate the bagged trees to gain reduction in variance. Since most committees in Bagging method always end up using the very strong predictor in the first split, the committees always end up with very similar trees with no reduction in variance compared to a single tree. Therefore by decorrelating, the variance can be reduced with a Random Forest model.

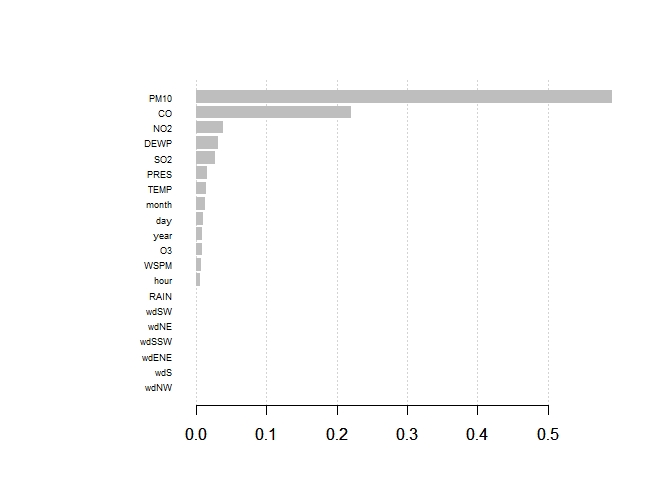
**Results**

We ran the model on 1) both predictors (weather and pollutants) and 2) on the weather data only. We did this to better understand whether the level of PM2.5 is influenced more by other pollutants or just by the weather data. (The step by step of the xgboosting method are inside the R file, and is not included here because of paper length limitations).

**Model 1: using all available predictors**

Result:

Model **MSE = ~259**, model **accuracy** using AQI categories **= 82%**

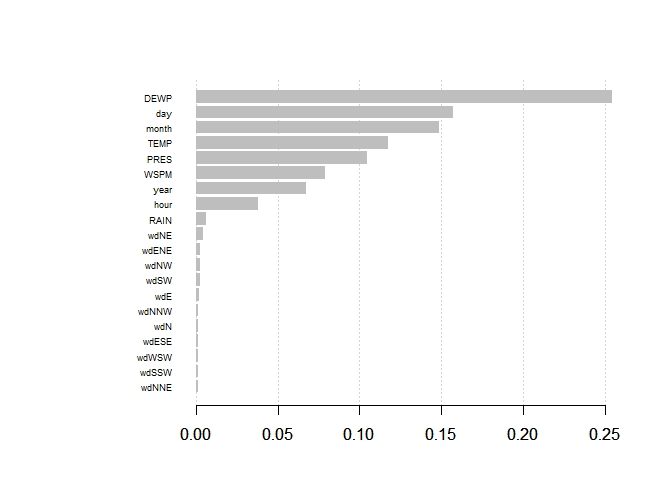
Relative importance chart of each variable: 

**Model 2: using only weather data**

Result:

Model **MSE = ~808**, model **accuracy = 66%**

Relative importance of each variable:



**Conclusions drawn from ML Methods**

From Model 1, we can conclude that pollutants serve as better predictive variables of the level of PM2.5 than the weather data. This is shown by 4 out of 5 pollutants (apart from O3) placed quite highly in the variables with relative importance. Apart from the pollutants, Dewpoints serve as the only weather data with relative importance.

We also tried to find out which of the weather data serve as most important predictors. Consistent with the findings from Model 1, Model 2 shows that Dewpoint serves as the most important predictor among the other weather predictors. Interestingly, Rainfall and Wind Direction seem to be insignificant predictors.

A screenshot of a cell phone

Description automatically generated

This finding is consistent with other research findings, such as the one shown in the graphic above, being observed in Lanzhou, China[[6]](#footnote-6). It shows that the drizzle or moderate rain actually does not change PM2.5 levels.

We found that the accuracy level is quite good in Model 1, reaching 82%. This means that the prediction largely falls into the right AQI (Air Quality Index) as shown in the Confusion Matrix Model 1. The matrix shows that most Air Quality Index categories predictions are relatively accurate, with >80% accuracy. However, the “Unhealthy for Sensitive Groups” category is particularly more challenging to predict. This is because the AQI conversion range from PM2.5 is not uniform throughout the scale. For categories with a narrower range, the predicted PM2.5 result has a higher chance to deviate from the true category.

Accuracy level is lower in Model 2, also with higher number of MSE, meaning that the weather data has lower performance in predicting PM2.5 compared to the pollutants. The AQI category prediction is most problematic again for the “Unhealthy for Sensitive Groups” category with a mere 36% accuracy. The confusion matrix suggests that our model tends to underestimate the AQI categorization, skewing the results to “Moderate.” For implementation, we should address this issue since prediction biased towards “Moderate” could harm people who might have a health condition and rely on the prediction. In other words, overestimating the AQI categories to the more unhealthy side is more conservative and likely better for the purpose of protecting public health. One way could be to slightly decrease the range of “Moderate” category in the predicted result to allow for skew toward the more conservative categorization.

As stated before, we need to take into account some bias and fairness in this model. These findings are only sourced from one location near Beijing, therefore the model might not be generalizable to other cities. To have better prediction, we need to try to include data from other cities or research, such as the one being done in Lanzhou. With multiple sources of data, we can confirm the accuracy and transferability of our model.

**How this Model is Different**

A common method to address pollution levels is the collection of real-time observation from expensive air quality monitoring equipment, similar to the sensors used in U.S. Embassies all over the world. Without software to analyze the collected data, this does not produce particulate matter level predictions, which would allow people to make important decisions on how to manage it (commute, wearing masks, etc.). Our prediction algorithm, if implemented, would not only produce useful information that would give people time to make decisions, but could even help cut costs by replacing the Embassy’s equipment if the necessary weather and pollution data are readily available.

On a different note, while pollution levels in cities are modulated by meteorological factors, different parameters have heterogeneous effects throughout the year depending on the conditions (seasons, altitude, terrain complexity, etc.)[[7]](#footnote-7). This creates a need for complex forecasting algorithms that can take into account a wide variety of variables. As a result, A major approach now used to predict PM2.5 is an artificial neural network (ANN), an intelligent system that has the capability of learning, adapting and creating internal connections among meteorological data. ANNs are appropriate models for highly nonlinear modeling, non-structured data and when no prior knowledge about the relationship between the parameters is assumed (e.g. text, image, etc). However, given there is no issue with the representation of this dataset, our xgboosting algorithm is sufficient to predict PM2.5 levels, as proven by the accuracy level.

**Recommendations for Implementation**

This analysis demonstrates that the use of our model based on the extreme gradient boosting method is relevant to predict PM2.5 levels from weather and pollution data in one location near Beijing. While our algorithm might not yet be generalizable to other locations, we think it would be beneficial to the U.S. Embassy to generate accurate predictions of PM2.5 levels as a supplement (or potentially as a cost-cutting replacement) to the real-time observations gathered from their sensory equipment. However, this implementation would require the Embassy to acquire more meteorological observation instruments to acquire the necessary variables in addition to software that would integrate with the observation hardware to run the machine learning algorithms for prediction. The other alternative is to use commercial vendors like AirVisual for the software, but they would need access to meteorological data from the Embassy that is currently unavailable. On a larger scale, once the transferability of our model is confirmed with the support of varying datasets from other cities, our prediction model could be utilized by the Chinese government to manage pollution levels in areas where meteorological measuring equipment is available but PM2.5 observation equipment may not be. However, this would also be challenging due to the political sensitivities of air pollution issues in China.

**Appendix 1:** List of variables used in the Aotizhongxin pollution data.

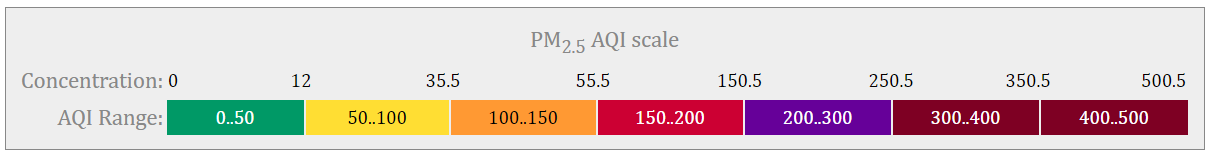
Number of observations: 35,064

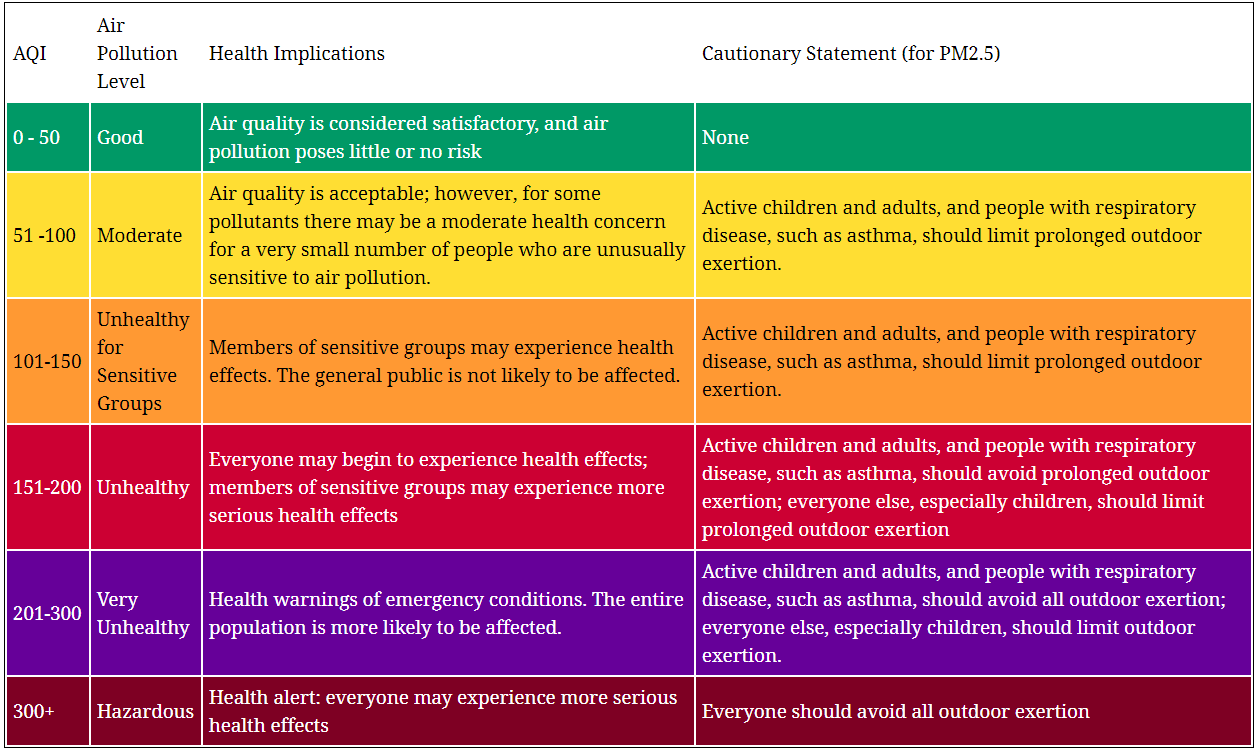
Dependent variable: PM2.5

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definition (unit of measurement)** | **Data type** | **% missing** |
| year | year of data in this row | numeric | 0 |
| month | month of data in this row | numeric | 0 |
| day | day of data in this row | numeric | 0 |
| hour | hour of data in this row | numeric | 0 |
| PM2.5 | PM2.5 concentration (ug/m^3) | numeric | 2.64 |
| PM10 | PM10 concentration (ug/m^3) | numeric | 2.05 |
| SO2 | SO2 concentration (ug/m^3) | numeric | 2.67 |
| NO2 | NO2 concentration (ug/m^3) | numeric | 2.92 |
| CO | CO concentration (ug/m^3) | numeric | 5.07 |
| O3 | O3 concentration (ug/m^3) | numeric | 4.9 |
| TEMP | temperature (degree Celsius) | numeric | 0.06 |
| PRES | pressure (hPa) | numeric | 0.06 |
| DEWP | dew point temperature (degree Celsius) | numeric | 0.06 |
| RAIN | precipitation (mm) | numeric | 0.06 |
| wd | wind direction | categorical | 0.23 |
| WSPM | wind speed (m/s) | numeric | 0.04 |

Note: We choose to use the date/time data (year, month, day, hour) as numerical variables instead of factor to speed up the computing time. For month, day, and hour this might be a decent assumption as pollution might be cyclical (i.e., affected by season, operating times of polluters, etc). For “year” using it as factor variables might be more appropriate to control for “shock” specific to the year, such as change in government policy that mandates stricter environmental standards. We did not find significant improvement running the model using these variables as factor and hence decided to keep them as numerical.

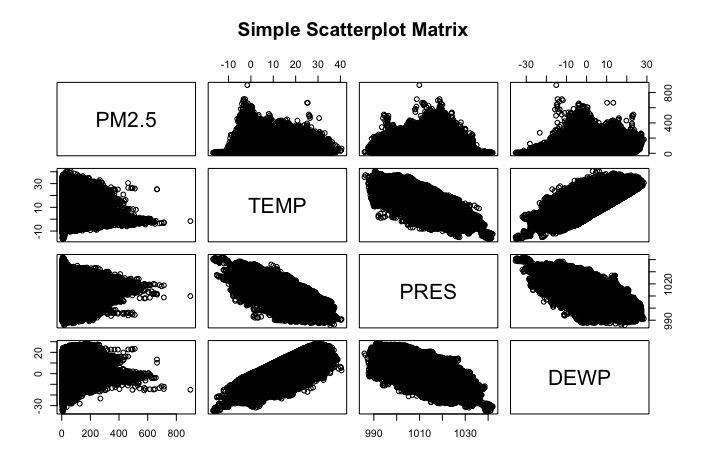
**Appendix 2:** PM2.5 AQI scale conversion

****

****

Source: <https://aqicn.org/calculator>

**Appendix 3**



**Appendix 4**

Confusion matrix of Model 1:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Prediction** | **Good** | **Moderate** | **Unhealthy for sensitive groups** | **Unhealthy** | **Very unhealthy** | **Hazardous** | **%**  **Accuracy** |
| Good | **620** | 144 | 0 | 0 | 0 | 0 | 81% |
| Moderate | 167 | **1276** | 118 | 11 | 0 | 0 | 81% |
| Unhealthy for sensitive groups | 0 | 182 | **559** | 141 | 0 | 1 | 63% |
| Unhealthy | 2 | 26 | 198 | **2197** | 102 | 1 | 87% |
| Very unhealthy | 0 | 0 | 0 | 89 | **663** | 29 | 85% |
| Hazardous | 0 | 0 | 0 | 2 | 21 | **279** | 92% |

**Appendix 5**

Confusion matrix of Model 2:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Prediction** | **Good** | **Moderate** | **Unhealthy for sensitive groups** | **Unhealthy** | **Very unhealthy** | **Hazardous** | **%**  **Accuracy** |
| Good | **336** | 128 | 1 | 0 | 0 | 0 | 72% |
| Moderate | 378 | **916** | 131 | 29 | 0 | 0 | 63% |
| Unhealthy for sensitive groups | 60 | 438 | **389** | 192 | 4 | 0 | 36% |
| Unhealthy | 13 | 146 | 351 | **2123** | 235 | 7 | 74% |
| Very unhealthy | 2 | 0 | 1 | 93 | **526** | 96 | 73% |
| Hazardous | 0 | 0 | 2 | 3 | 21 | **207** | 89% |

1. “Can Air Pollution Be Predicted?,” accessed November 16, 2019, https://www.iqair.com/us/blog/air-quality/can-air-pollution-be-predicted. [↑](#footnote-ref-1)
2. “Environments Tracking Air Quality - Health Impacts of Fine Particles in Air - CDC Tracking Network,” accessed November 16, 2019, https://ephtracking.cdc.gov/showAirHIA.action. [↑](#footnote-ref-2)
3. “Environments Tracking Air Quality - Health Impacts of Fine Particles in Air - CDC Tracking Network.” [↑](#footnote-ref-3)
4. “UCI Machine Learning Repository,” accessed November 30, 2019, https://archive.ics.uci.edu/ml/index.php. [↑](#footnote-ref-4)
5. Eli KintischApr. 19, 2018, and 10:00 Am, “Rooftop Sensors on U.S. Embassies Are Warning the World about ‘Crazy Bad’ Air Pollution,” Science | AAAS, April 19, 2018, http://www.sciencemag.org/news/2018/04/rooftop-sensors-us-embassies-are-warning-world-about-crazy-bad-air-pollution. [↑](#footnote-ref-5)
6. Feng, X., Wang, S., “Influence of Different Weather Events on Concentrations of Particulate Matter with Different Sizes in Lanzhou, China”, J Environ Sci, China, 2012. [↑](#footnote-ref-6)
7. Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk, “Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters,” Journal of Electrical and Computer Engineering, vol. 2017 [↑](#footnote-ref-7)