# Predicting PM 2.5 Index and Web Application Interface Developing

Lu Yin ly1123
Ruofan Wang rw2268
Shenghui Zhou sz2396

***Abstract*—Our team aims to make a precise prediction to the value of PM2.5 index in Guangzhou in China from the data set of PM2.5 measurement records in five years with 17 relative record features. In our project, Multi-linear regression, Decision Tree, Random Forest, XGboost and ARIMA are applied to fit the data set with promising results. On top of that, we construct a local multi-threaded asynchronous TCP server in Python, which gives precise predictions of PM2.5 value in the three following days and provides precise predictions of PM2.5 in Guangzhou.**

**Keyword: Multiple Linear Regression, Decision Tree, Random Forest, XGboost, ARIMA, Web Application Development**

## I. Introduction

### A. Motivation

In contemporary society, the environment problems especially heavy air pollution become one of the most severe menace for peoples health in China. Within a typical year in China, the haze (also called smog) weather accounts for almost 80 percent days in a year. In particular, the main pollution material is called PM 2.5, which as its name suggests has a diameter less than $2.5\mu m$ and can be easily inhaled into humans body and cause serious disease. Therefore, as a group of Chinese graduate students in NYU, our team tries to predict the future PM 2.5 index and to design a simple web interface to help people to plan their activities healthily and wisely.

### B. Goal

The goal of this project are two folds:

- Observe the PM 2.5 data from the online resources and then fits the data into the several models. We choose the model based on the MSE of the model
- Design a web application and interface by using Python in order to let users to plan their activities wisely.

## II. Data Visualization and Analysis

### A. Data understanding and visualization

In order to make a more precise prediction of PM measurement, we consider the data set released by University of California at Irvine Machine Learning Repository[3], which contains the PM2.5 measurement records starting from 2010 to the end of 2015 from three weather inspection stations that located in Guangzhou. The data set contains 52,584 rows and 17 features that helps to measure the value of PM2.5. However, there is roughly 60% of missing values of the data features in the data, which makes us to do data cleaning before we start to analyze the data set. The table below shows the missing values among the mentioned features in the data set and be denoted as NA.

TABLE I
Table of NA counts among the features

| Para | NA count | Para | NA count |
|---|---|---|---|
| $No$ | 0 | $year$ | 0 |
| $month$ | 0 | $day$ | 0 |
| $hour$ | 0 | $season$ | 1 |
| $PMCS$ | 20232 | $PM5MS$ | 31489 |
| $PMUSPost$ | 20232 | $DEWP$ | 1 |
| $HUMI$ | 1 | $PRES$ | 1 |
| $TEMP$ | 1 | $cbwd$ | 1 |
| $Iws$ | 1 | $precip$ | 1 |
| $Iprec$ | 1 | $dtype$ : | int64 |
| $gzshape$ | (52584, 17) | | |
| $NApercent$ | 0.5988 | | |

We plot three figures that showing the PM2.5 distribution in three stations, which are the city station,5th Middle School and US Post, across month within the five years. From the plots

we can find similar distribution of PM2.5. Since the location of 5th Middle School shares the most popularity and data with less missing, we decide to make prediction that focus on the data collection from this particular area.

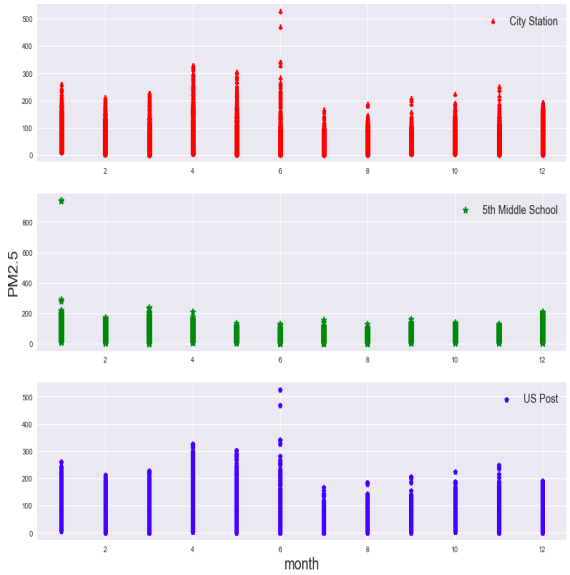


Fig. 1. PM2.5 distribution among three stations

### B. Data processing

We find two severe problems when we plot the distribution of the PM2.5 in the raw data set, the figure is shown below:

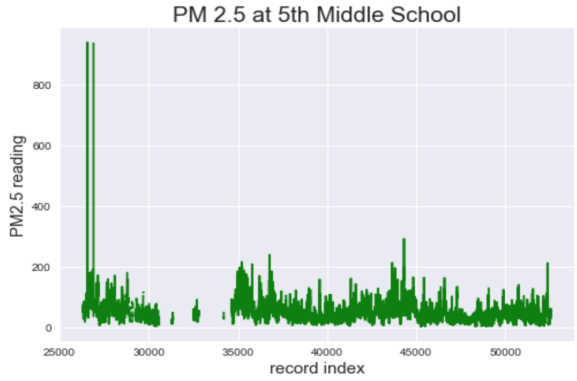


Fig. 2. PM2.5 distribution at 5th middle school (raw data)

- Many extrema outliers in the data set can be observed in the data set. The outliers are fatal to our analysis when we deliver our model and likely to lead to inaccurate results of prediction. In order to avoid this problem, we standardized our data and smooth our data with moving median. Figure shows the distribution of PM2.5 after smoothing and standardizing.

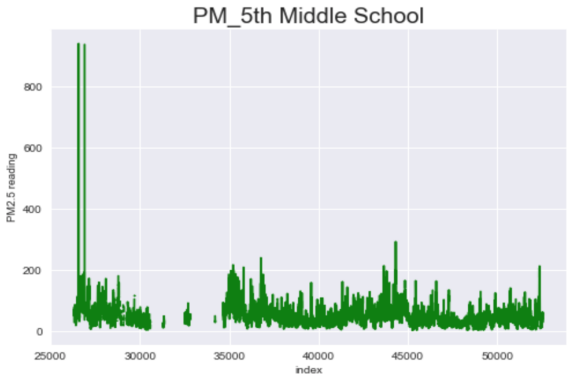


Fig. 3. PM2.5 distribution at 5th middle school (data without outliers)

- A clear time period gap is shown in the data set We can observe a time slot which contains nearly no data with the index number from 30000 to 35000 but relatively consistent afterwards. There will be bad result if we build time series model with large gap missing in the data set. Hence, we remove all the data with the index before 35000 and eventually get a data set with consistent records by time.

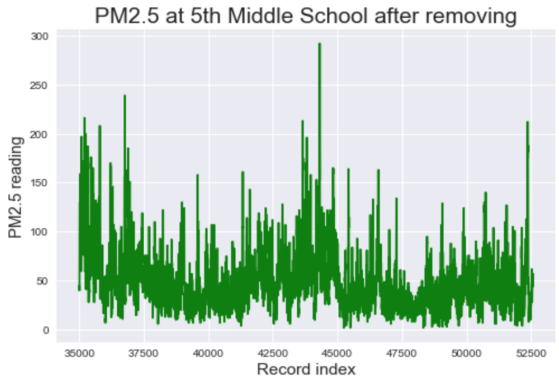


Fig. 4. PM2.5 distribution at 5th middle school (cleaned data)

The figure below shows the data distribution after all data processing

## III. Model

Regression model

*1) Multi-linear Regression:* At the beginning of the data modeling, we decide to implement Multi-linear Regression(MLR) as our baseline model since all the important feature in our data set are time related. From our expectation there is a strong relationship between our target variable with the features along time. We will concentrate on the value of MSE when we estimate the model we build and deliver our

models to improve our analysis. We firstly drop some features to reduce collinearity before we build multi-linear regression:

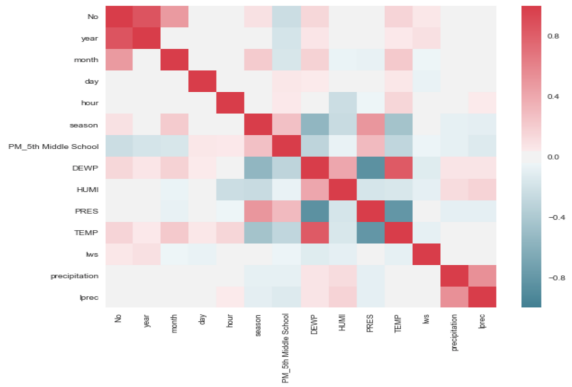


Fig. 5. Correlations between features

The comparison between the prediction values and actual values is shown below figure with the value of MSE as 622.9357:

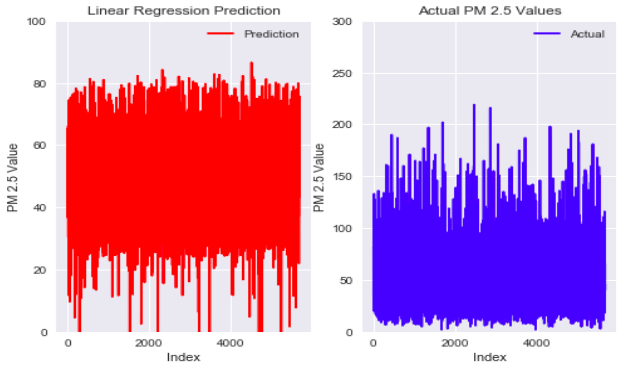


Fig. 6. Comparison between the prediction values and actual values

*2) Tree based models:* The low performance of multi-linear regression shows that there is no necessarily linear relation between target variable and features along. Hence, we trialed tree based models in our report to get a better result.

We delivered several methods of tree based models ranging from Decision Tree, Random Forest and eventually find XG-Boost Regressor as the best tree based model by comparing MSE of each model with 108.7630.

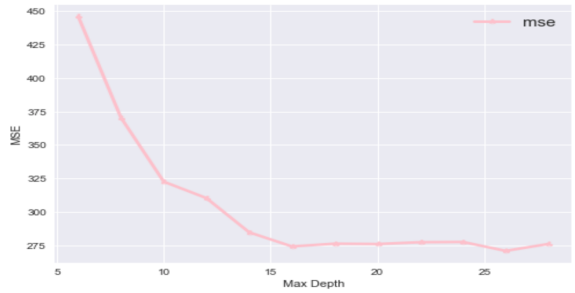


Fig. 7. MSE by max depth on Decision Tree

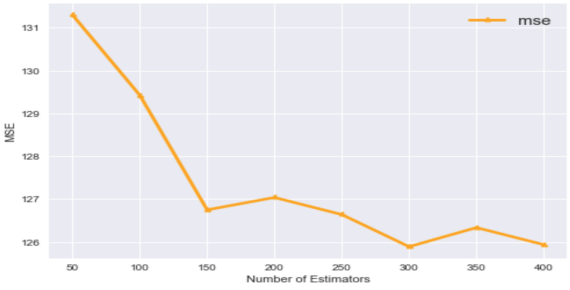


Fig. 8. MSE by max depth on Random Forest

### A. *Time series model*

For the purpose of getting more promising model, we plan to apply time series model which perform better with time series data.

*1) ARIMA Model:* For the purpose of getting more promising model, we plan to apply time series model which performs better with time series data. In the ARIMA model, we firstly need to build a linear regression model with time lagged predictors. After testing, we find that some weather conditions are observed has a higher relationship with present PM2.5 index.

We take lagged humidity as an example. The figure shows strong correlation between PM2.5 and lagged humidity against different time lag (in hours) and four days lagged humidity is more suitable to predict PM2.5. Furthermore, there are several features like pressure, accumulated precipitation and wind direction and speed were also included into linear regression model with lagged humidity. The check of residual is shown below which shows as stationary. Furthermore, we also plot a QQ plot that prove the validation of our assumption of the regression model.

On top of that, we build ARIMA to fit our data set. From the plot we can find that residuals are now stationary. We identify the optimal parameters for ARIMA model according to the

AIC results, and finally select ARIMA(1,0,1) $*$ (0,1,1,12) as the best fit for our data set. The result can be find in the following figure and AIC with the value of 121031.0877 which is relatively small.

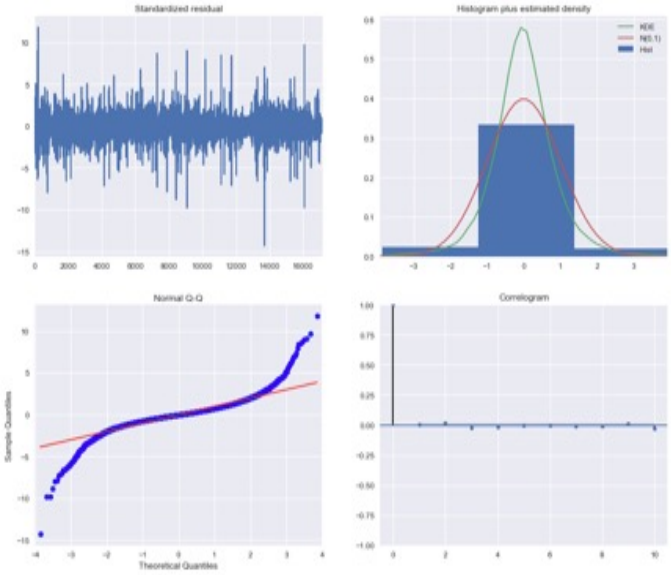


Fig. 9. results of ARIMA

## IV. Website Design and Optimization

As the final goal of this project, our team designs a web application to show the predictions of PM 2.5 to potential customers. More importantly, we implement the website with different labels to realize our business goals.

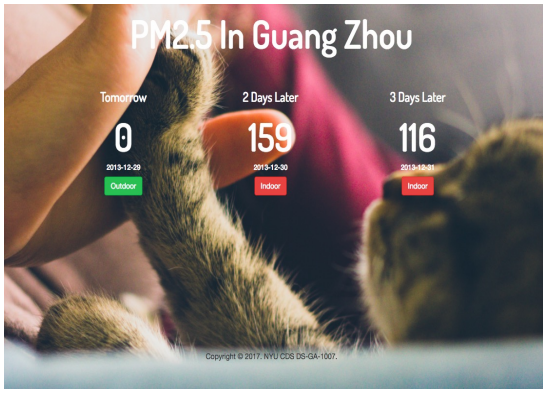


Fig. 10. Interactive website

If the PM2.5 Index is higher than 100, it turns out to be red. Index higher than 100 means the outdoor activities could be harmful for human bodies. Therefore, the button would direct the users to those sellers who offer healthy products to clean air and absorb harmful PM2.5.

On the other hand, the green button means index of PM 2.5 is lower than 100 and the outdoor activities are recommended. Then, this button will direct the users to workout applications and websites for outdoor activities. Here, it leads users to the TripAdvisor Guangzhou site.

With these implementations of the predictions, our team could combine our results with the real-world business problems and help people manage their indoor and outdoor activities accordingly .

## V. Conclusion

For this project, we first use the multilinear regression model as our baseline model. By doing the linear regression models, we identify several significant features. However, the linear models have poor performance when fitting the time lagged data. Therefore, start from there, our team turns to explore tree-based models. As the result shows, the tree based model much better than the linear models. The method we use is to combine several tree models with different weights. On top of that, our team explore the time lagged data by using time series models, the ARIMA Model. Based on the evaluation of MSE for all those models, we choose the XGBoost regression as our the optimal model.

By using the ARIMA model, our team first uses the linear model to check the residual value of each time points. As our result shows, the residue is stationary. Then by comparing the parameters, our team finally choose the order of (1,1,1) as our best fit model.

Getting the good prediction result is not our final goal. As a group of ambitious graduate students, we try to make the business value out of our predictions. By designing the Website, we direct the users to different websites based on different PM 2.5 predictions and realize the business opportunity.

## References

[1] Brook RD, Rajagopalan S and Pope CA 3rd, *Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association.*, US National Library of Medicine National Institutes of Health, 2010.

[2] Yu-Fei Xing, Yue-Hua Xu and Min-Hua Shi, *The Impact of PM2.5 on the Human Respiratory System*, Journal of Thoracic Disease, 2016.

[3] Data Source, University of California: Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities