

Air Pollution in Seoul

Chiu Fan Hui

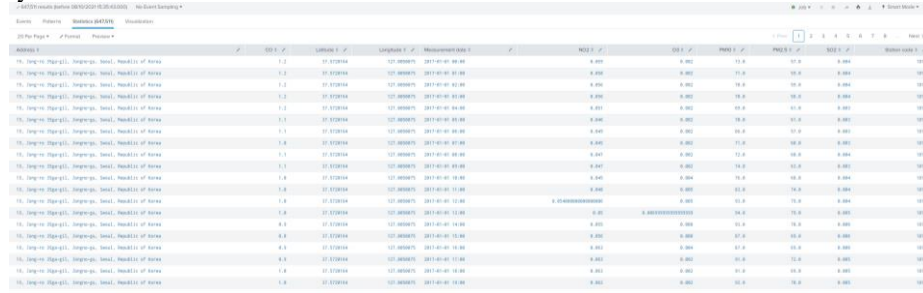
1 Problem Statement and Background

1.1 Problem Statement

Recently, the air quality is getting worse and worse. With many serious air pollutions, people have become more aware of the importance of air protection. Among them, PM2.5 is one of the important factors that cause haze. Only when the content of PM2.5 in the air drops, can the haze situation be alleviated. Therefore, the dataset used in this study is the 2017-2019 air quality data of Seoul, South Korea. The main purpose of this study is to analyze and predict PM2.5 in district 105, so as to determine the changes in PM2.5 in the next 200 days.

1.2 Background

The initial data contains 647511 data including 11 attributes such as address, geographic location and pollutants (Figure 1). The pollutants contained in the Dataset are CO, NO2, O3, PM10, PM2.5 and SO2, and the monitoring interval is one hour. Therefore, it is necessary to perform data clean before analyzing this set of data to remove redundant data, thereby simplifying data and reducing operational complexity.



Address	Station	Date	Time	CO	NO2	O3	PM10	PM2.5	SO2
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	00:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	01:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	02:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	03:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	04:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	05:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	06:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	07:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	08:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	09:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	10:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	11:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	12:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	13:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	14:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	15:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	16:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	17:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	18:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	19:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	20:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	21:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	22:00	0.000	0.000	12.0	10.0	0.000	0.000
105-1, Jangjeon-dong, Seoul, Republic of Korea	105-1	2017-01-01	23:00	0.000	0.000	12.0	10.0	0.000	0.000

Figure 1. Raw data set

2 Methods

2.1 Data Preparation

The following part uses clustering, prediction and forecasting for data analysis and data mining.

Clustering is the method of separating a population or set of data points into multiple groups so that data points in the identical group are more similar than data points in other groups. To put it another way, the aim is to identify groups with related features and assign them to clusters. There are two algorithms used for clustering in this report, which are K-Means and DBSCAN.

The clearly marked groups in the data can be found by the K-means algorithm, which is very effective for identifying unknown groups in complex data and judging the types of groups existing in the data.

DBSCAN is a data clustering approach that uses density-based spatial grouping of applications with noise. DBSCAN combines together points that are close to each other based on a distance measurement (typically Euclidean distance) and a minimal number of points based on a set of points. The spots that are in low-density regions are likewise marked as outliers.

Prediction involves estimating the outcome of unknown data.

Forecasting is a sub-discipline of prediction in which it can predict its future based on time series data. A algorithm called Kalman filter is used for forecasting in this report.

There are two algorithms used for clustering in this report, which are Liner Regression and Random Forest Regressor.

Linear Regression fits a line to the observed data to describe the connection between variables. A straight line is used in linear regression models, whereas a curved line is used in logistic and nonlinear regression models. Regression can be used to estimate how a dependent variable will vary as the independent variable(s) change.

Random forest allows multiple decision trees to be integrated into the final decision, which can be used for classification tasks or regression tasks. Among them, classification tasks include discrete output or predictive classification, while regression tasks include predicting continuous output such as time and price.

2.1.1 Data Cleaning:

The data cleaning operation needs to be completed on the basis of the initial data firstly, so that the amount of system operation can be reduced in the subsequent data analysis and prediction process, and the operation efficiency can be improved. Filtering out invalid data is the first part of data cleaning. This step directly removes data that has no effect on pollutant values such as address and geographic location. Then I expressed all the data of the same pollutant every day in the form of average. The data that have completed data cleaning process (Figure 2).

1000 results 20170101 00:00:00 to 20191231 22:00:00

Run Data Preview 12

Time	Station_Latex	PM2.5	PM10	CO2	NO2	NO3	CO
2017-01-01	100	41.54100000000000	91.27100000000000	0.7610000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-02	100	34.10100000000000	126.11100000000000	0.7410000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-03	100	44.10100000000000	98.10100000000000	0.8010000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-04	100	31.10100000000000	90.00100000000000	0.7610000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-05	100	17.10100000000000	37.10100000000000	0.7010000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-06	100	13.10100000000000	34.10100000000000	0.6910000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-07	100	23.10100000000000	40.10100000000000	0.7410000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-08	100	19.10100000000000	35.10100000000000	0.7110000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-09	100	40.10100000000000	95.10100000000000	0.7810000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000
2017-01-10	100	15.10100000000000	31.10100000000000	0.6810000000000000	0.0010000000000000	0.0010000000000000	0.0010000000000000

1 2 3 4 5 6 7 8 9 10 Next

Figure 2. Data set after filtering

2.1.2 Outliers:

There are 14 outliers detected in the dataset (Figure 3). However, the outliers are not removed. It is because the outliers are legitimate observations and they do not affect the result of the following analysis.

Data and Outliers

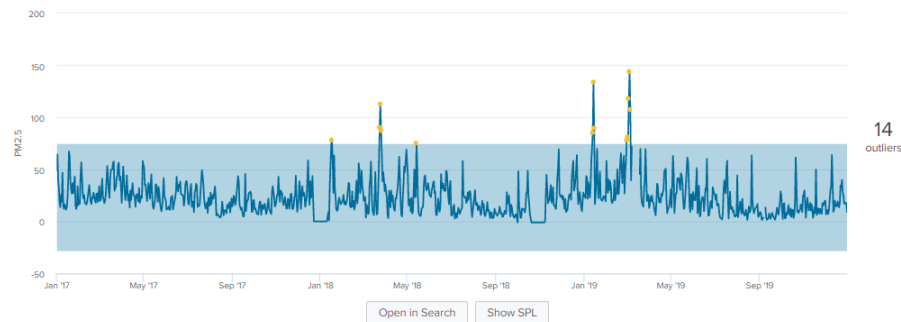


Figure 3. Outliers in dataset

2.1.3 Missing data and null value:

There is no missing data and null value found in the dataset.

2.2 Data analysis and data mining

The following part uses clustering, prediction and forecasting for data analysis and data mining.

2.2.1 Clustering using K-means:

The cluster silhouette_score of clustering using K-Means algorithm is 0.4696 and it is not far from 1 (Figure 4).

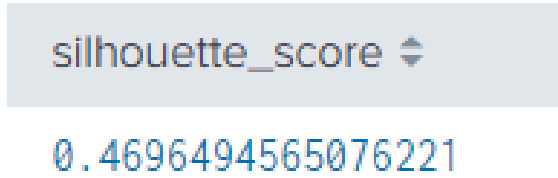


Figure 4. silhouette_score of clustering using K-Means algorithm

2.2.2 Clustering using DBSCAN:

The cluster silhouette_score of clustering using DBSCAN algorithm is 0.859 and it is close to 1 (Figure 5).



Figure 5. silhouette_score of clustering using DBSCAN algorithm

2.2.2 Prediction using Liner Regression:

The predicted line of PM2.5 and the actual line of PM2.5 is very different (Figure 6). The predicted PM2.5 and the actual PM2.5 in scatter chart is also very different (Figure 7). The R^2 of prediction using Liner Regression is 0.4695 and it is far from 1 (Figure 8).

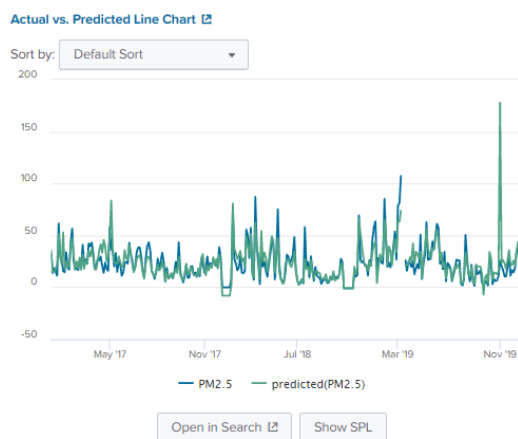


Figure 6. Line chart of prediction using Liner Regression

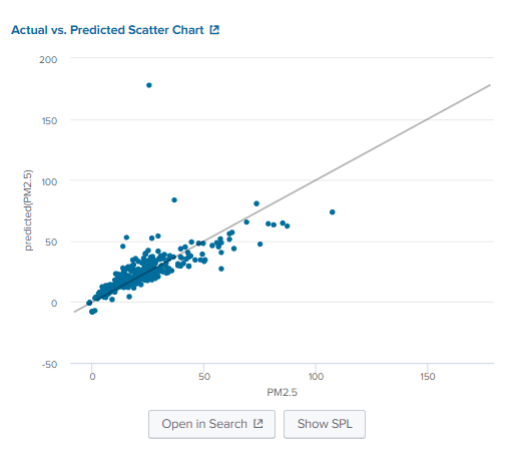


Figure 7. Scatter chart of prediction using Liner Regression

R² Statistic
0.4695

Figure 8. R² of prediction using Liner Regression

2.2.2 Prediction using Random Forest Regressor:

The predicted line of PM2.5 and the actual line of PM2.5 is similar (Figure 9). The predicted PM2.5 and the actual PM2.5 in scatter chart is different (Figure 10). The R² of prediction using Random Forest Regressor is 0.8139 and it is close to 1 (Figure 11).

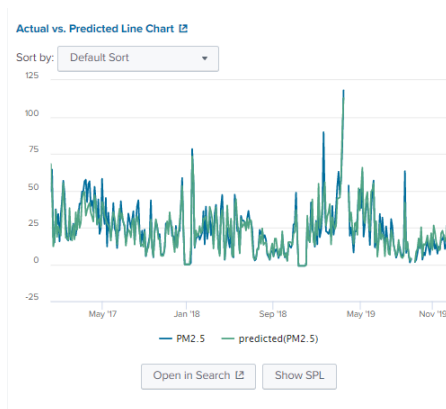


Figure 9. Line chart of prediction using Random Forest Regressor

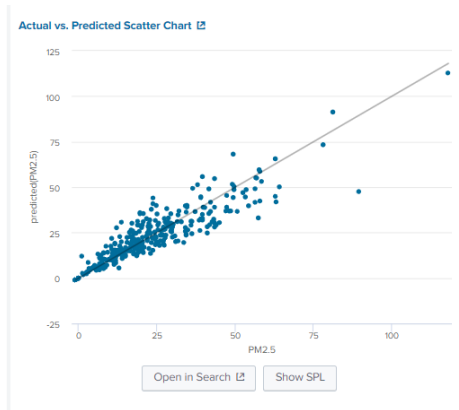


Figure 10. Scatter chart of prediction using Random Forest Regressor

R² Statistic

0.8139

Figure 11. R² of prediction using Random Forest Regressor

2.2.2 Forecasting using Kalman filter:

The R² of forecasting using Kalman filter is 0.4598 and it is far from 1 (Figure 12).

R² Statistic

0.4598

Figure 12. R² of forecasting using Kalman filter

3 Evaluation

For clustering, the silhouette_score of DBSCAN algorithm is 0.3894 higher than the silhouette_score of K-Means algorithm. Better clustering configuration is represented by the silhouette_score that closer to 1. Therefore, DBSCAN algorithm is recommended for clustering.

For prediction, the R² of Random Forest Regressor algorithm is 0.3444 higher than the R² of Liner Regression algorithm. R² that closer to 1 represent better model. Therefore, Random Forest Regressor algorithm is recommended for prediction.

For forecasting, Kalman filter algorithm is not recommended for forecasting. It is because the R^2 of the algorithm is far from 1 which means it is not a suitable model.

4 Tools

Splunk was used in this research, and it is mainly used for big data search and monitoring. The information stored by Splunk can be correlated and indexed by it, and the data can be searched and reports can be generated and data can be visualized.

5 Lessons Learned

In this research experience, I have a deeper understanding and mastery of the data analysis knowledge in the course middle school, which includes the use conditions and methods of different algorithms and how to choose a more suitable algorithm to complete the research. Through continuous practice and testing, I have also increased my understanding of how to use Splunk. In addition, in the process of completing the research this time, my time planning skills and learning ability have also been enhanced.

6 Conclusion and future work

To conclude, DBSCAN algorithm is recommended for clustering and Random Forest Regressor algorithm is recommended for prediction.

7 Weekly member activities

The whole project is completed by myself, I participated in the entire process of data processing and gained a deeper understanding and mastery of each step.