5002 Project Report

Chen Donger 20551304, Wu Xiting 20549143
Group 59

1. Introduction

Over the past years, there is increasingly severe air quality problem in Beijing. As we know, air pollutant like PM2.5 is harmful to human's body. Avoiding inhaling air contaminants is the key to preventing smog damage to body. Therefore, it is important to predict PM2.5 and other air particles. In this project, we are requested to predict concentration levels of several pollutants over the coming 48 hours for 35 stations in Beijing, China. Air quality changes very quickly and pollutant will spread to surrounding area with complex spatial dependence, which is the difficulty of this project. Following figure shows the process that we solve this problem.

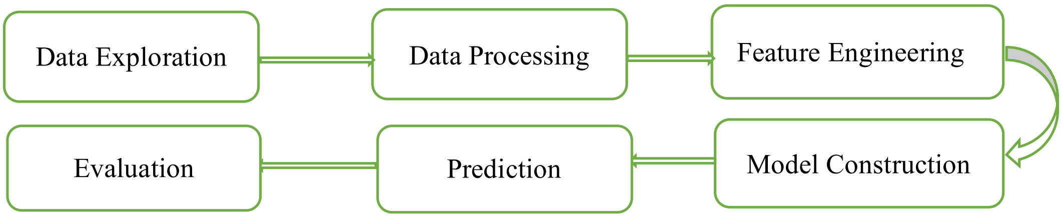


Figure1.1 The process of the project

2. Data Exploration

2.1 Scatter plot between label and feature

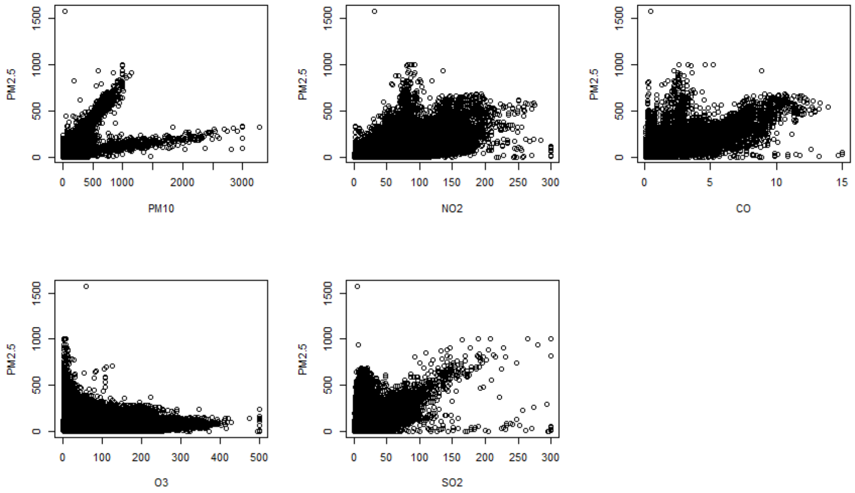


Figure2.1 The Scatter Plot of PM2.5 with other pollutant

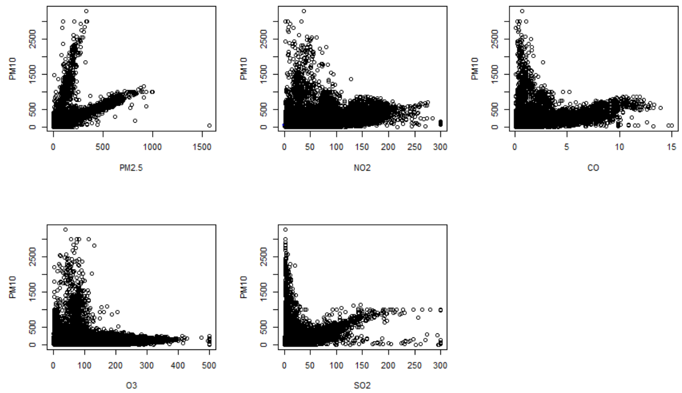


Figure2.2 The Scatter Plot of PM10 with other pollutant

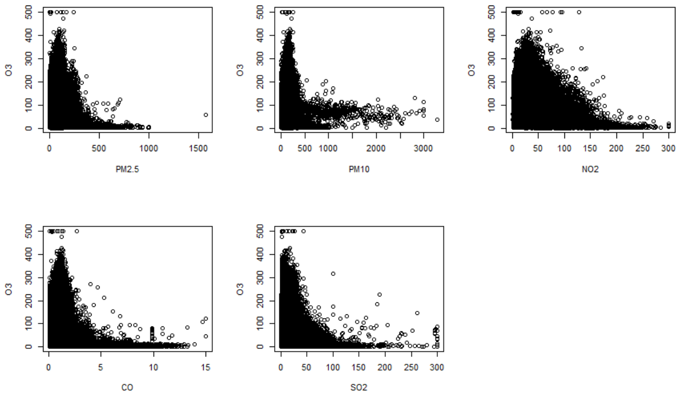


Figure2.3 The Scatter Plot of O3 with other pollutant

According to Figure2.1 and Figure2.2, we can find that PM2.5 and PM10 might have some linear relationship. If the concentration of PM2.5 is high, the concentration of PM10 will also high. However, the relationship of other pollutant is not clear enough in scatter plot. So we will also do other kinds of analysis.

2.1 Correlation between air quality features

Table2.1 correlation matrix

	PM2.5	PM10	NO2	CO	O3	SO2
PM2.5	1.000000	0.706965	0.568204	0.717363	-0.086850	0.571312
PM10	0.706965	1.000000	0.407165	0.457216	-0.027593	0.377578
NO2	0.568204	0.407165	1.000000	0.573579	-0.505092	0.397035
CO	0.717363	0.457216	0.573579	1.000000	-0.253588	0.473479
O3	-0.086850	-0.027593	-0.505092	-0.253588	1.000000	-0.144037
SO2	0.571312	0.377578	0.397035	0.473479	-0.144037	1.000000

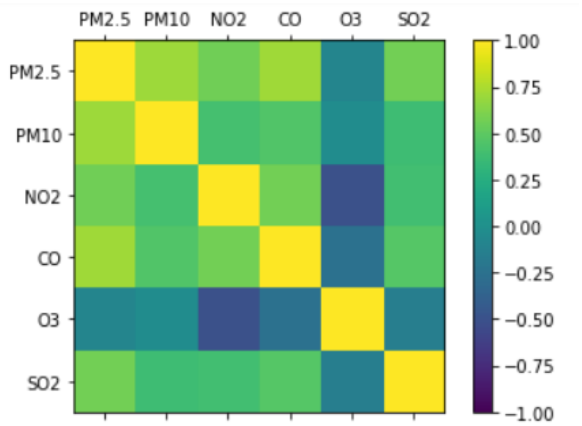


Figure 2.4 Correlation between air quality features

After watching Table2.1 and Figure 2.1, PM2.5, PM10, NO2, CO and SO2 have a certain degree of correlation and are positively correlated. But O3 is negatively correlated to PM2.5, PM10, NO2, CO and SO2, which is similar to scatter plots above.

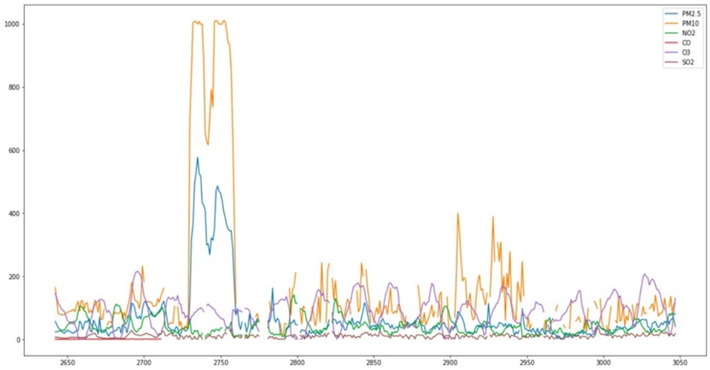
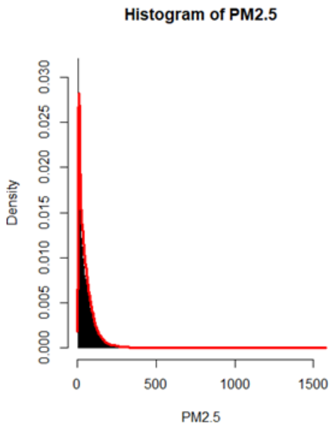


Figure2.5 Air quality of a station in May and June

According to Figure2.5, we can see that PM2.5, PM10, NO2, CO, O3 and SO2 are not stationary and they change very quickly. The curves of PM2.5, PM10, NO2, CO and SO2 seem to have same trend.

2.2 Distribution of PM2.5, PM10 and O3



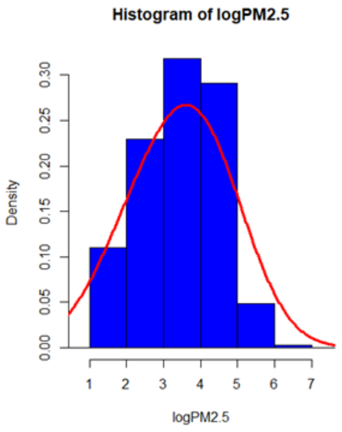


Figure2.6 Histogram of PM2.5

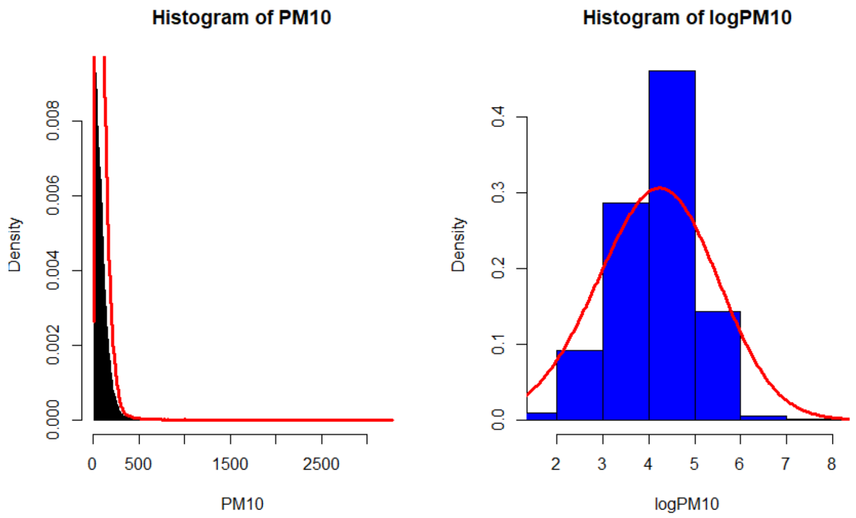


Figure2.7 Histogram of PM10

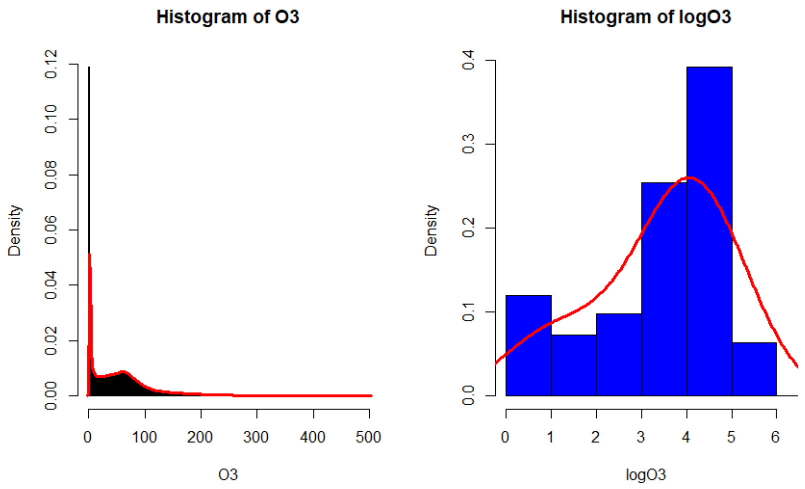


Figure2.8 Histogram of O3

Observed from Figure2.6, Figure2.7 and Figure2.8, the distribution of PM2.5 and PM10 are skewed, which is not suitable for regression. But O3 almost follows a normal distribution. After transferring to log(PM2.5) and log(PM10), their distributions almost are normal distribution. Therefore, when we train the model and predict, we will use log function to normalize PM2.5 and PM10.

3. Data Processing

3.1 data cleaning

Since the real word data is not clean, we need to find the dirty data and clean them before we train the model.

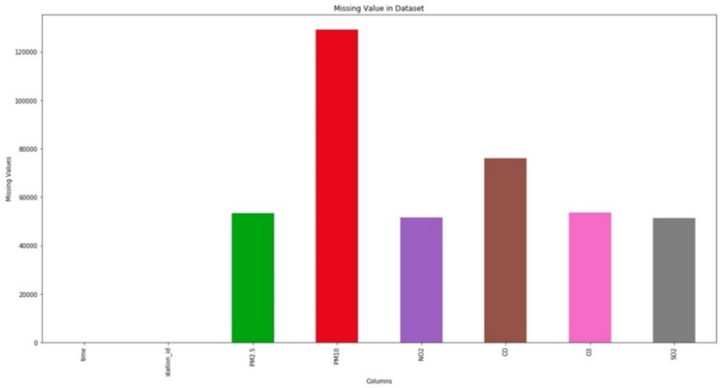


Figure3.1 Sum of missing values in air quality

According to Figure3.1, we can see that there are a lot of missing values in air quality data. For example, there are more than 120000 missing values in PM10. We can't fill these missing values by some simple methods like using mean value or 0 to replace because they can't reflect the real situation of these missing values. We plan to fill them by several methods.

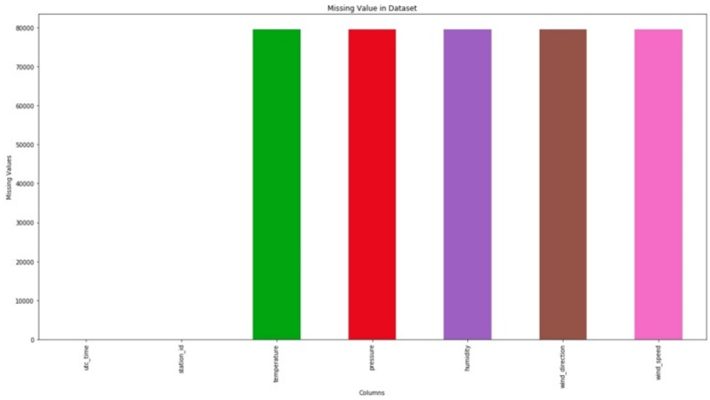


Figure3.3 Sum of missing values in gridWeather

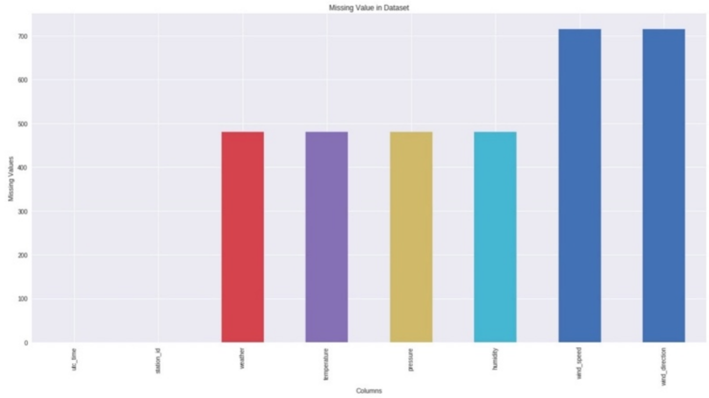


Figure3.4 sum of missing values in observedWeather

According to Figure3.3 and Figure3.4, both grid weather and observed weather have some missing values. But compared with air quality data, the missing rate is not high. So we can fill these missing values according to the adajacent values by using linear method.

In addition, there are some hidden missing values. A day has 24 hours, but we find that there is some missing time in airquality, gridWeather and observedWeather data. For example, in Figure2.5, we can see that the line plot of air aquality data is not continuous, which means in some time points, the data is missing. Because we will exact some features related to time, it is necessary for us to fill the time and make it completed. In addition, the time zone of some data is not the same, some of them are UTC time and some are Beijing time. The UTC time will be replaced by Beijing time.

3.1.1 Filling missing value with linear method

For a single attribute X, if there are consecutive missing values that don't exceed three, a linear method is used as bellow. Suppose X_i is missing. $If\ X_{i-1}\ and\ X_{i+1}$ are not missing,

$$X_i = \frac{X_{i-1}+X_{i+1}}{2} \tag{1}$$

If X_{i-1} is also missing but $X_{i-2}\ and\ X_{i+1}$ are not missing,

$$X_i = \frac{X_{i-2}+2X_{i+1}}{3} \tag{2}$$

$$X_{i-1} = \frac{2X_{i-2}+X_{i+1}}{3} \tag{3}$$

If X_{i+1} is also missing but $X_{i-1}\ and\ X_{i+2}$ are not missing,

$$X_i = \frac{2X_{i-1}+X_{i+1}}{3} \tag{4}$$

$$X_{i+1} = \frac{X_{i-1}+2X_{i+2}}{3} \tag{5}$$

If X_{i-1}, X_{i+1} are also missing but $X_{i-2}\ and\ X_{i+2}$ are not missing,

$$X_{i-1} = \frac{3X_{i-2}+X_{i+2}}{4} \tag{6}$$

$$X_i = \frac{2X_{i-2}+2X_{i+2}}{4} \tag{7}$$

$$X_{i+1} = \frac{X_{i-2}+3X_{i+2}}{4} \tag{8}$$

We use this method to fill the missing values of grid weather and observed weather air. Most of the missing values can be filled by using this method.

We also use this method to fill the missing values of air quality data. But since the large volume of missing data, there are still a lot of missing values left after doing this.

3.1.2 Filling missing value with baseline model

To fill the left missing values of air quality data, we train three baseline models to predict PM2.5, PM10 and O3, respectively. We choose LightGBM as our baseline model. The train data contains 158 features which are the maximum, median, and minimum of PM2.5, PM10, NO2, CO, O3 and SO2 over past 168 hours, 72 hours and

24 hours, station id, station type, longitude, latitude, hour, weekday, monthday, month, holiday, season, weather, windspeed, wind direction, temperature, pressure, humidity.

To fill null values of PM2.5, we treat the rows that PM2.5 is missing as the testing data. Similarly, to fill null values of PM10, we treat the rows that PM10 is missing as the testing data. To fill null values of O3, we treat the rows that O3 is missing as the testing data. Then we put these test data to the model to predict the missing PM10,PM2.5 and O3. We fill the missing PM2.5, PM10 and O3 in original tables with predicted PM2.5, PM10 and O3.

3.1.3 Remove Duplicates

There are over 6000 duplicates in air quality tables. Therefore we delete those duplicate rows.

3.2 Data Combination

The format of data we have is different. For example, we have three csv files for air quality data. The columns name of airQuality_201802-201803.csv and airQuality _201804.csv are different. In addition, the order of station id is different between airQuality_201802-201803.csv and airQuality_201804.csv.

At first, we merge airQuality_201701-201801.csv, airQuality_201802-201803.csv and airQuality_201804.csv. Secondly, we merge observedWeather_201701-201801.csv, observedWeather_201802-201803.csv and observedWeather _201804.csv. Thirdly, we add station location, longitude and latitude to airQuality and observedWeather. Fourthly, we group airQuality by stations. And then we merge gridWeather_201701-201803.csv and gridWeather_201804. Lastly, we add longitude and latitude of Beijing grid weather station to gridWeather.

4. Feature Engineering

4.1 Feature Extraction

4.1.1 Space Features

There are 35 stations. We extract features for different stations. Different stations have different concentration levels of several pollutants. So we construct space feature like station id, longitude and latitude. With the rapid increase in the number of motor vehicles in cities and the increase in traffic congestion, motor vehicle exhaust has become one of the main sources of air pollution in large and medium cities in China. Therefore there are significant differences in atmospheric pollution between the suburbs of Beijing, urban areas and areas near traffic. So we add a new feature —station type.

4.1.2 Temporal Features

The air quality changes greatly during the day and there are some periodical changes. The air quality is always better in the afternoon. So we create a feature — hour. And we want to use some feature to replace original time, so we create time feature like monthday and month. There is a difference in transportation between workday and holiday. So we exact a new feature — workday_holiday to describe whether that day is holiday or workday. We also add a feature — weekday to describe the day-of-week.

When winter is coming, heating is provided throughout Beijing. So coal burning will cause a significant increase in air pollutants. PM2.5 in spring mainly comes from the contribution of sand dust in the north and the contribution of farmland straw burning in surrounding areas. In autumn, due to strong solar radiation, atmospheric oxidation is enhanced, and photochemical smog often occurs. At the same time, atmospheric diffusion conditions are not good to cause pollutants to aggregate, which increases the chance of conversion of gaseous pollutants to secondary particulate matter, resulting in severe pollution. We think air pollutant is seasonal. In the summer temperature conditions are not easy to invert and more frequent rainfall and windy weather are conducive to the diffusion and removal of PM2.5, so the PM2.5 concentration in summer is the lowest. So we create a season feature.

4.1.3 Weather Features

According to related research, frequent rainfall and windy weather are conducive to the diffusion and removal of air pollutant. When it rains or snows little, the air quality can't be improved immediately. The windless weather without wind and rain will also cause more serious pollution caused by cold air. So weather is important for air pollutant. As left photo of Figure4.1 shows, for each station, we adapt the closed observed weather to represent its weather.

If the temperature and pressure make the air unable to form convection, the air pollutant will be difficult to spread. Moist air allows suspended pollutants to adsorb other pollutants, which is more likely to cause pollutant accumulation. Therefore we choose temperature, pressure and humidity as features. As right photo of Figure4.1 shows, for each station, we use the nearest grid weather to represent its temperature, pressure and humidity.

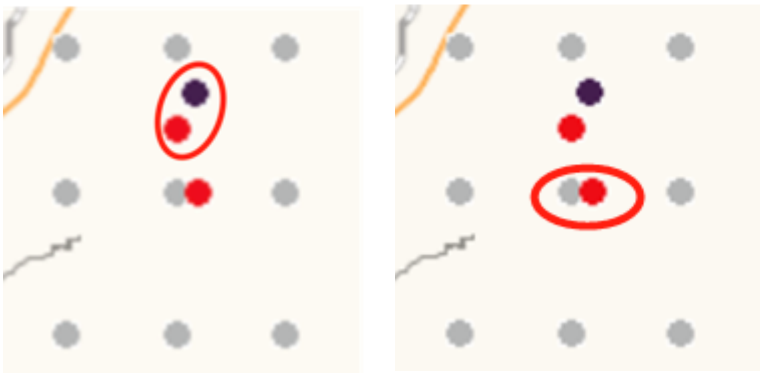


Figure4.1 station map

Wind is an important factor affecting the spread of air pollutant. So we choose windspeed and wind direction as features. In addition, we divide the wind direction into eight directions and 45 degrees in each direction. For wind direction, we also use One-Hot Encoding.

4.1.4 Air Quality Features

To predict PM2.5, observed from Figure4.2, we can see that PM2.5 of the past 72 hours has a strong autocorrelation. So we choose PM2.5 over the past 72 hours as features. In addition, Both the concentration of pollutant and the diffusion of wind are directional. We consider air pollutant nearby. We divide the wind direction into eight directions the wind direction of each area is determined by the mode. Therefore for each station, we create a feature that is last one hour PM2.5 of its nearest air quality station in 10 kilometers in that direction. If there is no air quality station in 10 kilometers in that direction, we use last one hour PM2.5 of target station.

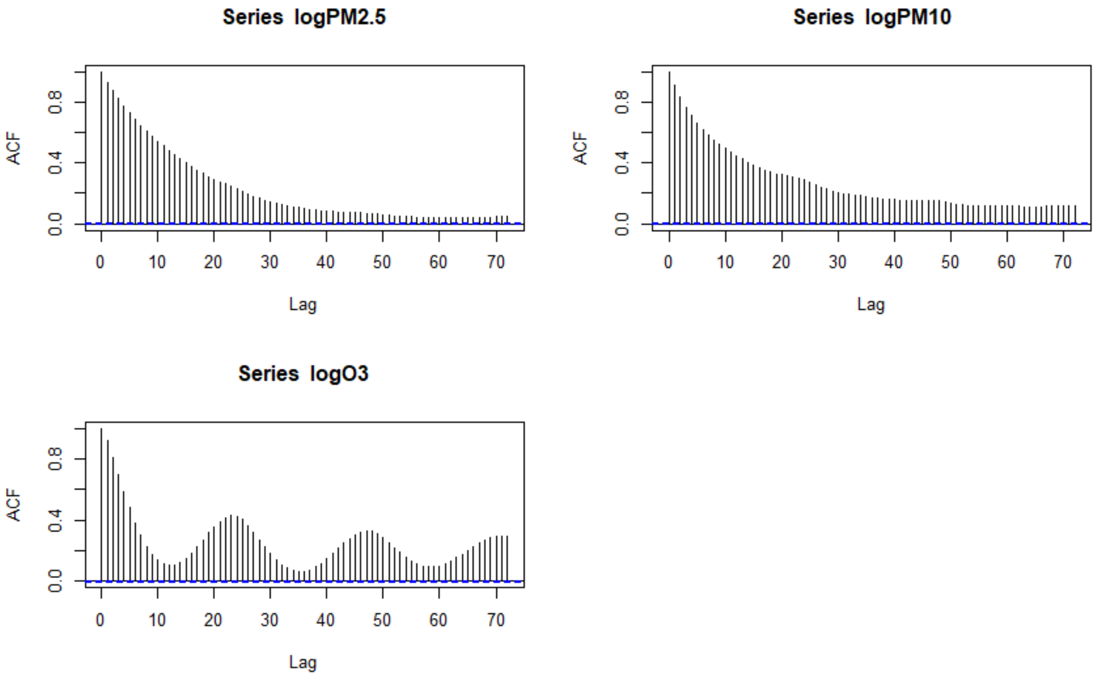


Figure4.2 ACF of log(PM2.5), log(PM10) and log(O3)

To predict PM10, observed from Figure4.2, we can see that PM10 of the past 72 hours has a strong autocorrelation. So we choose PM10 over the past 72 hours as features.

The same to PM2.5, we create a feature that is last one hour PM10 of its nearest air quality station in 10 kilometers in that direction. If there is no air quality station in 10 kilometers in that direction, we use last one hour PM10 of target station.

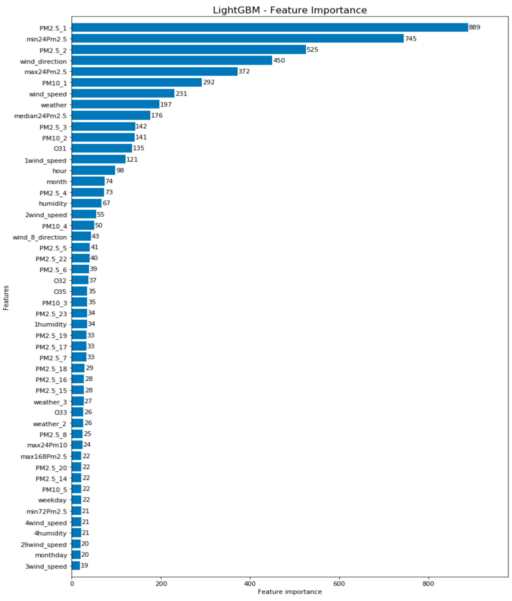
To predict O3, observed from Figure4.2, we can see that O3 of the past 72 hours has a strong autocorrelation. So we choose O3 over the past 72 hours as features. The same to PM2.5, we create a feature that is last one hour O3 of its nearest air quality station in 10 kilometers in that direction. If there is no air quality station in 10 kilometers in that direction, we use last one hour O3 of target station.

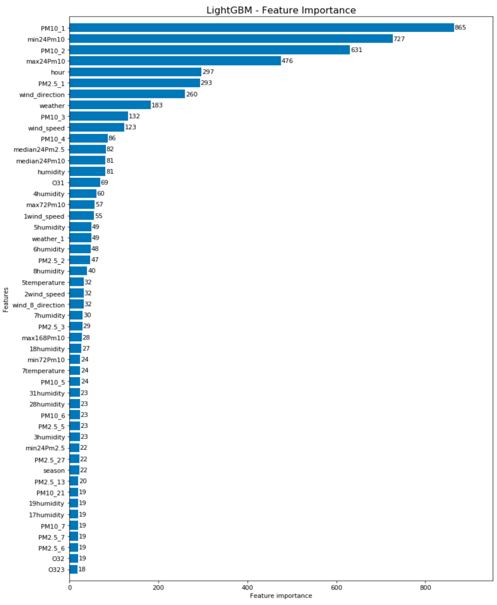
The maximum, median, and minimum of PM2.5, PM10 and O3 are also selected to be features.

4.2 Feature Selection

As mentioned above, we have hundreds of features. The model we build needs to further select features to abbreviate the runtime of the model. And some features have no significant impact on forecasting. Therefore we select the features based on the feature importance of LightGBM. Form the baseline model Figure4.3, we can see some of the features have very low importance, and that means these features have little contribution to construct the tree model. Consider the Occams Razor Theory, we discard this sort of feature.

In addition, we did some experiment to select features. At first, we run a model contains all the features we generated. Then we delete some features, if the error becomes bigger, we will put this feature back, if the error becomes smaller, we will not use this feature in the next model. We repeat doing this procedure until the error no longer decrease. For example, we drop the feature air quality(PM2.5,PM10,O3) of eight different directions in last hour in our final model because the error decrease after we drop it.





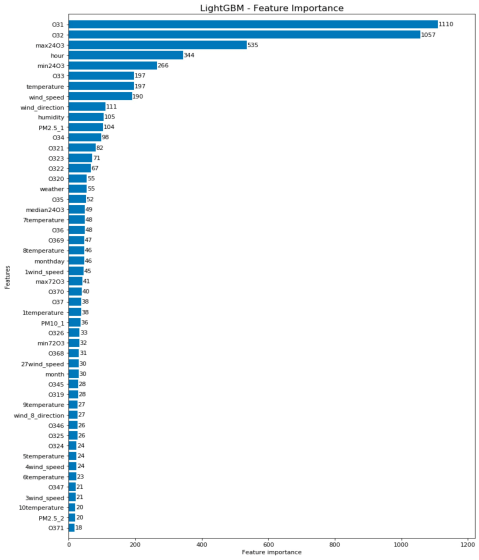


Figure4.3 feature importance of PM2.5, PM10 and O3

5. Model Construction

5.1 Validation Set and Training Set

Validation is a common method for determining the values of hyperparameters. We use train_test_split in sklearn.cross_validation to split the training data and validation data, the rate is 8:2. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an Independent dataset. So we use cross-validation to find best parameters.

5.2 Model Selection

This project is a regression problem. We can use many models to solve it, like GBDT models and Neural network models. We finally chose lightGBM. LightGBM is a gradient boosting framework that uses tree based learning algorithms. It has the following advantages: Faster training speed and higher efficiency; Lower memory usage; Better accuracy; Capable of handling large-scale data.

5.3 Model Building

Because different contaminants have different distributions, each contaminant needs to be trained separately. So for PM2.5, PM10 and O3, we train three different models. When we train the model, we do a log transformation on the labels of PM2.5, PM10. So we perform the exponential transformation to get the final result.

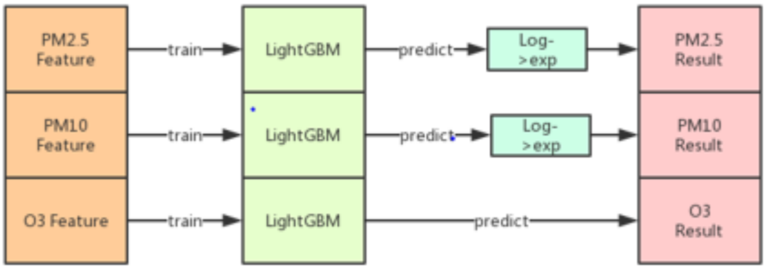


Figure5.1 The process of model

As mentioned above, we finally choose space feature, temporal feature, weather feature and air quality feature to train the model.

- **Space feature:** station id, station type, longitude, latitude.
- **Temporal feature:** hour, weekday, monthday, month, holiday_workday, season.
- **Weather feature:** weather, windspeed, wind direction, temperature, pressure, humidity and the weather of the past 72 hours.
- **Air quality feature:** PM2.5, PM10, O3 of the past 72 hours and the maximum, median, minimum of PM2.5, PM10, O3 of past 168 hours, 72 hours and 24 hours.

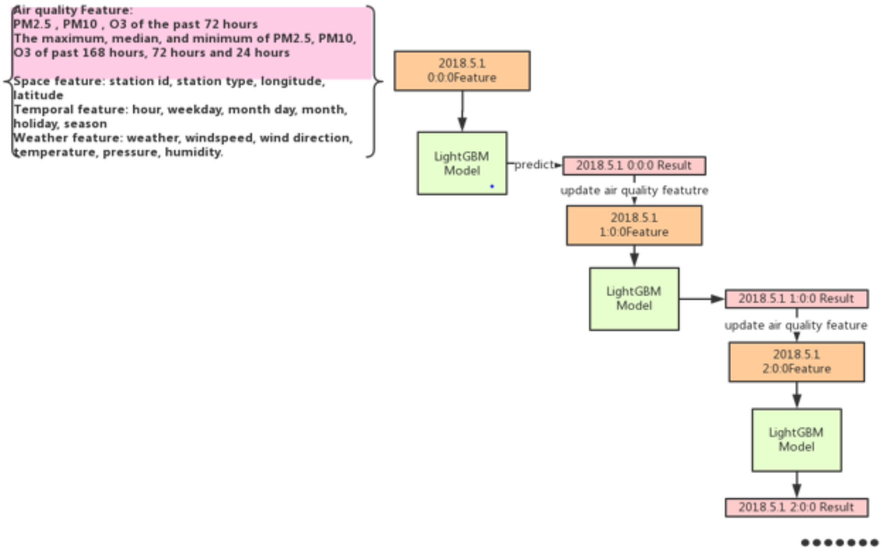

Air quality Feature:
PM2.5 , PM10 , O3 of the past 72 hours
The maximum, median, and minimum of PM2.5, PM10, O3 of past 168 hours, 72 hours and 24 hours
Space feature: station id, station type, longitude, latitude
Temporal feature: hour, weekday, month day, month, holiday, season
Weather feature: weather, windspeed, wind direction, temperature, pressure, humidity.
2018.5.1 0:0:0Feature
LightGBM Model
predict
2018.5.1 0:0:0 Result
update air quality featutre
2018.5.1 1:0:0Feature
LightGBM Model
2018.5.1 1:0:0 Result
update air quality feature
2018.5.1 2:0:0Feature
LightGBM Model
2018.5.1 2:0:0 Result
•••••••

Figure5.2 example of sliding window

The feature we use includes the past hours' PM2.5 and PM10 and O3. So we need to use our prediction results to update the data for May. For example, if we want to predict the PM2.5 of at 15:00:00 1/5/2018, the feature of this time point includes the past 72 hours' PM2.5, but we don't have the real PM2.5 data from 00:00:00 1/5/2018 to 14:00:00 1/5/2018. We need to use our predict results to update these PM2.5 data.

To solve this problem, we only predict a hour's PM2.5 at a time and then update our data. After update, we can predict next hour's PM2.5. We will predict 48 times in this project. This process is identical to predict PM10 and O3.

According to Figure5.2, after predicting PM2.5 at 0 o'clock on May 1st2018, we will use the prediction to update air quality feature to predict PM2.5 at 1 o'clock on May 1st2018. For PM2.5 at 1 o'clock on May 1st2018, we have new PM2.5 of past 72 hours and new maximum, median, minimum of PM2.5 past 168 hours, 72 hours and 24 hours.

6. Conclusion

In this project, the most difficult parts are fill missing values and feature engineering.

We tried a lot of methods to fill missing values, like replacing with mean value or previous value. But in some cases, such method is not reasonable. So finally we use linear method and the baseline model to fill them.

In feature engineering part, we exact many features. To test whether these features are useful or not, we did a lot of experiments. And this is a time-consuming work.

We also meet some difficulties in model construction part. For example, we need to update the testing feature according to the test result last several hours. But finally we solve it successfully.

We spent a lot time on this project, and also learn a lot from it. In the feature, we will try some other models like LSTM and will also extract more useful features like the frequency domain features.

Reference

[1] https://baijiahao.baidu.com/s?id=1610208209814240444&wfr=spider&for=pc
[2] https://www.leiphone.com/news/201808/Al3ERxLCTUpabtaI.html

Task Assignment

Chen Donger 20551304:
Code: Data Preprocessing, Data Merging, Feature Engineering, Feature Selection, Model Construction
Report: Feature Selection, Model Construction, Conclusion

Wu Xiting 20549143:
Code: Data exploration, Fill missing values, Baseline model, Model Testing
Report: Introduction, Data Exploration, Data Preprocessing, Feature Extraction, Model Construction