Air Quality Prediction

Abstract

For data preprocessing part, we firstly fill in the missing values by using the time and space characteristics. The outliers are corrected using data from the closest site. Finally, we do the log10 and units of measurement processing. For feature engineering, we construct features about weather, historical air quality and time. Finally, we build several candidate models for training and prediction and select three models on the basis of their performance by using cross validation method.

1. Introduction

1.1 Background

Over the past years, air pollution has become more severe in many large cities, such as Beijing. As we know, Harmful health effects of particle matter suspended in the atmosphere are well established. Mass concentration of particles with an aerodynamic diameter less than 2.5/10 lm (PM2.5, PM10) and O3, is usually used as a standard measure of air pollution. Therefore, accurately monitoring and predicting the concentration of these particles have become increasingly essential. If it were possible to predict these episodes one or two days in advance, more efficient actions could be taken in order to protect the citizens. Our study is based on data obtained at 35 official monitor stations and observerd/grid weather stations in the city of Beijing. In this task, we aim to predict concentration levels of several pollutants, including PM2.5, PM10 and O3 for the next 48 hours at 35 different air quality stations in Beijing.

1.2 Dataset Description

We use the data from 2017-01 to 2018-01 as our training set. We have three data resources: air quality, grid weather and observed weather. Considering weather stations and air quality stations are separated and the data format is not uniform, data preprocessing and feature engineering is of significance for us to predict the air quality for the next 48 hours.

- Air Quality Data: We have three csv file about air quality data from 2017-01-01 to 2018-04-30, containing air quality data and time series. Here we will pay attention to the PM2.5, PM10 and O3 attributes.
- Observed Weather Data: This file describes weather data at observed weather stations. There are human errors or equipment faults occur from time to time.
- Grid Weather Data: The csv file is from satellite image and other meteorology and atmosphere data, including the information of temperature, humidity, pressure, wind speed, wind direction and weather.

2. Data Preprocessing

2.1 Missing Value

2.1.1 Air Quality Missing Value

After initial data exploration, we find there are lots of missing values in these sheets. In sheet "airQuality_201701-201801", the missing value is like what below:

Table 1. Missing value of Air Quality Data

Pollutant	The number of missing value	Percentage
PM2.5	20389	0.065557
PM10	83263	0.267718
O3	20421	0.065660

- **For PM2.5 and O3 missing value (use spatial feature)**

These two types of missing values are relatively small. To fill these missing values in air quality data set, we build a list for each air station which contains 12 air stations that are close to the specific one among 35 air stations. When there is a missing value, we use the closest one in the list to replace the null. If all these 12 stations miss the same feature in the same time, we use the average all over the year to fill the null.

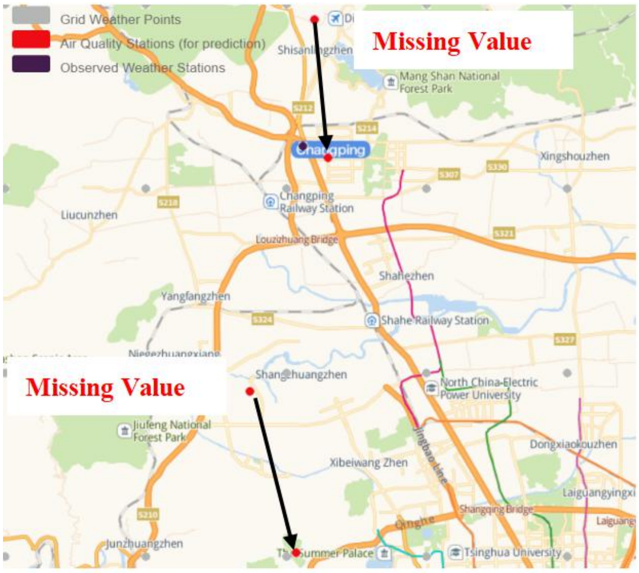


Fig. 1 Fill Missing Value for PM2.5 and O3

- **For PM10 missing value (use time feature)**

As shown in the table 1, the percentage of missing values in PM10 is much larger. Also, we find that there is a strong linear relationship between PM2.5 and Pm10 and the correlation coefficient is 0.88. Therefore, we build a linear regression model to predict the missing value of PM10.

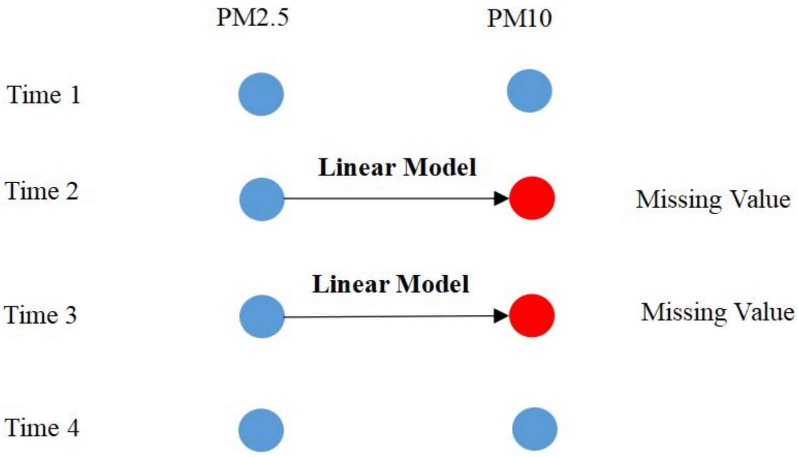


Fig. 2 Fill Missing Value for PM10

2.1.2 Observed Weather Missing Value

Besides, there's also some missing values in observed weather data set. We use the grid weather data which is closet to the missing value to fill the null.

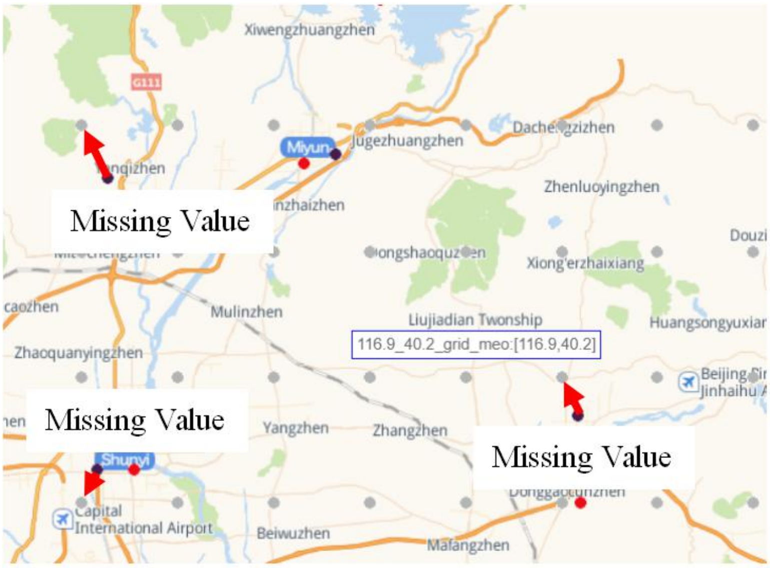


Fig. 3 Fill Missing Value for Observed Weather

2.2 Outlier

After use “describe” function in these data sets, we found some outliers in weather data. For variable “wind_direction”, those equal to “99017” are set to “0”. Here, we replace all the outlier with the value of its closet grid weather station. The outlier detection is as follows:

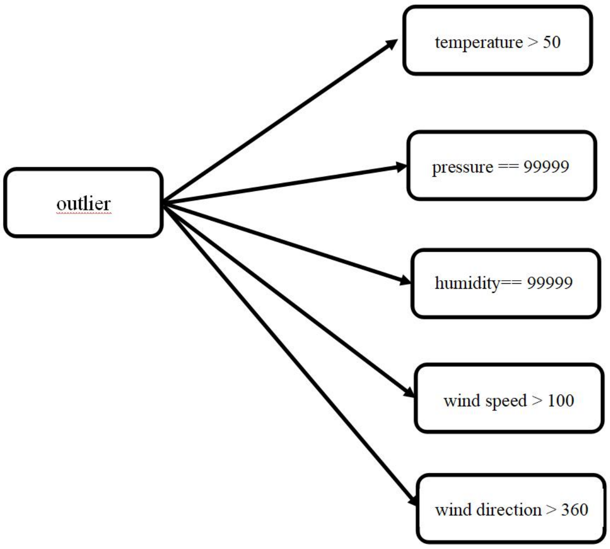


Fig. 4 Outlier Detection

2.3 Standardize the Units of Measurement

The units of wind speed feature in observed weather data is “m/s”, while the units of wind speed in grid weather data is “km/h”. So we need to do a transformation:

$$\mathrm{m}/s = km / h * 3.6$$

2.4 Log10 Processing

We check the distributions of our variables. For example, we find the distribution of PM2.5 is skew, shown in figure. So we apply a “log10” transformation on these variables. Then its distribution is more like normal distribution. As we know, data coming from a normal distribution will help the model come out a better performance, so we apply the “log10” transition to all the numerical variables.

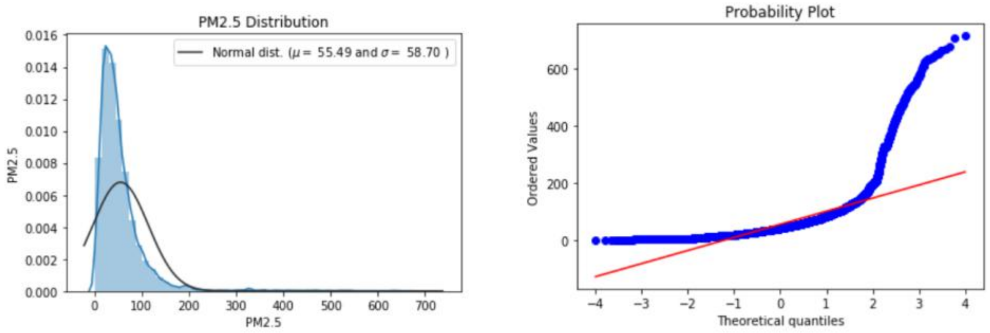


Fig. 5 PM2.5 Distribution before Processing

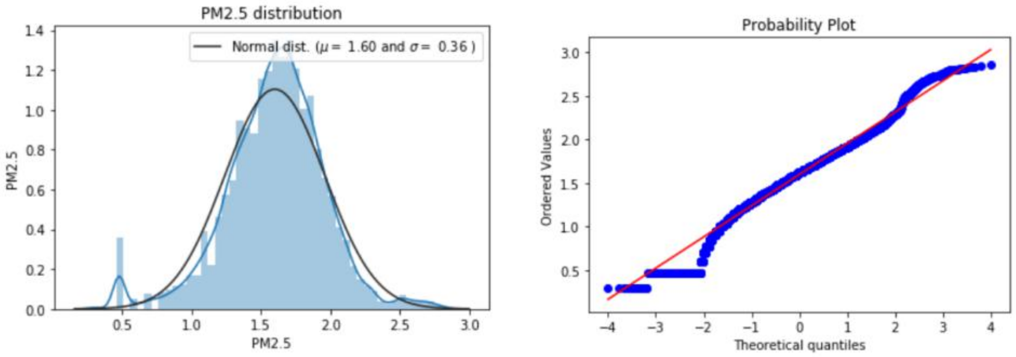


Fig. 6 PM2.5 Distribution after Processing

3. Feature Engineering

There are 18 observed weather stations and 650 grid weather stations and we need to construct features using these data. Here we consider weather data, historical air quality data and time data as our features.

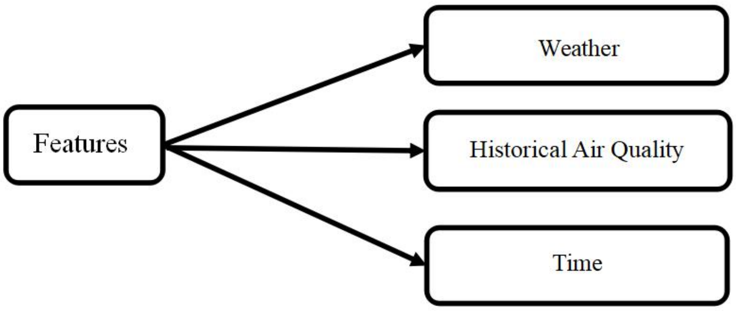


Fig. 7 Feature Engineering

3.1 Weather Features

To predict the air quality in 35 air quality station, we have to construct weather features for them. As the grid weather stations are dense and for each air station, there must be a grid weather station that is very close to it, we decided to use the weather features from the closet grid weather station as weather features in air station.

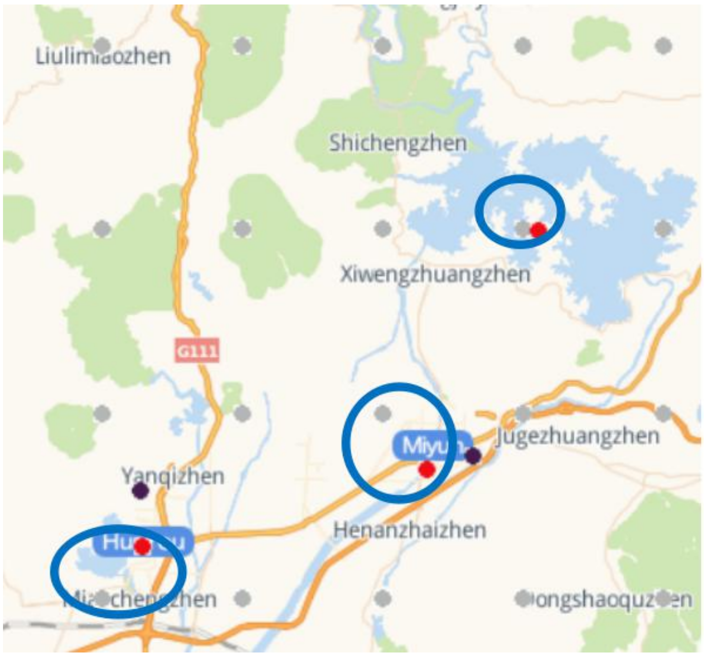


Fig. 8 Weather Features for Each Air Stations

We have temperature, humidity, pressure, wind speed and wind direction in our data. Especially for wind direction data, we need to some process because it represents a value of wind direction angle. The wind blowing from the east comes from the sea, and it will be clean. On contrary, the wind blowing from the west comes from desserts and grassland, and it will be full of pollutants. Considering this, we divide the wind direction into 8 categories.

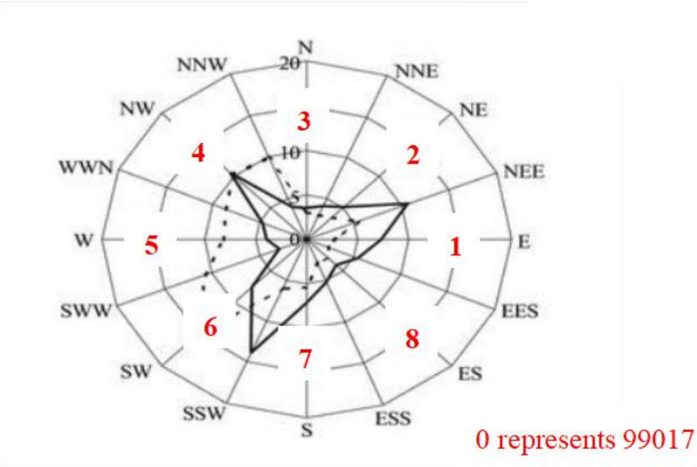


Fig. 9 Wind Direction Feature

What’s more, we consider that the value of the pollutant concentration at a certain time is the accumulation of weather effects in the previous period, e.g. 8 hours. So when we build the weather feature, we need to take the weather in the previous time period into account. Here we choose to use the mean of temperature, humidity, pressure and wind speed and the mode of wind direction in the past 8 hours as the weather data for each time point. This method can reflect the cumulative effect of weather on pollutants.

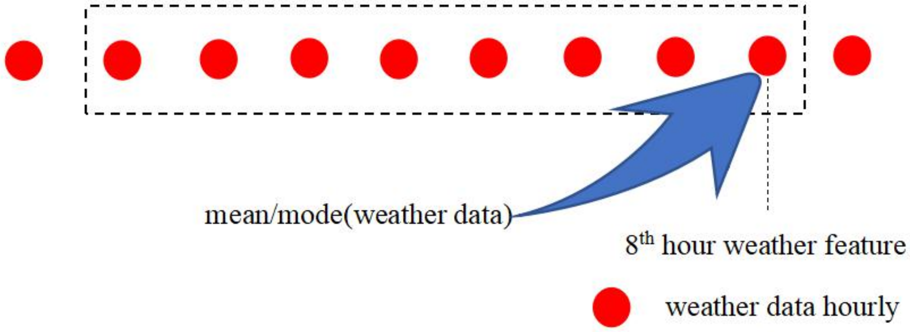


Fig. 10 Deal with Weather Cumulative Effect

3.2 Historical Air Quality Features

There is some information about air quality features. We want to generate some features from previous air quality data to capture the trend of air quality and their magnitude in a period. We add the concentration of PM2.5, PM10 and O3 for the previous hour to the historical air quality features. (For the test set, we use the prediction output of the previous data as the feature of the next data). Also, We compute the mean, median, standard deviation, maximum, minimum and the range of PM2.5, PM10 and O3 in the past 48 hours.

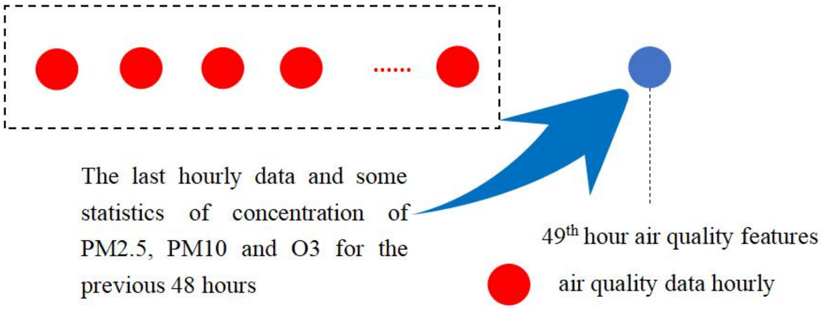

The last hourly data and some statistics of concentration of PM2.5, PM10 and O3 for the previous 48 hours
49th hour air quality features
air quality data hourly

Fig. 11 Historical Air Quality Features

3.3 Time Features

After reviewing the data, we found that time has a great impact on the concentration of pollutants. Therefore, we create the following features based on time information.

Table 2. Time Features

Time Features	Meaning	Description
weekday	Determine the day of the week	0-Sunday, 1-Monday,...,6-Saturday
rush hour	Determine whether current time is rush hour or not	00:00-02:00,09:00-12:00 in UTC_Time is rush hour
midnight	Determine whether current time is midnight or not	15:00-22:00 in UTC_Time is midnight
month	Determine the month of current time	1-Jan., 2-Feb., ..., 12-Dec.

Note: There are 8 hours difference between Beijing time and UTC Time. And considering 5.1-5.2 is Chinese Labor Day(holidays), we change the weekday feature of 05.01-05.02 into 6 and 0 in the test set.

3.4 Dummy Variables

We have several categorical variables, i.e. wind_direction, weekday, midnight, rush_hour, month. We convert them into dummy variables. For example, we can use (1,0,0,0,0,0,0) to denote the Monday and (0,0,0,0,0,0,1) to denote Sunday.

4. Model

We select some models as our candidate models. And we train candidate models and find the models with good performance. Then we average them to do the prediction work. Because PM2.5, PM10 and O3 have different patterns, we will train the model for PM2.5, PM10 and O3 respectively.

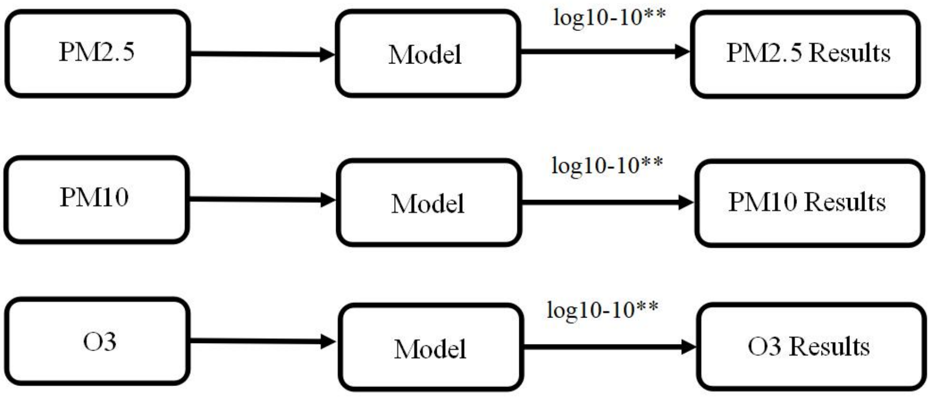


Fig. 12 Model training

In the process of our model training, we use mean square error as a measurement by cross validation method (k=5).

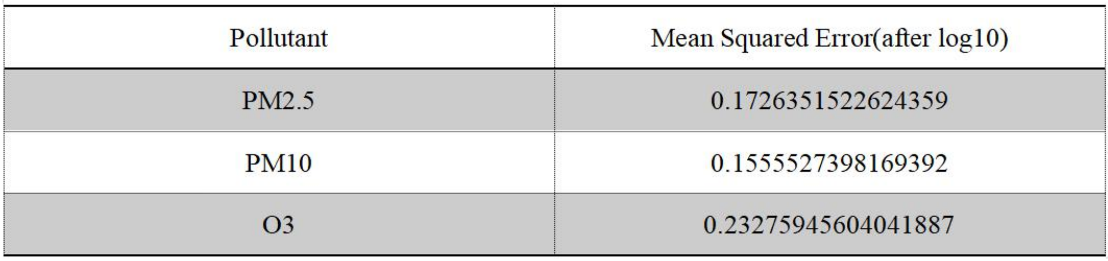
4.1 Candidate Model

➢ **Gradient Boost**

GBoost is often used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

We get the GBoost results by using cross validation method:

Table 3. Mean Squared Error by using GBoost

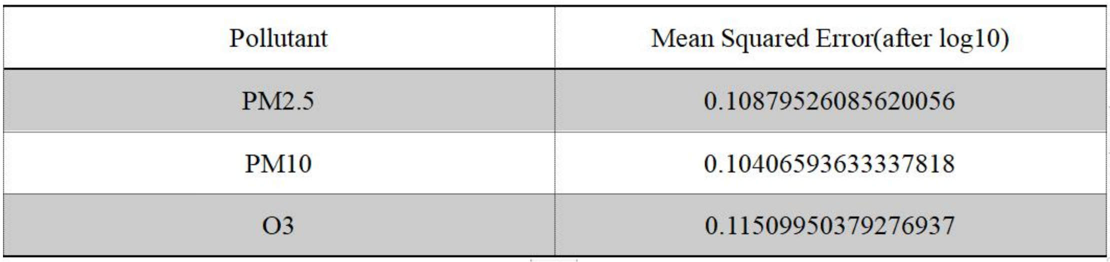


Pollutant	Mean Squared Error(after log10)
PM2.5	0.1726351522624359
PM10	0.1555527398169392
O3	0.23275945604041887

➢ **XGBoost**

We get the results by using cross validation method:

Table 4. Mean Squared Error by using XGBoost

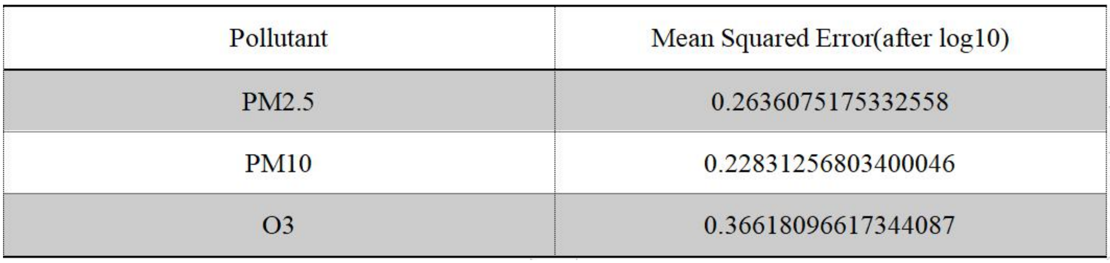


Pollutant	Mean Squared Error(after log10)
PM2.5	0.10879526085620056
PM10	0.10406593633337818
O3	0.11509950379276937

➢ **LightGBM**

We get the results by using cross validation method:

Table 5. Mean Squared Error by using LightGBM



Pollutant	Mean Squared Error(after log10)
PM2.5	0.2636075175332558
PM10	0.22831256803400046
O3	0.36618096617344087

➢ **MLP**

MLP is a class of feedforward artificial neural network. An MLP consists of three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

We get the results by using cross validation method:

Table 6. Mean Squared Error by using LightGBM

Pollutant	Mean Squared Error(after log10)
PM2.5	0.1726351522624359
PM10	0.21237649138393666
O3	0.3241699061719891

4.2 Model Selection and Prediction

Based on the performances of several models, we choose XGBoost, LightGBM and Gradient Boost as the final model. We average them with a certain weight. And then we can get the prediction output.

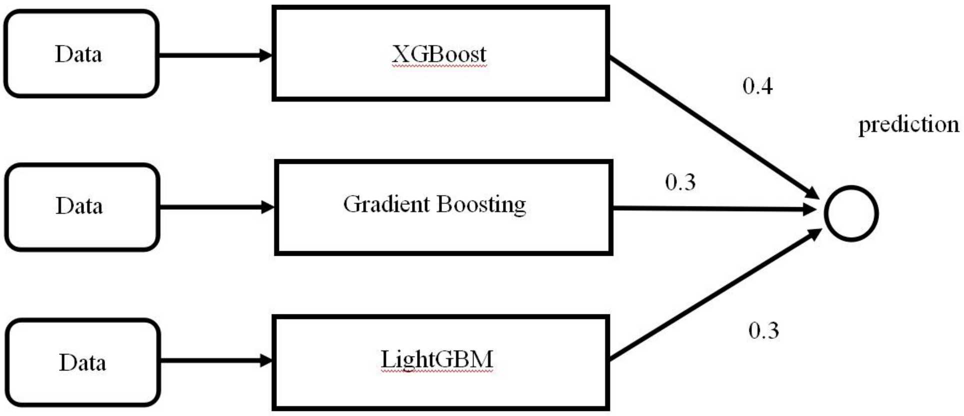


Fig. 13 Model Selection