参赛队员姓名：刘子灏

中学：北京海淀国际学校

省份：北京

国家/地区：中国

指导教师姓名：徐瑶

指导教师单位：北京海淀国际学校

论文题目：PM2.5 Density Prediction based on a Two-Stage Rolling Forecast Model using LightGBM

PM2.5 Density Prediction based on a Two-Stage Rolling Forecast Model using LightGBM

Zihao Liu

ABSTRACT

At present, air pollution is a primary issue of the world. Particularly, PM2.5 pollution can cause severe impact on economy and human health, so developing an accurate PM2.5 prediction model becomes a hot topic. Up to now, researchers had developed PM2.5 forecasting methods based on decision tree models, RNN models, and hybrid models. Previous works also discovered plentiful features, such as seasonal data and weather forecasting data, that help increase the accuracy of PM2.5 prediction. To improve the model accuracy, we developed a LightGBM-based PM2.5 prediction model that has two innovations: 1) our model studies how special events (e.g. diplomatic visits, sport events, and government meetings) influence PM2.5 variation. 2) our model adopt the strategy of two-stage rolling forecasting so that it can achieve high accuracy without relying on weather forecasting data.

Keywords: Short-time Fourier transform, LightGBM, PM2.5, Rolling forecast, Time series, What-if analysis

CONTENTS

I. INTRODUCTION

Since the writing of Silent Spring in 1962, environmental protection has become a primary issues all over the world, especially in developing countries. In particular, PM2.5 pollution is one of the environmental issues that have caused severe impact to people's daily life [1]. Although PM2.5 can directly cause health problems on the population, it also has impacts on economy. In fact, a study conducted by Xie *et al.* [11] indicates that health and economy issues at become substantial when PM2.5 concentrations are high at a provincial level. Specifically, it was projected that China will have to spend additional $25.2 billion on health expenditure for diseases brought by PM2.5 pollution. However, if PM2.5 is properly manipulated with air pollution control technology, by 2030 the number of patients due to PM2.5 will reduce by 75%, reducing PM2.5 health expenditure $6.5 billion. Accordingly, it is vital to develop solutions to the PM2.5 issues for their severe impacts in health and economy.

To minimize the damage of PM2.5, researchers in this field have developed various models based on the theory of time series to forecast air quality. Currently, most of the proposed models either use simple regression models, deep neural networks, or hybrid models that make use of both.

Typical regression models used for PM2.5 prediction include linear regression, decision tree, and random forest; but most authors adopt tree-based models to forecast future PM2.5 concentration. For example, Zhang *et al.* [2] proposed a LightGBM forecasting model that utilized past air quality data recorded by monitoring stations and meteorological data provided by weather forecasting. Similarly, Zhang *et al.* [3] suggested a PM2.5 forecasting model based on random forest algorithm. In fact, Lee *et al.* [4] analyzed the performance on PM2.5 prediction of all tree-based models such as XGBoost, LightGBM, and random forest and obtained promising results, explaining why tree-based prediction models are common in the field of PM2.5 forecasting.

Deep learning models include the traditional feed-forward neural networks, convolutional networks, and recurrent neural networks, but most deep-learning-based research done on PM2.5 forecasting make use of RNN-based networks. For instance, the STE (Spacial-Temporal Ensemble) model which makes use of temporal features, proposed by Wang and Song [5] is an LSTM-based algorithm. Likewise, Ong *et al.* [6] proposed an RNN-based PM2.5 quality prediction model with improved training methods. Particularly, they developed a new pre-training method that allows the model to make more accurate predictions. Essentially, researchers tend to use RNN-based networks to develop deep learning PM2.5 prediction models.

Not only did researchers developed model solely based on tree-based models or recurrent networks, but also created hybrid models that use both. For instance, Zheng *et al.* [13] proposed a hybrid PM2.5 forecasting model in which a specific mechanism is set up to combine predictions made by a linear regression model and a neural network. Similarly, Qi *et al.* [14] proposed a PM2.5 prediction model based on graph convolutional neural networks and LSTM. Furthermore, the methods Qi *et al.* [14] developed are also applicable to the prediction of other particles' densities. Likewise, Zhang *et al.* [15] proposed a hybrid prediction model that combines convolutional networks and recurrent networks. Particularly, Zhang *et al.* discovered that their model produces more accurate results than others when there are large fluctuation of values in the data.

Previous researches done on the field of PM2.5 forecasting have produced plenty of discoveries. For example, Zhang *et al.* [2] enhanced model accuracy by integrating the model with data obtained from weather forecasting stations. Likewise, Zhao *et al.* [7] discovered that seasonal data will improve the accuracy of their linear-regression-based PM2.5 prediction model. Similarly, Zhang *et al.* [3] proposed a random-forest-based PM2.5 prediction model that can effectively handle large-quantity data. Although these researches have already yielded remarkable results, their models require knowing weather forecast data and concentrations of pollutants other than PM2.5. In addition, current forecasting models did not take the impact of government's policy on PM2.5 concentration into account. Furthermore, previous works on PM2.5 prediction do not perform what-if analysis on their prediction models. As a result, we develop solutions to each of these issues.

In the face of the aforementioned potential improvements, we propose our own PM2.5 prediction model[1]. Similar to previous works, our model adopts the LightGBM framework for its efficiency and flexibility. [9] In our study, we use the PM2.5 data set published by Liang *et al.* [8] for its granularity: the data set contains hourly records of PM2.5 and other meteorological parameters between the year 2010 and 2015. The density of this data set allows us to predict PM2.5 concentration in terms of hours. Nevertheless, the attributes in the original data set is not sufficient to train a prediction model with high accuracy, so we used a variety of techniques to generate more features. To let our model study the relationship between the attributes and

[1] The source code of our model is available at https://github.com/TravorLZH/pm25

their past states, we generate lagged features for PM2.5 concentrations and meteorological parameters. According to Zhang *et al.* [2], statistical features over rolling window are also helpful in PM2.5 prediction, so we generate them as well. In addition to the above time-domain features, we generate frequency-domain features via short-time Fourier transform on the time domain of the PM2.5 time series so that our model can study the seasonality of PM2.5 concentration, therefore improving its accuracy in prediction.

In their feature engineering process, Zhang *et al.* [2] generate a date-time feature called "is_weekend" which denotes whether the day predicted is weekend. This innovation motivates us to dig further in date-time feature generation. By tracking the government's event calendar, we are able to study the impacts of government's policies on PM2.5 since governments always make event-specific policies during these special time intervals. In fact, government's policies have substantial impacts on PM2.5 concentrations. For example, in 2013, the Chinese government imposes "Air Pollution Prevention and Control Action Plan," aiming to reduce PM2.5 concentrations of China's major cities by more than 10%, and investigation led by Zhang and Di [12] discovers that China's PM2.5 concentrations decreases substantially during 2013 and 2017. To track Beijing's special events, we manually collected date intervals during which special events occurred in Beijing from 2010 to 2015 using search engine and created a new data set. Not only does this "special event" data set contains the date intervals, but also labels the recorded special events by type. Types of recorded special events include celebration of special festivals (e.g. 2015 military parade to celebrate the allied victory over fascism), diplomatic visits (e.g. Michelle Obama's visit to China in 2014), and sport events that occurred in Beijing. Since governments often made special policies during these events, incorporating "special event" features into our data set allows the prediction model to learn how government's policies affect PM2.5 concentration.

Because our work is independent of weather forecasting, our testing set does not contain meteorological features. In order for our model to predict PM2.5 concentration, we use the following steps to handle this conflict: first, we use a separate model to predict meteorological features in the testing set. Then, we generate statistical and lagged features from the predicted meteorological parameters. Lastly, we use the technique of rolling forecasting to predict PM2.5 concentrations.

To test the effectiveness of our proposed new features, we perform control experiments. Particularly, we train two models: one with the new features and one without. Subsequently, we compare them using evaluation metrics such as mean absolute error, root mean square error, and symmetric mean absolute percentage error defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - p_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2} \tag{2}$$

$$SMAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - p_i|}{(|y_i| + |p_i|)/2} \tag{3}$$

To determine how influential governments' policies are during special event, we evaluate the performance of the two models during special event time intervals. In addition, we perform feature importance analysis on the PM2.5 prediction model to determine what features contribute most to our forecasting model.

II. PROBLEM SETUP

In this study, we aim to create a PM2.5 prediction model that guides local governments to make effective policies to improve PM2.5 quality while being independent of weather forecasting data. To implement, we develop a PM2.5 prediction model that requires knowledge of government's event calendar. Since governments make policies to control traffic and other aspects of a city when the city is holding special events, accessing government's event calendar allows our model to study the impacts of such policies. Not only do governments make special policies during special events but also enact policies during holidays, so we also incorporate "is_holiday" attribute denote whether the entry was recorded in a holiday. In order for our model to predict PM2.5 without weather forecasting data, we adopt the technique of rolling forecasting. As shown in Figure 1, our model first predicts meteorological features of one future hour, and then predicts PM2.5 concentrations using these predicted values. Subsequently, we will compare the prediction results of the control group and the experimental group using evaluation metrics such as MAE, RMSE, and SMAPE.

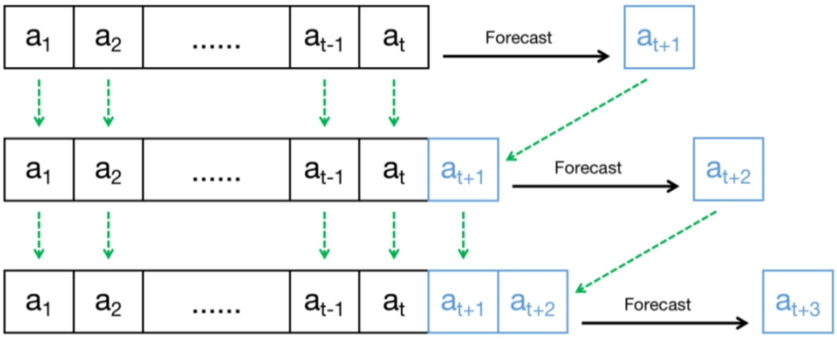


Figure 1. Illustration of rolling forecasting

III. DATA SET

A. Data description

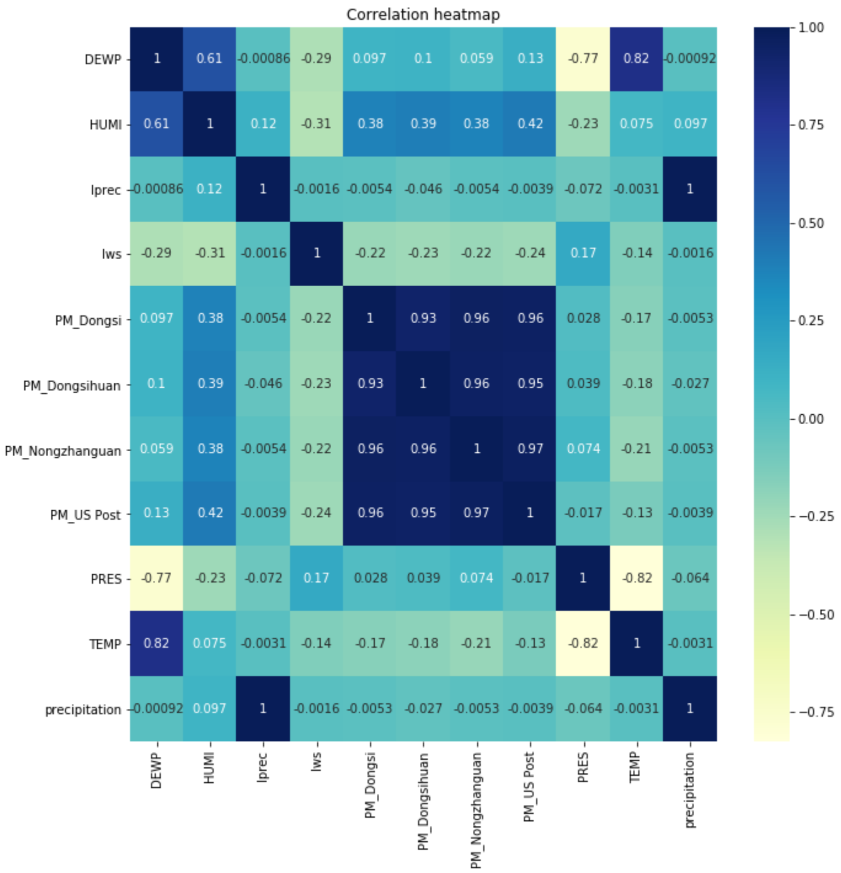


Figure 2. Correlation matrix of the data set

In this study of PM2.5 forecasting, we used hourly recorded PM2.5 data provided by Liang *et al.* [8]. According to Table I, the data set contains PM2.5 concentration measured from four different stations: Dongsi, Dongsihuan, Nongzhanguan, and U.S. diplomatic post. In addition to PM2.5 attributes, the data set also contains several hourly recorded meteorological attributes: wind speed, wind direction, precipitation, air pressure, humidity, dew point, and temperature.

However, this data set contains redundant attributes, anomalies, and missing values [8]. As a result, we perform some preprocessing before conducting further analysis.

Table I. ATTRIBUTES OF THE ORIGINAL DATA SET

Classification	Attributes	Description
Date-time attributes	year, month, day, hour	The exact time of the data entry
	season	Current season
Meteorological attributes	HUMI	Humidity
	TEMP	Temperature
	PRES	Atmospheric pressure
	precipitation, Iprec	Precipitation
	Iws, cbwd	Wind speed and wind direction
	cbwd	Wind direction
PM2.5 concentration values	PM_Dongsi	PM2.5 from Dongsi observatory
	PM_Dongsihuan	PM2.5 from Dongsihuan observatory
	PM_Nongzhanguan	PM2.5 from Nongzhanguan observatory
	PM_US Post	PM2.5 from U.S. diplomatic post

B. Data exploration

In truth, data preprocessing described in the next section allows us we can dig deeper into our data set since we have eliminated anomalies and imputed missing values. Exploring the data set allows us to discover various properties of PM2.5 concentrations and other meteorological attributes.

1) Relationship between dew point and humidity

In our data set, dew point and relative humidity are stored as DEWP and HUMI, and according to Figure 3, distribution of DEWP and HUMI appear to be very similar. In addition, correlation analysis in Figure 4 reveals that there is a positive correlation between DEWP and HUMI. That is, dew point grows large as relative humidity increases, and dew point becomes low as relative humidity decreases. This explains why both dew point and relative humidity are often used to reflect the amount of moisture in the atmosphere [10].

2) Temperature distribution in Beijing

The data set we use also contains a complete hourly record of Beijing’s temperatures during 2010 and 2015. Studying its distribution allows us to discover some weather facts in Beijing. According to Figure 3, there are two maxima in the distribution of temperature: one at approximately 22 degree Celsius above zero, the other at approximately 3 degree Celsius below zero. Since Beijing in the northern hemisphere, its temperature in summer is greater than that in winter. As a result, we conclude that Beijing is very cold in winter while not so hot in summer.

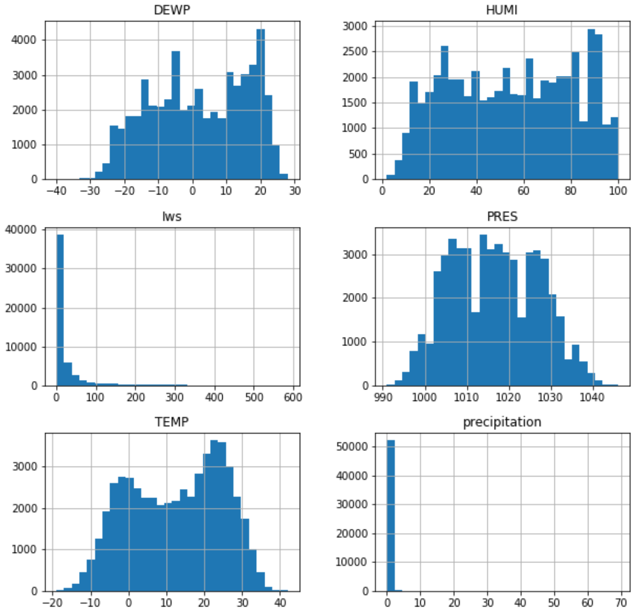


Figure 3. Histograms of different meteorological parameters in the imputed data set

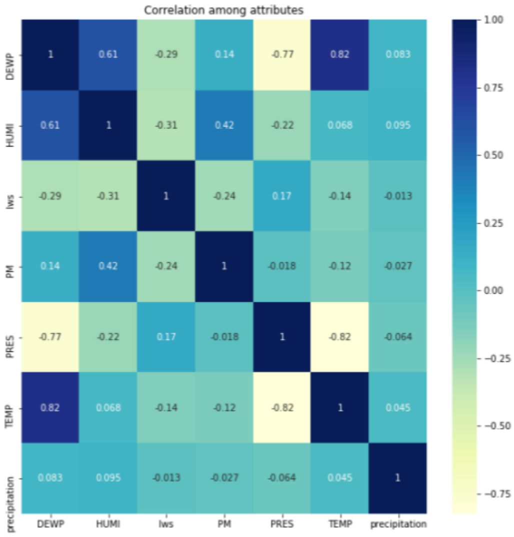


Figure 4. Correlation analysis on the imputed data set

3) Distribution of PM2.5 in Beijing

According to Figure 5, Beijing's PM2.5 concentrations appear to follow an exponential distribution. That is, the frequency of PM2.5 records decrease exponentially as the PM2.5 values increase. This means that most of the PM2.5 concentrations are low, implying that Beijing experiences little PM2.5 pollution most of the time during 2010 and 2015.

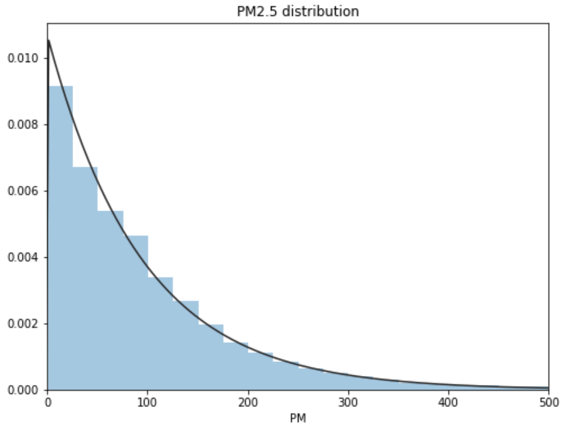


Figure 5. Distribution of PM2.5 in the imputed data set

IV. METHODS

A. Data preprocessing

In our study, we use regression models to predict future PM2.5 concentrations. Specifically, we are using our regression models to predict future PM2.5 values via studying the original data set. To make our model efficient, we need to reduce the dimensions of our daa set. In order for our model to make more accurate predictions, we need to ensure our data set does not contain any anomalous or erroneous values. In this section, we scrutinize the data set provided by Liang *et al.* [8] and develop our own solutions to reduce data dimensions, process outliers, and impute missing entries.

1) Handling redundant attributes

According to Figure 2, “Iprec” and “precipitation” attributes are very similar. In fact, we perform a more definitive comparison, realizing that more than 95% values of “Iprec” and “precipitation” are identical. As a result, we detach “Iprec” column from the data set. Figure 1 also implies that the correlations among PM2.5 values measured from different stations are highly similar to each other, so we decide to keep only one PM2.5 record. By the analysis in Figure 6, we decide to preserve the U.S. diplomatic post’s version of the PM2.5 record since its data are the more complete than those of other observatories.

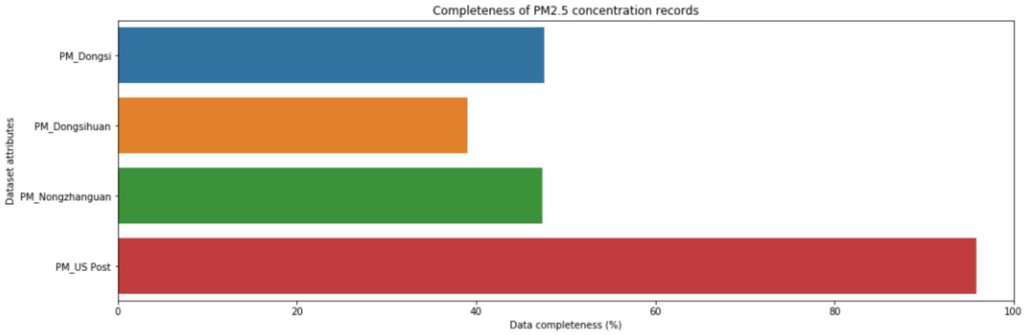


Figure 6. Completeness of PM2.5 data

2) Outlier removal

The data set also contains anomalies. For example, we discovered that the data set contains precipitation record of 999990 millimeters, which is certainly impossible to achieve. As a result, we set up a scheme to identify anomalous values. Similar to Zhang *et al.* [2], we regard PM2.5 data values greater than 500 $\mathrm{ug/m^3}$, precipitation values exceeding 400 millimeters, wind speed that tops 500 m/s, and air pressures that go beyond 2000 kPa as anomalies. After identifying these anomalies, we replace them with NAN and refill them using imputation techniques.

3) Missing values imputation

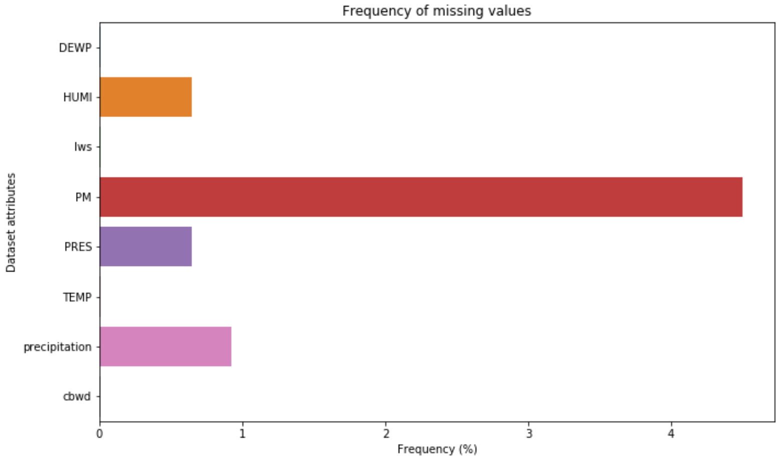


Figure 7. Percentage of missing values in each attribute

Not only does the data set we use encompass anomalies, but also missing values [8]. According to Figure 7, the number of missing values differ among each attribute, so we need to impute them separately. Since missing values are distributed randomly in the data set, we adopt different methods to handle them. Common methods for imputation include Next Observation Carry Backward (NOCF), Last Observation Carry Forward (LOCF), and linear interpolation [16]. Specifically, We use NOCF to impute missing values occurring at the beginning of the data set's record. For missing values at the tail of the record, LOCF is applied. At last, we handle missing values in between data entries with linear interpolation.

4) Processing categorical features

Not only does our data set contain numerical features but also categorical features. That is, some attributes of our data set can only have discrete values instead of continuous real numbers. For instance, date-time features can only have integer values. Although these categorical features can only be integers, they can be easily integrated into our regression model since integers are also numbers. However, the data set also contains categorical features that are not integers. For instance, wind direction attribute "cbwd" only stores string values (namely, NE for northeastern wind, SE for southeastern wind, etc.). As a solution, we map these string values to distinct integers so that wind directions can be integrated into our prediction model.

B. Feature engineering

Liang *et al.* [8] have shown that PM2.5 concentrations are highly related to meteorological parameters, which motivates us to use meteorological parameters as one of the input variables for our PM2.5 prediction model. However, only using these attributes cannot produce high-accuracy prediction , and our means to improve model accuracy is to perform feature engineering. Time series itself can tell many information beyond the data values themselves. To study the relationship between PM2.5 concentrations and special occasions such as government meeting, holidays, diplomatic visits, and sport events, we generate additional date-time features with the help of external tools. To study the relationship between time series and itself, we extract lagged features and statistical features using sliding window mechanism. To study the seasonality of time series, we generate frequency-domain features via short-time Fourier transform. All features generated in our feature engineering process are listed in Table III.

1) Holiday and special event features

One main goal of our work is to study the influence of government's policies on PM2.5 concentration curve, and our means to investigate it is by creating date-time features. Motivated by Zhang *et al.*'s "is_weekend" features [2], we generate "is_holiday" attribute via using the external package "chinese-calendar"[2] so that our model can study the correlation between PM2.5 concentrations and holiday. In addition, we manually collected time intervals during which special events such as government meeting, sport events, and diplomatic visits occurred

[2] It's an open source project at https://github.com/LKI/chinese-calendar

in Beijing during 2010 and 2015 from year tables provided by Baidu Baike. An excerpt of these collected events is available in Table II. To integrate this special event data set into our feature data set, we create a Boolean column named "is_special_event" to denote whether the specific hour falls within any of the special event intervals of our special event data set.

Table II. SELECTED SAMPLES FROM SPECIAL EVENT DATA SET

Category	Event description	Starting date	Ending date
meeting	First meeting of 12th National People's Congress	Mar 5, 2013	Mar 20, 2013
visit	South Korea's President Lee Myung-bak visits China	Jan 9, 2012	Jan 11, 2012
sport	Opening of 2010 Chinese Football Association Super League	Mar 27, 2010	Mar 28, 2010
meeting	Second round of Sino-US military and economic dialogue	May 24, 2010	May 25, 2015

2) Lagged features and statistical features

Not only do we integrate features from external sources into our data set but also decompose the original data attributes to obtain new features. Correlation analysis in Figure 4 reveals that time series in our data set are not completely independent of each other, and auto-correlation analysis in Figure 8 shows that time series are not independent of themselves either. To let our model study this relationship, we generate lag features with a period of 48 hours. Zhang *et al.* [2] shows that statistical features are helpful in PM2.5 forecasting, so we also generate statistical parameters such as mean, minimum, and maximum over lagged features.

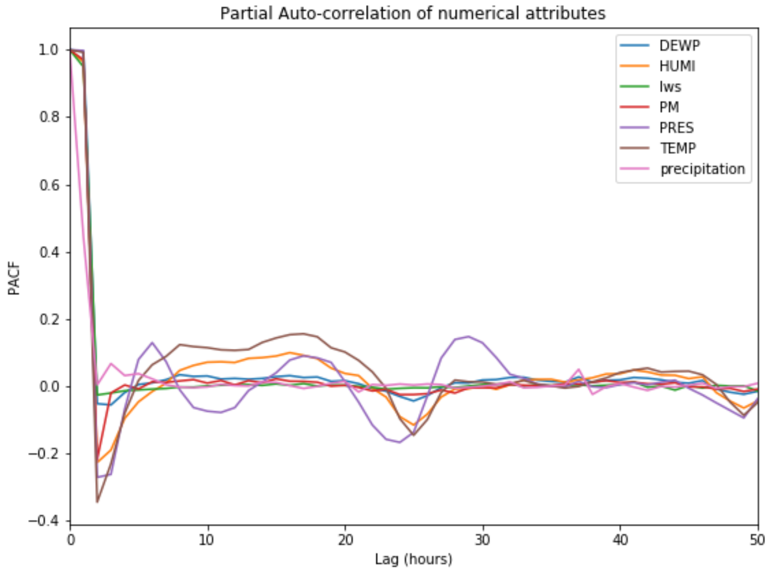


Figure 8. Partial Auto-correlation analysis on the data set

3) Frequency-domain features

Time series varies with seasonality. A prediction model that understands the seasonality of PM2.5 time series may produce more accurate results than a model that does not study PM2.5's seasonality, and one way to extract seasonality of a time series is to analyze its spectrum. That is, we perform short-time Fourier transform on PM2.5 time series and incorporate the generated coefficients into each data entry. According to Figure 9, Fourier coefficients of PM2.5 time series decay as the frequency increases, indicating that Fourier coefficients at high frequencies contain less useful information than those at low frequencies. As a result, we truncate the spectrum and only store Fourier coefficients of frequencies less than or equal to 0.02 units into our data set.

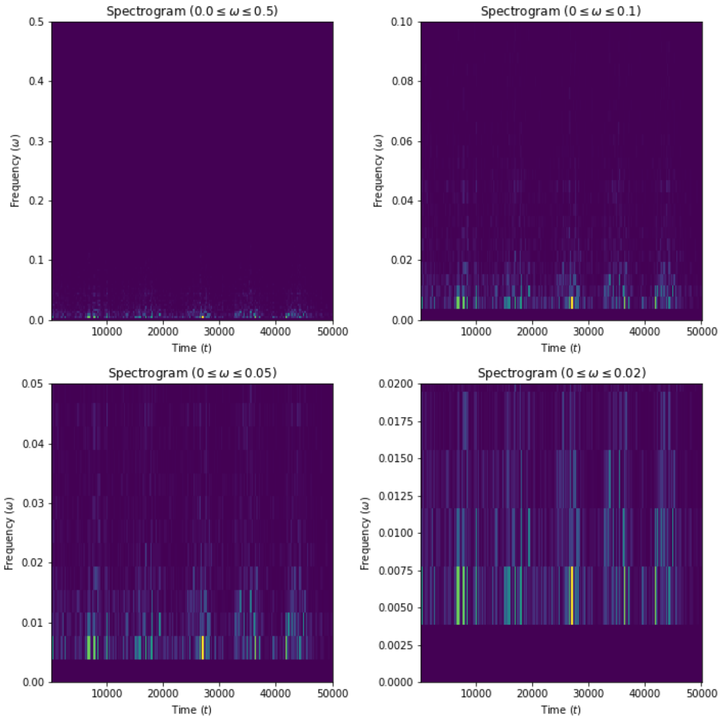


Figure 9. Spectral analysis on PM2.5 time series

Table III. LIST OF NEWLY GENERATED FEATURES

Category	Feature	Description
Date-time features	is_holiday	Whether the entry is recorded during a holiday
	is_special_event	Whether Beijing held special events
Lagged features	{}_1, {}_2, {}_3 ... {}_48	Lagged meteorological and PM2.5 features with a size of 48 hours.
Statistical features	{}_mean, {}_min, {}_max	Local mean, minimum, and maximum on lagged features
Frequency-domain features	stft_1, stft_2, stft_3 ... stft_134	Truncated spectrum of PM2.5 time series

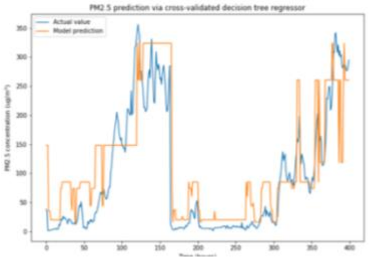


Figure 10. PM2.5 prediction from decision tree

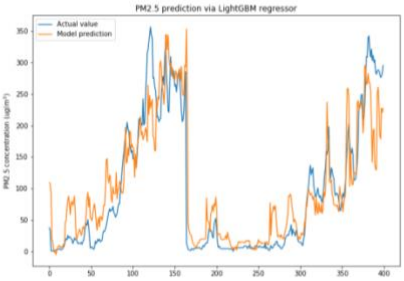


Figure 11. PM2.5 prediction from LightGBM

4) Features from "tsfresh" package

In addition to the aforementioned features, we also considered incorporating features generated by a feature extraction package named "tsfresh." According to Christ *et al.* [17], this package extracts features from the data set using scalable hypothesis test. That is, a concrete mathematical method is used to determine whether a specific feature is relevant to (i.e. helpful for creating prediction on) the specified target variable. Examples of extractable features include lagged features, and wavelet features [18]. In our case, "tsfresh" extracts 122 relevant features for PM2.5 prediction. However, principal component analysis on these newly generated features reveals that the dimensions of "tsfresh" features can be virtually be reduced into 13. By analyzing the these 13 principal components, we discovered that they are highly related to features that are already in our data set. For instance, as shown in Figure 12, the largest principal component is almost identical to PM2.5 concentration trend. Therefore, we choose not to incorporate "tsfresh" features into our project.

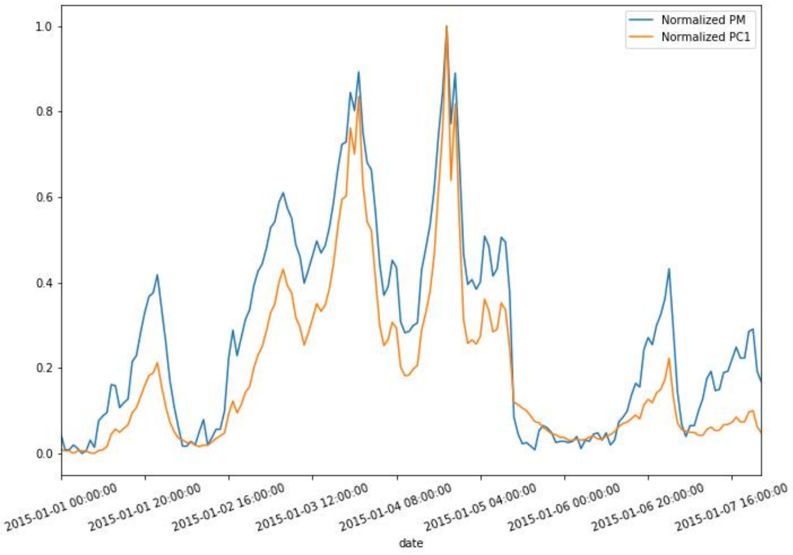


Figure 12. Selection of normalized PM2.5 curve and normalized PC1 curve

C. Forecasting Method

1) Model selection

Introduced by Ke *et al.* [9], LightGBM is a machine learning model based on gradient-boosting decision tree (GBDT) and XGBoost. Compared to traditional regression models such as decision tree, LightGBM is more efficient in training, especially when the amount and dimensions of the data are large. In fact, as shown in Figure 10 and Figure 11, LightGBM produces better prediction when the data contains many fluctuations. As a result, we decide to build our prediction model using LightGBM as the basis for its efficiency and accuracy.

2) Configuration

To test whether rolling forecast helps the model provide more accurate PM2.5 prediction, we set up a baseline group that does not use rolling forecast. Instead, the baseline group uses data set features of the current day to predict PM2.5 on the next day. To verify how government policies affect the PM2.5 curve, we perform comparison on the control group and the experimental group. As shown in Table IV, the only difference between the control and experimental models is that the experimental group inputs holiday and special event features whereas the control group does not.

Table IV. CONFIGURATION OF THE EXPERIMENT

Model	Input features
LightGBM without rolling forecast (baseline)	Features from original data set, lagged features, and statistical features. No rolling forecast
LightGBM (control)	Features from original data set, lagged features, and statistical features
LightGBM+holiday and special event (experimental)	Features from original data set, lagged features, statistical features, and holiday and special event features

3) Two-stage Forecasting

For each model in our experiment, the same process is followed so that the only differences between the models are in the input features and whether using rolling forecast. To effectively verify the impact of government policies, we decide to let the control model predict PM2.5 values between Aug 20, 2015 and Sept 10, 2015. Because our study is independent of weather forecasting data, the prediction model itself needs to forecast meteorological parameters before predicting PM2.5 concentration values. Hence, a two-stage rolling forecast process is developed for the model to make predictions:

1. The model learns the training set containing the observed data and generated features
2. As illustrated in Figure 13, the model first predicts meteorological features, then generate lagged features and statistical features, and, at last predicts PM2.5, and this

process is executed for 24 times so that the hourly PM2.5 concentrations of the next day is predicted.

3. After one day, the model generates features from actual observation of meteorological features and PM2.5 data during the day and store them into the data set.
4. Retrain the model with the updated training set so that it becomes ready to predict PM2.5 of another new day.

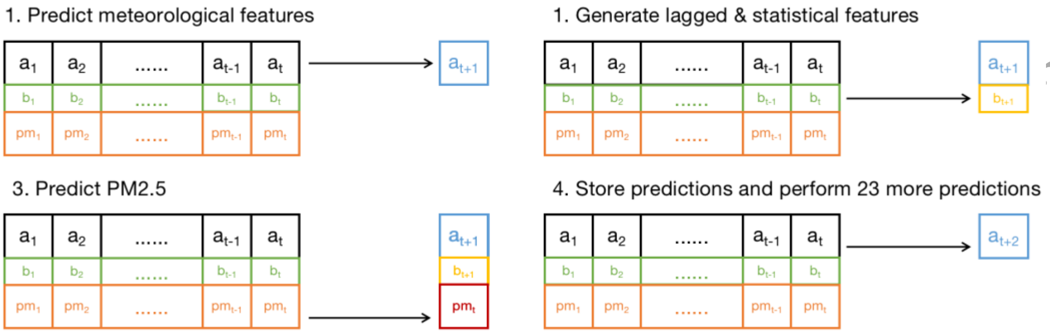

1. Predict meteorological features
a_1 a_2 ...... a_{t-1} a_t
b_1 b_2 ...... b_{t-1} b_t
pm_1 pm_2 ...... pm_{t-1} pm_t
a_{t+1}
1. Generate lagged & statistical features
a_{t+1}
b_{t+1}
3. Predict PM2.5
a_{t+1}
b_{t+1}
pm_t
4. Store predictions and perform 23 more predictions
a_{t+2}

Figure 13. Procedure for PM2.5 prediction

V. Performance Evaluation

After performing a series of experiments, we obtain three arrays of PM2.5 prediction: one from the baseline group that does not use rolling forecasting, one from the control group, and the other from the experimental group that incorporates holidays and special event features. As shown in Figure 14, the control group is already able to make such an accurate prediction compared to the PM2.5 values actually recorded from the observatory, implying that the performance of our groups in the experiments cannot be solely determined by the visualizations. Consequently, we compare them using the evaluation metrics mentioned above in Eq. (1), (2), and (3).

A. Model comparison

In the study, we compare our prediction models with the LightGBM proposed by Zhang *et al.* [2] that uses weather forecasting data as inputs in addition to lagged features and statistical features. Although Zhang *et al.* used a different data set and configuration in their research. The comparison is still helpful on measuring how rolling forecast improves model accuracy. As Table V suggests, although incorporating weather forecasting indeed helps improve the accuracy of model prediction, the rolling-forecasting-based model we proposed yields substantially better results. In spite of the fact that the MAE scores worsens when comparing the baseline group and the control group, differences in RMSE and SMAPE scores between the baseline group and the control group reveals are substantially greater than those between Zhang *et al.*'s models. This phenomenon indicates that it is possible to make more substantial improvements on model accuracy without learning the weather forecasting data. In our case, we use rolling forecasting to fill in meteorological parameters and eventually achieve a more accurate prediction. Also illustrated in Table V is that the experimental group makes substantially more accurate PM2.5 prediction than does the control group. To determine how a change in government's event calendar, we perform what-if analysis on the experimental group.

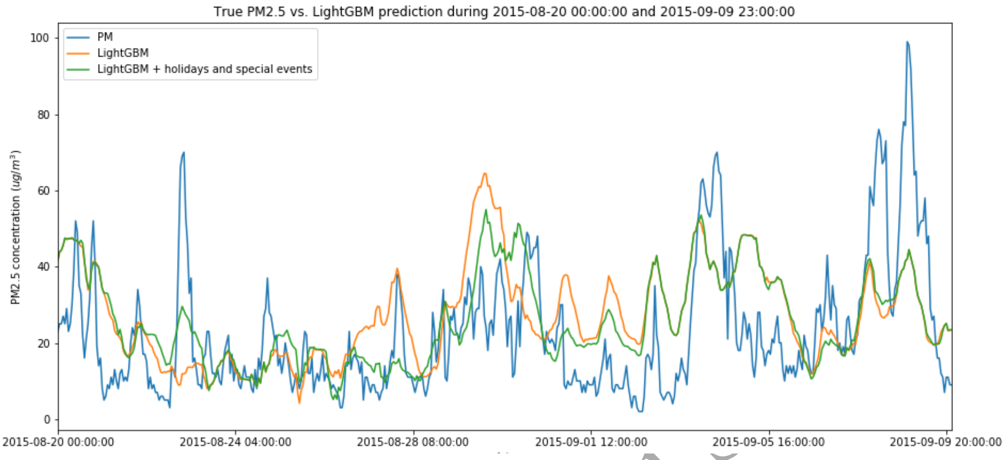
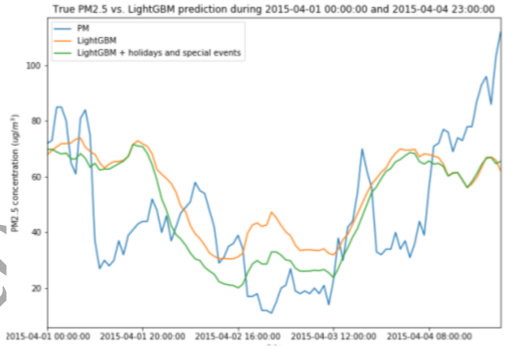


Figure 14. Comparison between the true PM2.5 values and the models' predictions

Table V. COMPARISON AMONG THE MODELS

Model	MAE	RMSE	SMAPE
LightGBM without rolling forecast (baseline)	12.9255	18.1800	0.5695
LightGBM (control)	13.2621	17.1323	0.5453
LightGBM + holidays and special events (experimental)	11.7570	15.2813	0.4994
Zhang *et al.*'s model [2]	26.4359	32.8711	0.4229
Zhang *et al.*'s model without weather forecast [2]	26.6824	33.8922	0.4298



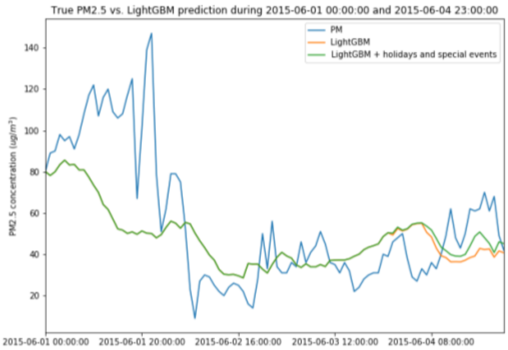


Figure 15. PM2.5 prediction during April 1, 2015 and April 5, 2015

Figure 16. PM2.5 prediction during June 1, 2015 and June 5, 2015

B. What-if analysis

The term "what-if analysis" originally refers to detecting the impact of changing the cell values in a data sheet. In our study, we change the special event attribute to determine to which degree government event calendar assists PM2.5 prediction. In particular, we decide to compare the prediction results during two events: The IAAF World Championships BEIJING 2015 during Aug 22, 2015 and Aug 30, 2015 and Military parade celebrating the 70th anniversary of victory over fascism from Aug 27, 2015 to Sept 3, 2015. As specified in Table VI, we let the aforementioned experimental model predict PM2.5 using an altered testing set that sets the special event flag to zero so that we can monitor how sensitive the model is to the nuance in government's event calendar.

Table VI. CONFIGURATION FOR WHAT-IF ANALYSIS

Model	Testing set setup
LightGBM + holidays and special events (normal)	Unchanged
LightGBM + holidays and special events (what-if)	All "is_special_event" are set to 0 (i.e. no special event occuring)

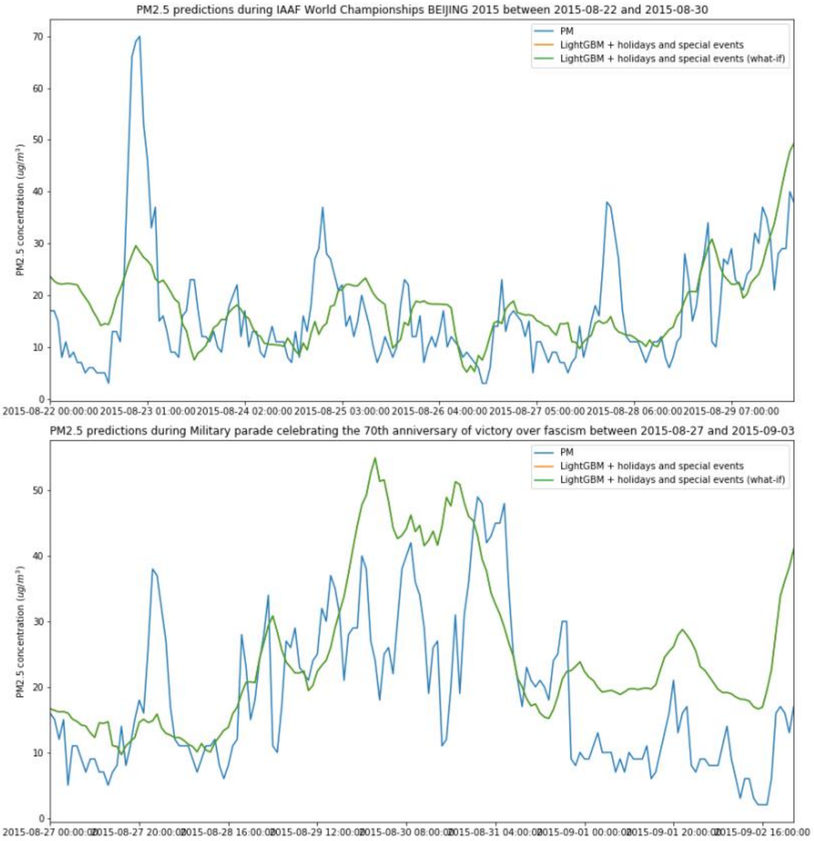


Figure 17. PM2.5 predictions during special events

According to Figure 17, the predicted curve of the control group and that of the experimental group completely overlap each other. Furthermore, as suggested in Table VII, although the scores of the normal group are slightly lower than those of the what-if group during each special event interval, these differences are still not sufficient for us to deduce the impact of government's policies on PM2.5. Therefore, the holiday and special event features did little in changing the model's performance. Further analysis in the next section offers a more detailed explanation.

Table VII. MODEL COMPARISON DURING SPECIAL EVENTS

Model	MAE	RMSE	SMAPE
IAAF World Championships BEIJING 2015			
LightGBM + holidays and special events (normal)	47.6043	55.3941	0.6930
LightGBM + holidays and special events (what-if)	47.6710	55.4920	0.6940
Military parade celebrating the 70th anniversary of victory over fascism			
LightGBM + holidays and special events (normal)	81.5279	103.6433	0.8684
LightGBM + holidays and special events (what-if)	81.5623	103.6565	0.8688

C. Feature importance

One benefit brought by LightGBM other than efficiency and accuracy is that it allows users to study how much a specific feature of the data set contributes to the model prediction, and this functionality allows us to explain why there were only tiny differences in accuracy between our control group and experimental group.

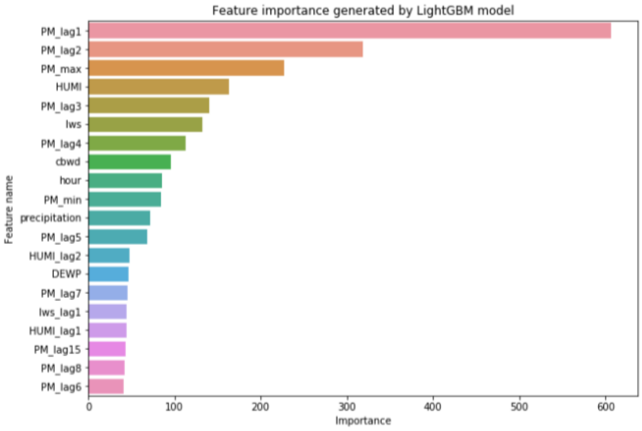


Figure 18. Feature importance plot

As shown in Figure 18, the key features that guide our LightGBM model to predict PM2.5 concentrations are lagged features and statistical features. This implies that nudges in holiday and special event features causes only little change in the prediction, which also means that lagged features and statistical features already compile sufficient information used for model prediction. Accordingly, using lagged features and statistical features alone can be already satisfactory to create accurate PM2.5 prediction models.

VI. CONCLUSION

In our study, we propose a LightGBM model to produce PM2.5 prediction by processing high-dimensional data set. Specifically, we remove redundant attributes, filter out outliers, and impute missing values on this original data set created by Liang *et al.* [8]. Subsequently, we perform feature engineering on the preprocessed data set. For instance, we generate holiday and special event features using external packages and data sets to study how government policies impact PM2.5. In addition, we generate lagged features and statistical features using the sliding window principle. Furthermore, we generate frequency-domain features of the PM2.5 in order for the model to study the spectral characteristics of PM2.5 density values.

Following the feature integration, we conduct experiments on the model to testify the usefulness of holiday and special event features. Because our model does not use weather forecasting data, we adopt the two-stage rolling forecasting strategy to predict meteorological features and PM2.5 values. Additionally, for each 24 hours, we retrain our models with newly observed features to prevent the model's prediction from deviating too much from the true PM2.5 curve.

After experimentation, we perform comparison among previous models and those we propose. An overview in evaluation metrics reveals that rolling forecast brings more substantial improvement to the prediction accuracy of the model than using weather forecasting data. A comparison between the control group and the experimental group in the overall testing set and special event intervals shows that holiday and special event features improves the model accuracy. Nevertheless, altering special event features in the testing set did not cause changes in predictions. In fact, this phenomenon is answered by the feature importance analysis. As LightGBM model's functionality suggests, what contribute to the most to the PM2.5 prediction are lagged features and statistical features, meaning that changes in holiday and special event features will not modify the overall trend of the model prediction.

Although holiday and special event features in our study are unable to allow the prediction model to study the impact of government's policies on PM2.5 variation, the rolling forecasting technique put forward in this research creates a substantial improvement on the model accuracy, thus can be useful for future research in the field of PM2.5 prediction.

VII. REFERENCES

[1] Wang, Y., Wild, O., Chen, H., Gao, M., Wu, Q., Qi, Y., … Wang, Z. (2020). Acute and chronic health impacts of PM2.5 in China and the influence of interannual meteorological variability. Atmospheric Environment, 229, 117397.

[2] Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., ⋯ Huang, L. (2019). A Predictive Data Feature Exploration-Based Air Quality Prediction Approach. IEEE Access, 7, 30732‑30743.

[3] Zhang, C., & Yuan, D. (2015). Fast Fine-Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark. In 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom) (pp. 929–934).

[4] Lee, J., Hong, Y., Lee, Y., Kim, H. S., Song, C. H., Kim, D. Y., & Jeon, M. (2019). Empirical Analysis of Tree-Based Models for PM 2.5 Concentration Prediction. In 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1–7).

[5] Wang, J., & Song, G. (2018). A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. Neurocomputing, 314, 198–206.

[6] Ong, B. T., Sugiura, K., & Zettsu, K. (2016). Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. Neural Computing and Applications, 27(6), 1553–1566.

[7] Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period PM2.5 prediction based on multivariate linear regression model. PLOS ONE, 13(7).

[8] Liang, X., Li, S., Zhang, S., Huang, H., & Chen, S. X. (2016). PM2.5 Data Reliability, Consistency and Air Quality Assessment in Five Chinese Cities†. Journal of Geophysical Research, 121(17), 10220–10236.

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3149–3157).

[10]Lawrence, M. G. (2005). The relationship between relative humidity and the dewpoint temperature in moist air - A simple conversion and applications. Bulletin of the American Meteorological Society, 86(2), 225–233.

[11]Xie, Y., Dai, H., Dong, H., Hanaoka, T., & Masui, T. (2016). Economic Impacts from PM2.5 Pollution-Related Health Effects in China: A Provincial-Level Analysis. Environmental Science & Technology, 50(9), 4836–4843.

[12]Zhang, Q., & Di, G. (2020). 中国清洁空气行动对 PM 2.5 污染的影响. SCIENTIA SINICA Terrae, 50(4), 439–440.

[13]Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting Fine-Grained Air Quality Based on Big Data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2267–2276).

[14]Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. Science of The Total Environment, 664, 1–10.

[15]Zhang, Q., Wu, S., Wang, X., Sun, B., & Liu, H. (2020). A PM2.5 concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations. Journal of Cleaner Production, 275, 122722.

[16]Swalin, A. (2018, March 19). How to Handle Missing Data. Retrieved September 11, 2020, from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

[17]Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing, 307, 72–77.

[18]Christ, M., Kempa-Liehr, A. W., & Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. ArXiv Preprint ArXiv:1610.07717.

VIII. Acknowledgment

Thanks Dr. Juliana Xu for coaching me the techniques of conducting researches. During the quarantine due to COVID-19, she provides me with flexible arrangement to support my research and ensures that I have the best access to Internet during my investigation. In addition, I deeply appreciate my mother for providing me with helpful resources and information, which broaden my view on the environmental perspective of this research, and continuing to encourage me whenever I meet difficulties. Without her, I might not even be able to complete this project. Thanks for all teachers and parents who have cared and aided me during this project!