COMP6237 predictive project: Predicting the air quality

YUEFU LIU, YULIN LI, WEIQING DING, DANFENG GUO, PENGXIANG WEI
University of Southampton
yl2e18@soton.ac.uk,yl8n18@soton.ac.uk,wd2m18@soton.ac.uk,dg1a18@soton.ac.uk,pw1a18@soton.ac.uk

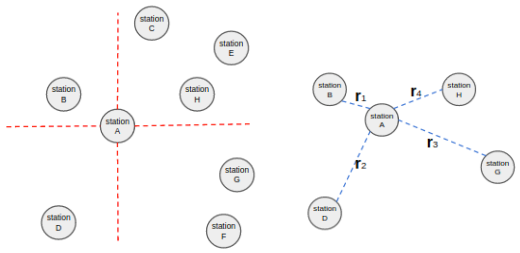
1 INTRODUCTION

This project builds an air quality prediction system for the KDD cup 2018 competition [1]. We predict the concentration level of PM2.5, PM10, and NO2 over the upcoming 48 hours in London. From our prediction results of London, we have achieved performance far exceeds the top 10 teams, which predicted both Beijing and London.

2 TASK EXPLANATION

2.1 The raw data and difficulties

KDD provides us the grid weather data, which contains 861 grid points, and the air quality data of three pollutants, which contains 24 stations. Fig. 1(a) shows an example. There are 13 stations that need to be predicted, while the other 11 stations do not need.



(a) Location of station A (b) Four neighbours

Figure 1: Example of station A

The raw data is noisy, unstable and full of missing values. Moreover, the air quality changes irregularly and the time series is difficult to model. However, the biggest difficulty is that the forecast covers all future 48 hours. Since the predicted value of each hour is based on the predicted value of the previous hour, as the predicted time continues to retreat, the uncertainty of the prediction itself will continue to accumulate, assuming that the initial error of the prediction in the next 1st hour is only 10%, then the error of the 48th prediction will reach a staggering level.

2.2 Evaluation method

Since we need to evaluate different models, we need to have a unified indicator and test set. In this task, we use the SMAPE [2] (symmetric mean absolute percentage error) as our unified indicator to evaluate the error during our training and testing (Github: [3]).

$$SMAPE = \frac{1}{n}\sum_{n}^{t=1}\frac{|F_t - A_t|}{(A_t + F_t/2)}$$

This evaluation function is based on the Norm-1 MAE evaluation function, Which is divided by the average of the predicted and true values. Using SMAPE can prevent the abnormal value unique to the air quality problem from influencing the overall score. for example, when the air quality changes to a large value abruptly, the denominator term can weaken the influence effectively.

For the unified dataset, after preprocessing and exploration, we just select the data of 13 stations that need to be predicted. Because our machine can't process more data (20GB), and we found that only trained with 13 stations can get the same results of all stations. Also, we separate the processed data into a training dataset and testing dataset. In order to maximize the independence of the test set, we selected the last 20% of the data for each station as the test set, which is around the last two months of the entire time span.

3 DATA PREPROCESSING

3.1 Preparations

3.1.1 Find neighbour stations. The first step in our preprocessing is to calculate the neighbors in four directions of each station based on longitude and latitude. Figure 1 shows an example which separates all neighbors into four directions and chooses the nearest station of each direction. In this step, we also calculate and record the distance of each nearest neighbour: r_1, r_2, r_3, and r_4 ($r \neq 0$).

3.1.2 Merge data and handle negative values. Then we merge air quality data of 13 stations and other 11 stations and sort them by time. Since there are some negative values in the raw air quality data, and we know that there is no negative pollution concentration in real life, we replace all negative values with zero.

3.1.3 Calculate average change%. Then we calculate the average of change%, which is the rate of change of pollutant concentration per hour compared to the previous hour. In this step, considering the different average change% in fresh condition and pollution condition, we set the concentration threshold of 20 to separate these two conditions. As a result, around 90% of the non-null data meets the condition that the change% is less than (average change%)×2.

3.1.4 Group by station. In order to facilitate the subsequent operations, we store the data in the form of a dictionary according to different stations. Each station ID is used as the key, and the dataframe corresponding to the station is used as the value.

3.2 Fill missing values

3.2.1 Merge duplicated stations. The latitude and longitude of the two groups of stations are completely coincident, so they are actually two same stations. Then we merge these two groups of stations into two stations, in the process, merging their values in missing items and subtracting the duplicated stations.

3.2.2 Weighted average of neighbours' values. For the remaining missing values, we use the weighted average of the same pollutant concentration at the same time in their nearest neighbors, and their weights are inversely related to the distances. Taking Fig. 1(b) as an example, the weighted average value of station A (V_A) can be

described as the weighted average of values of station B D G H:

$$K = Max(r_1, r_2, r_3, r_4) + Min(r_1, r_2, r_3, r_4)$$

$$V_A = \frac{V_B(K - r_1) + V_D(K - r_2) + V_G(K - r_3) + V_H(K - r_4)}{(K - r_1) + (K - r_2) + (K - r_3) + (K - r_4)}$$

After this calculation, we also consider the pollutant concentration change%. If the calculated result is too large or too small, we would adjust the result according to the average change% calculated by the last section, which already covers 90% non-null values.

3.2.3 According to the previous and next hour. Then we fill the missing values according to their previous and next hour and take the average of these two values. After this step, there are still a few missing values, and we do the weighted average step again.

4 DATA EXPLORATION

The main purpose of this section is to explore the relationship between different features or the combinations of features.

4.1 Temperature, Pressure and Humidity

Initially, we calculate the daily average concentration of three pollutants. By using the daily average data, we explore the relationship between the environment data(temperature, pressure, and humidity)and the concentration.

4.1.1 description. The data exploration results show that the concentration changes of the three pollutants in the day show a distinct bi-modal structure, and the peak data is concentrated in 7:00 to 9:00 and 17:00 in the day. At the same time, from about 22:00 to 3:00 in the next morning, the pollutant concentration was at the lowest level of the day. From the experimental conclusions drawn from figure 2 and figure 5(b), we can see that the highest temperature of the day is roughly between 10:00 and 15:00 in the day, which is a trough of the three pollutant concentrations in the day. That is to say, the temperature and the concentrations are negatively correlated. Furthermore, the trend of atmospheric pressure per day tends to be the same as the trend of pollutant concentration, but the peak delay of the atmospheric pressure is about 2 to 3 hours. Additionally, the trend of the humidity in a day is roughly opposite to the trend of temperature. Humidity reaches the lowest point of the day between about 12 and 15 o'clock, and then gradually rises until the same time period of the next day reaches the lowest point again. When we average the data by month, we can find that the conclusions about temperature and humidity are unchanged. However, the atmospheric pressure becomes an irrelevant feature here and is no longer a convergence relationship.

The details can be seen in the figure 3. The concentration of three pollutants in June, July and August is the lowest in the year. From October, the concentrations began to rise, peaking in January of the following year, and then decreasing month by month, reaching a low concentration in July of the following year. The annual change rate of PM2.5 concentration is the highest, up to 92.3%, while the annual change rate of nitrogen dioxide and PM10 is relatively low, 70.6% and 46.3%, respectively.

4.1.2 conclusion. In a short summary, during the day, the concentrations is negatively correlated with temperature, showing the same trend as atmospheric pressure, and positively correlated with

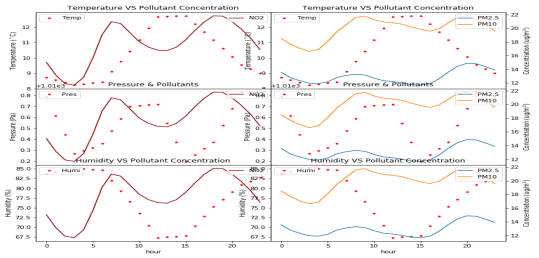


Figure 2: Temperature, Pressure and Humidity relations with Pollutant concentration

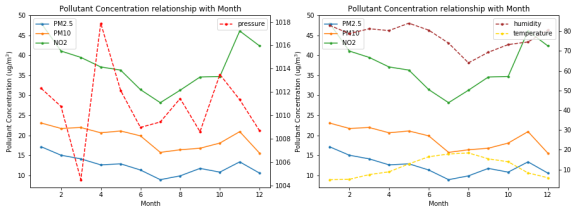


Figure 3: monthly change of pollutant concentration

humidity. But in a year, atmospheric pressure has no relationship with the concentrations. The sharp rise in pollutant concentrations since October may be caused by heating. The power demand caused by heating and the strong demand for gas has further led to a higher concentrations in winter to spring.

4.2 Holidays, Weekends and Workdays

A total of 3 different holiday attributes are recorded in the data, which are weekdays, weekends, and holidays. We classify the data according to the holiday attributes and then comparing the differences. The concentrations on holidays is expected to change significantly from the working day.

4.2.1 description. The experimental resultsis: working day pollutant concentration > average pollutant concentration > weekend pollutant concentration > holiday pollutant concentration. More specifically, the concentrations of NO_2, PM2.5 and PM10 were significantly reduced during the holidays, which respectively reduced by 40.97%, 35.16% and 37.34% compare to the average data of the working day.

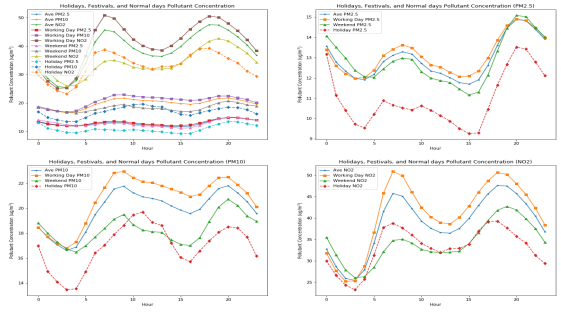


Figure 4: holidays, weekends and workdays comparison

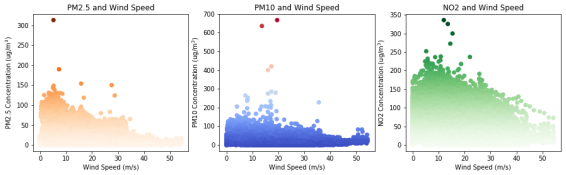
4.2.2 conclusion. To sum up, on the holidays and weekends, the concentration of the three air pollutants is reduced to different degrees relative to the working day, and the decrease of PM2.5 and PM10 is more obvious. This shows that the holiday attribute will be a very important attribute when we consider the impact of various features on the prediction results.

4.3 Wind Speed and Direction

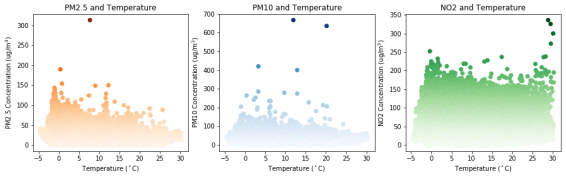
Wind speed and direction are also factors we believe might have impact on pollutants. We separately extracted the wind direction, wind speed, and the data of the concentrations, and reconstituted 5 sets of different vectors. Then we generate a heatmap of the five sets of vectors to compare the correlation coefficients between the five sets of vectors.

4.3.1 description. From the heatmap in figure 6(a) and figure 5(a) we can see that,
there is a negative correlation between wind direction and wind speed, and the concentration of the three pollutants. Among them, the influence of wind speed on the concentration of three pollutants is obvious. For NO_2, PM10 and PM2.5, the correlation coefficients are -0.34, -0.2 and -0.26, respectively. The wind direction does not seem so obvious for the change of NO_2, only -0.031, but there is a significant negative correlation between PM10 and PM2.5, and the correlation coefficients are -0.24 and -0.27, respectively.

4.3.2 conclusion. According to the correlation coefficient results displayed by heatmap, the two features of wind direction and wind speed can be used to predict the concentration of three pollutants. Therefore, when selecting data features in the next step, wind direction and wind speed are features that are worth choosing.



(a) wind trend

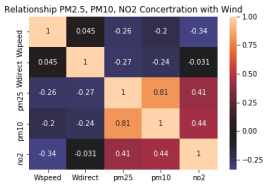


(b) temperature trend

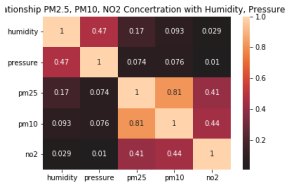
Figure 5: data distribution

4.4 Correlation coefficient between humidity, pressure and pollutants

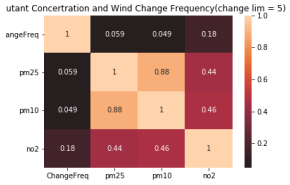
In this step, we will further explore the heat coefficient map of the correlation coefficient between humidity and atmospheric pressure based on the exploration method of the previous step.



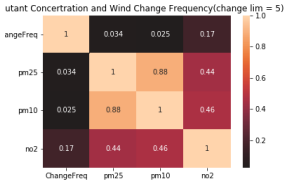
(a) wind speed and direction



(b) humidity and pressure



(c) limit=3



(d) limit=5

Figure 6: Correlation coefficient heat map

4.4.1 description. The results of the data exploration showed that the correlation coefficient between barometric pressure and humidity, and the concentration of the three pollutants was not as high as we expected. According to figure 6(b), the combination with the highest coefficient is the influence of humidity on PM2.5, and the correlation value is 0.17. This indicates that humidity does have an effect on PM2.5, but the effect is not sufficient to make humidity a key feature in predicting contaminant concentrations. Secondly, none of the other two has a correlation coefficient exceeding 0.1. To this end, we separately show the annual variation of humidity, and found that the annual variation of humidity features is not obvious, and the data range fluctuates between 66% and 83%. This may be due to the fact that the United Kingdom, as a maritime climate country, receives the effects of the ocean and the humidity is often at a high level.

4.4.2 conclusion. The fifth step of data exploration shows that when constructing the pollutant prediction model in the next step, the two features of humidity and pressure may have a positive influence on the accuracy of the prediction result. Therefore, these two features can be ignored when building a prototype model. It is more reasonable to consider the model optimization afterwards.

4.5 Correlation coefficient between wind direction change frequency and pollutants

In the sixth step, we want to explore the influence of wind direction change frequency on pollutant concentration.We speculate that the rapid change of wind direction may lead to a decrease in pollutant concentration, so the frequency of wind direction change may be negatively correlated with the concentrations.

4.5.1 description. We create a new feature according to wind direction. The new feature describing how many degrees changes within an hour can be counted as a change in direction. A correlation coefficient heat map is generated using the new feature and the average daily pollutant concentration. We found that the changes on wind direction has no obvious effect on the concentrations. When the wind direction changes 15 degrees, the correlation coefficient is close to 0, and when the changes is 5 degrees, the

wind direction change frequency is positively correlated with the nitrogen dioxide concentration, but with PM2.5 and PM10. The correlation coefficient between them is still close to zero. When the threshold is continuously reduced to 3 degrees, the positive correlation between the concentration of the three pollutants and the frequency of change of the wind direction is more significant.

4.5.2 conclusion. For the data exploration of the sixth step, we find that the frequency of change of wind direction is indeed the data related to the concentrations Meanwhile, when the changing threshold is smaller, the positive correlation is obvious.

5 FEATURE ENGINEERING

5.1 Basic features

5.1.1 Six time features. Firstly, We extract five time features from the time record corresponding to each air quality: year, month, week and hour. In addition, because we know that different holiday situations lead to different travel modes of the crowd, we also added the feature of holidays in the UK, and divided into three cases: working day, weekend and special holidays.

5.1.2 Five weather features. Then we calculate the nearest grid node from each of the grid weather points based on the longitude and latitude of each station. After completing this step, we can treat the weather data of the nearest grid point as the weather data of this air station, and merge five weather features with existing air quality and Time features. These five weather features are temperature, air pressure, humidity, wind speed, and wind direction.

5.1.3 Station index feature. Since each station is located in a unique location, which represents a unique distance from the city center and corresponds to a unique mode. So we added a station index feature to represent different stations.

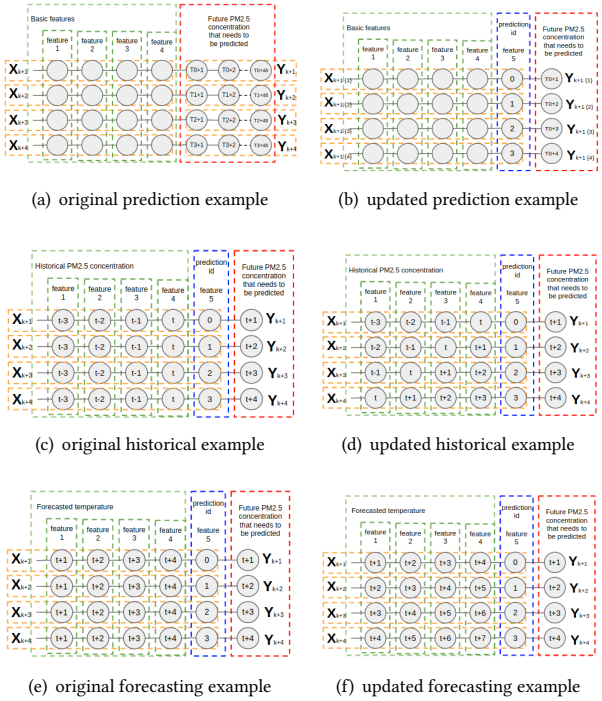
5.1.4 Four weather index features. We also added four weather index features. They are used to describe the proportional relationship between pressure and temperature (pressure/temperature), the proportional relationship between temperature and humidity (temperature/humidity), and the proportional relationship with PM2.5 concentration and PM10 concentration (PM25/PM10), as well as their accumulated concentration (PM25+PM10).

5.1.5 23 statistical features. To describe the long-term statistical features, after comparison, we selected the time window of two weeks. By calculating and extracting the statistical information of eight features in the last two weeks of each time point, we calculated the corresponding 23 statistical features. These 23 statistical features include the mean, maximum, and variance of the five weather features and three pollution concentrations over the last two weeks.

5.2 Timing features

5.2.1 Prediction id. The air quality needs to be predicted for 48 consecutive hours, which is shown as Fig. 7(a). Instead of letting the model train the next 48 hours of air data based on the features, which is only predicted once. According to our experiments, it is better to divide into 48 predictions, which is shown as Fig. 7(b).

Since we only predict one value at each time, we need to add a feature of prediction id to distinguish the predictions of different



(a) original prediction example (b) updated prediction example

(c) original historical example (d) updated historical example

(e) original forecasting example (f) updated forecasting example

Figure 7: Timing features

future time points under the same prediction moment. Here we have two options, one is to use the one-hot encoding, the other is to directly use 0-47 discrete values for encoding. In the beginning, we used the first method, but later we found that directly adopting the second method does not affect the results, and can greatly reduce the size of training and testing dataset.

5.2.2 96 historical features. Initially, it was to take the historical data of the past few hours as features directly, which is shown as Fig. 7(c). But the result was not good, which reflected the principle of GIGO (garbage in, garbage out). Then we updated these features, considering the target time of each prediction, and adjusting the location of features in each prediction, which is shown as Fig. 7(d). We were very careful to introduce predictions of future-time air quality and significantly increase the randomness of the original value to simulate the real predictions at each moment. The randomness introduced is much larger than the evaluation error we have in the testing dataset, so this simulation way can be considered as reasonable. And we compared and select the length of historical features as 12 hours. Finally, we have 12×8 features that represent the weather and pollution in the past half day.

5.2.3 60 weather forecasting features. Since we are unable to get the historical weather forecasting data directly, we also carefully used the processed weather data as the simulated weather forecasting data. Even though we introduced very large randomness in the simulation, our final model still achieved good results. If we can collect real weather forecasting data, certainly better results can be obtained. Similar to historical features, we take the same approach, the example of original forecasting features is shown as Fig. 7(e), and the updated forecasting features is shown as Fig. 7(f). The length is also 12 hours and finally we have 12×5 weather forecasting features.

5.3 Final Features

After careful evaluation, finally we didn't use some features such as extra pollution maps. Fig. 8 shows all of the final 198 features.

FeatureID	Description	FeatureID	Description
0	PM2.5 concentration	1	PM10 concentration
2	No2 concentration	3-5	temperature,pressure,humidity
6-7	Wind speed, direction	8	holiday
9-12	Month,week,day,hour	13	Station index
14-21	means of 0-7	22-28	Max of 0-6
29-36	Variance of 0-7	37-40	Index about weather
41-52	historical PM2.5 concentration	53-64	historical PM10 concentration
65-76	historical NO2 concentration	77-100	historical and forecasting temperature
101-124	historical and forecasting pressure	125-148	historical and forecasting humidity
149-172	historical and forecasting wind_speed	173-196	historical and forecasting wind_direction
197	prediction id		

Figure 8: FeatureID and description

6 MODELING

6.1 Linear regression model

6.1.1 Training and testing. In the linear regression model, the data of all 198 features were independent variables of X, and the data of PM2.5, PM10 and No2 concentrations were dependent variables of three different Y. In table 2, after training in the training dataset, we finally get relatively low errors in the testing dataset, which are 0.386 (PM2.5), 0.304 (PM10), and 0.272 (NO2).

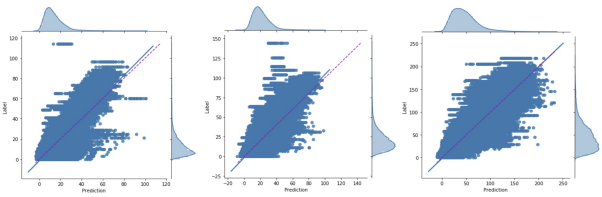


Figure 9: Performance in predicting PM2.5, PM10, and No2

6.1.2 Model exploration. This model has achieved relatively good prediction results. When we used the stochastic gradient descent algorithm to make the prediction, we found that the error would be 0.02-0.04 higher than this original model.

6.2 Neural Network model

6.2.1 Designing. In our Neural Network model, a regression predictive model which has two hidden layers, the optimizer is 'Adam'. During our exploration, we found that adding more hidden layer has a small impact on the performance but will slow down the calculation. Therefore, the model retains two hidden layers but would adjust the number of neurons based on feature selection.

6.2.2 Feature selection. Fig. 8 shows that there are 198 features. In this part of work we test the utility of the following features: pollutant concentration, weather features, and historical features respectively. Pollutant concentrations are feature ID 0 to 2, weather features are feature ID 3 to 8, historical concentration features are feature ID 41 to 76 and historical and forecasted weather features are feature ID 77 to 196. We extract 10 thousand data and separate them into four sets. The training episodes are 40 and the performance is shown in Table 1. Compared with other features, historical concentration features have the best performance.

error	All features	ID: 0-2	ID: 3-8	ID:41-76	ID:77-196
PM25	0.58	0.69	0.66	0.57	0.62
PM10	0.59	0.79	0.66	0.62	0.67
NO2	0.38	0.35	0.33	0.32	0.37

Table 1: Evaluation errors of different NN models

6.2.3 Model exploration. The performance of a neural network model is not good. Considering that the data are time series, we also tried to implement an LSTM (long short-term memory) model, which is a Recurrent Neural Network model. We tested the model by using original data rather than processed data. The result was not good, and the error was up to 1.65.

6.3 Xgboost model

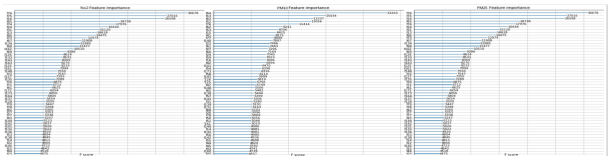
6.3.1 Optimization. In training phase, we use gridSearchCV (cross-validated grid-search) [5] to select the parameters and optimize the XGboost regression model. First, fixing learning rate and number of estimators for tuning tree-based parameters. Second, fixing the gamma parameter which can make the algorithm conservative. Third, fixing alpha value which is L1 regularization term on weight. Fourth, fixing subsample and colsample_bytree which Denotes the fraction of columns to be randomly sampled for each tree and Denotes the fraction of observations to be randomly sampled for each tree, also lower values make the algorithm more conservative and prevents over-fitting. Finally, we get the best parameter in learning_rate=0.001, n_estimators=3000, gamma=0.8, reg_alpha=0.001, subsample=0.8, colsample_bytree=0.5, which lead to 0.02 ~ 0.10 SMAPE improvement on validation dataset.

6.3.2 Feature selection. Firstly, we train our model with all the features, and we get a feature importance map for 3 different models in figure 10(a). Then we try to remove the feature whose F_score is under 3000. For PM2.5, we remove the feature like holiday, month, week, day. We only leave hour as the time feature. Also, removing historical and forecasting pressure, temperature and wind_direction feature. For PM10, we remove some features, like index about the weather, some of the historical and forecasting humidity. For No2, we remove some features about PM2.5, PM10 concentration.This lead to 0.02-0.04 improvement on SMAPE of the validation set.

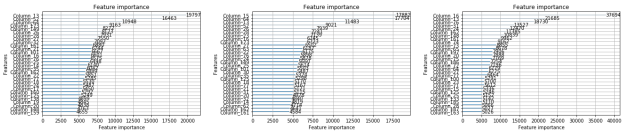
6.3.3 Conclusion. After selecting features and Finally, the results of Xgboost model improve from 0.02 to 0.33 and it is shown in table 2. We can find what the feature about PM2.5, No2, and PM10 concentration have an important weight. Also, temperature, pressure, and humidity are also important. And the weekday, holiday do not have a strong influence on the PM2.5. We can infer that air quality is not strongly influenced by 2.5. For PM10, out of shock, the PM10 concentration is not important but the historical PM10 and weather detail has an important weight. We can infer that PM10 is the main particle in London. For No2, we can find that the PM10 concentration will influence the concentration of No2. And the xgboost model is fit for predicting PM10 and NO2.

6.4 Lightgbm model

We optimized our Lightgbm model based on the same grid search approach used in the xgboost model. For the different parameters



(a) xgboost feature importance of predicting PM2.5,PM10,NO2



(b) lgb PM2.5 (c) lgb PM10 (d) lgb NO2

Figure 10: Feature importance of xgboost and lightgbm

of the Lightgbm model, we perform grid search on different parameters in order according to the official guide [4]. The feature importance (top30) of our optimized model is shown in Fig. 10(b), 10(c), and 10(d). We can see that the station index feature plays an important role in these three models, especially in predicting the PM2.5 concentration. In other words, PM2.5 is the most difficult to predict. Through data exploration, we know that the reason may be that the changing of PM2.5 is the least obvious one.

The errors in the testing dataset is shown in table 2. Compared with XGboost model, Lightgbm model had faster training speed and better testing results. Among four individual models, it achieved the best scores. Compared with XGboost model, which trained more than 33 hours, our Lightgbm model just need to train in 1 hour, even the size of the training data is around 10 GB.

6.5 Model ensemble

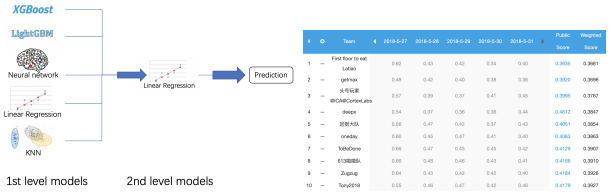
After the experiment of training with different models, we can get the results on the testing dataset, which is shown in table 2. Among the three pollutants, all of the models achieved higher scores in predicting the concentration of NO2. According to data exploration, we know that the change range of NO2 is more significant than the other two pollutants, so this helps our models to catch the pattern of the changing. If we consider further, the main reason for NO2 is vehicle exhaust, which makes the factors such as commuting time more important, so its mode is easier to find out by the same model.

MODEL	Evaluation	No2	PM2.5	PM10
LinearRegression	SMAPE	0.272	0.386	0.304
NeuralNetwork	SMAPE	0.38	0.58	0.59
XGboost	SMAPE	0.342	0.466	0.384
Lightgbm	SMAPE	0.258	0.336	0.292
Ensemble	SAMPE	0.241	0.301	0.282

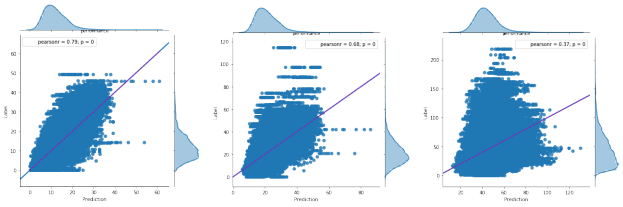
Table 2: Evaluation error of each model

6.5.1 Blending. In order to ensemble all the model together. We use the simple ensemble method, blending. Table 2 shows the blending result. We use the first level models predictions as the input for the second level model. Using this way, the second level model will basically use the first level model predictions as features and learn where to give more weight. To use this technique we also need to use the 1st level models and make predictions on the testing dataset so we can use them on the second level model. We can also pass complete validation set with extra features, like first level model predictions, to the second level model. Therefore, the model can do more work on finding solution and optimization. According to the results in Table2, we try to use a linear regression model in the second level to combine the first level model predictions. The ensemble structure is shown as figure 11(a). The linear regression model will combine the other ones to make an overall better prediction hopefully. We find that the result of NN is not good, so we only combine Xgboost, Lightgbm, KNN, linear regression.

6.5.2 Conclusion. After ensemble, the performance also shown in Fig. 11(c), 11(d), and 11(e). The closer the points are to the middle dashed line the better are the predictions. The ensemble model has a better performance than others.



(a) Ensemble architecture (b) Performance of top 10 teams



(c) Performance of PM2.5 (d) Performance of PM10 (e) Performance of NO2

Figure 11: Ensemble model

7 CONCLUSION

Figure 11(b) shows that top 10 teams achieved the error around 0.4 (Beijing + London), and we achieve better error around 0.28 in the testing dataset (London). Due to the limitation of our running machine, we couldn't handle data with a larger size, however, we consider many possible improvements based on our final model. Our final model can be improved by adding additional position features, adding additional automatic randomly generated training data, and using other ensemble approaches. Moreover, we can use the model iteration method to iteratively get better performance. We can continuously use the final prediction of the model to replace the step of future-related concentration features in the section 5.2.2.

REFERENCES

[1] 2018. 2018 KDD CUP of fresh air. https://biendata.com/competition/kdd_2018/. (2018). accessed 10-May-2019.

[2] 2018. Symmetric mean absolute percentage error. https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error. (2018). accessed 10-May-2019.

[3] 2019. air-quality-prediction. https://github.com/data-mining-not-found/air-quality-prediction. (2019). accessed 10-May-2019.

[4] 2019. Parameters Tuning. https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html. (2019). accessed 10-May-2019.

[5] Weizeng Wang, Yuliang Shi, Gaofan Lyu, and Wanghua Deng. 2017. Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform. *DEStech Transactions on Computer Science and Engineering* aiea (2017).