Table of contents

Data (Collection)

Disaster Prevention Weather Data (Seoul)

Source: Meteorological Data Open Portal

(https://data.kma.go.kr/cmmn/main.do.

accessed 10 Dec 2023.)

Data : Disaster Prevention Weather Data

Hourly Measurement data of disaster prevention weather by districts in Seoul

Period : 2016.1.1 ~ 2020.12.31

Variables : temperature, wind speed, wind direction, precipitation

* Local air pressure, humidity, etc. had high multicollinearity with other variables, so only four variables above were selected

Seoul Air Quality Data

Source: Air Korea

(https://www.airkorea.or.kr/web/sidoQualityCompare?itemCode=10008&pMENU_NO=102.

accessed 10 Dec 2023.)

Data : Air Quality Data

Hourly measurement data of air quality monitored by districts of Seoul

Period : 2016.1.1 ~ 2020.12. 31

Variables : NO_2 , PM10, PM2.5

* NO_2 was selected as a variable that affects the predictors: PM10 and PM2.5

Seoul Greenspace data

Source: Seoul Basic Statistics Public Data Portal

(https://data.seoul.go.kr/dataList/368/S/2/datasetView.do.

accessed 10 Dec 2023.)

Data : Greenspace Data

Green space-related data for each district of Seoul prepared annually

Period : 2016 ~ 2020

Variables: Green area by road, square green area, number of street trees, number of berry trees, number of zelkova trees, number of cherry trees, number of ginkgo trees

* Four selected tree types were selected in order of the largest number of trees in Seoul's streets

Data (Pre-processing)

Change data unit

- Converted hourly weather data to daily data
- Daily average : Temperature, wind speed, wind direction, precipitation, NO2
- Daily Maximum : PM10, PM2.5

Data matching & Integration

- Since Air Korea and disaster prevention weather stations are not the same, data is integrated based on the 'district'
- The corresponding disaster prevention weather stations were grouped based on the district of 25 monitoring stations in Air Korea

Weather data and Greenspace data integration

- Integrate combined weather data and green space data
- Green area related data were grouped based on the district column of the integrated weather data

Area processing of greenspace data

From 2016 to 2020:

- Calculate the ratio by dividing the green area along the sidewalk and the square green area by the area of the district
- Calculate the density of street trees by dividing the number of street trees by the green area along the road

Adding COVID-19 related variables

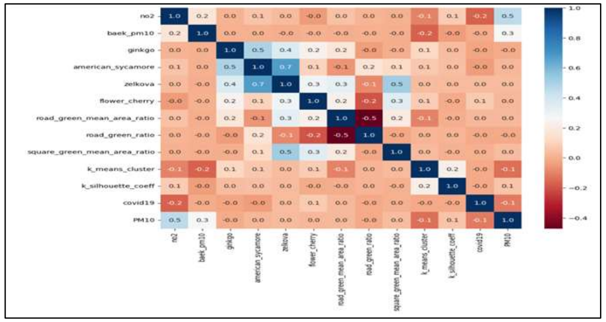
- Based on the Ministry of Health and Welfare, since January 1st, 2020, the date and time column is 1 and 0 for the rest of the date and time
- Among the years included in the analysis, 2020 was different from other periods due to COVID-19, so variables are selected to reflect this in the forecast

Data (Description)

Dataset

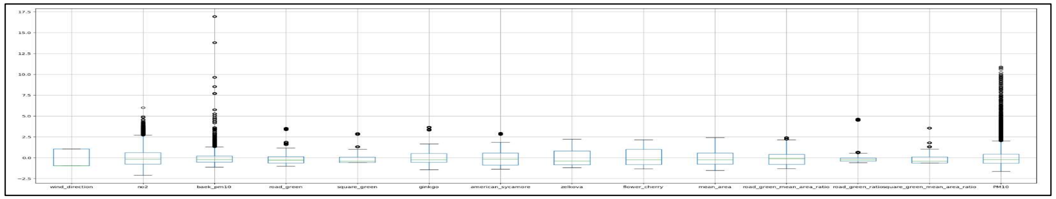
Variable	Unit	Source	Description
temperature	℃	Meteorological Data Open Portal	Converted regional hourly data to daily average data
wind_direction	°		0: 0° – 180° (easterly wind) 1: 180° - 360° (westerly wind)
wind_speed	m/s		Converted regional hourly data to daily average data
rainfall	0 or 1		0: Rainy X 1: Rainy
no2	μg/m^3	Air Korea	Converted regional hourly data to daily average data
PM10	μg/m^3		
PM2.5	μg/m^3		
zelkova	trees / day	Seoul Basic Statistics Public Data Portal	
ginkgo	trees / day		
flower_cherry	trees / day		
american_sycamore	trees / day		
road_green_mean_area_ratio	㎡ / ㎡		Converted regional annual data to daily average data
road_green_ratio	(trees / day) / ㎡		
square_green_mean_area_ratio	㎡ / ㎡		
covid19	0 or 1	Ministry of Health and Welfare	0: Before the COVID-19 outbreak (~2019.12) 1: After the COVID-19 outbreak (20.01~)

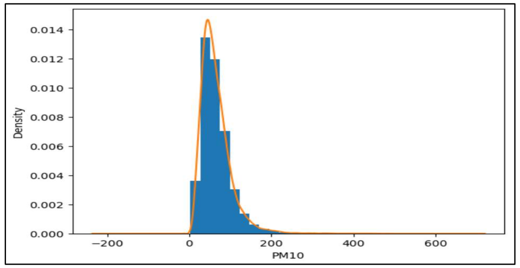
Check correlation coefficient & multicollinearity

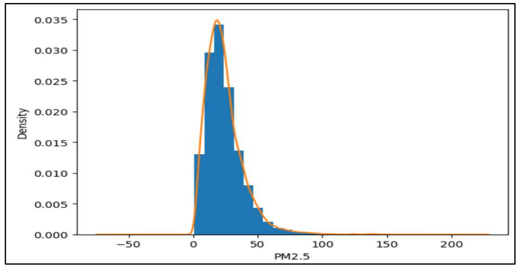


	Feature	VIFscore
0	no2	6.478875
1	baek_pm10	2.646961
2	ginkgo	6.543249
3	american_sycamore	9.483502
4	zelkova	10.977438
5	flower_cherry	4.279494
6	road_green_mean_area_ratio	4.461421
7	road_green_ratio	1.725992
8	square_green_mean_area_ratio	2.463852
9	covid19	1.266677
10	PM10	4.151445

Check Box Plot and Distribution







Train/Test Dataset classification

```
X_train, X_test, y_train, y_test =train_test_split(X, Y, test_size=0.2, random_state=2021, stratify=Y)
```

Train/Test 8:2 separation, apply ‘stratify = y’ to set y to be more balanced

Methodology

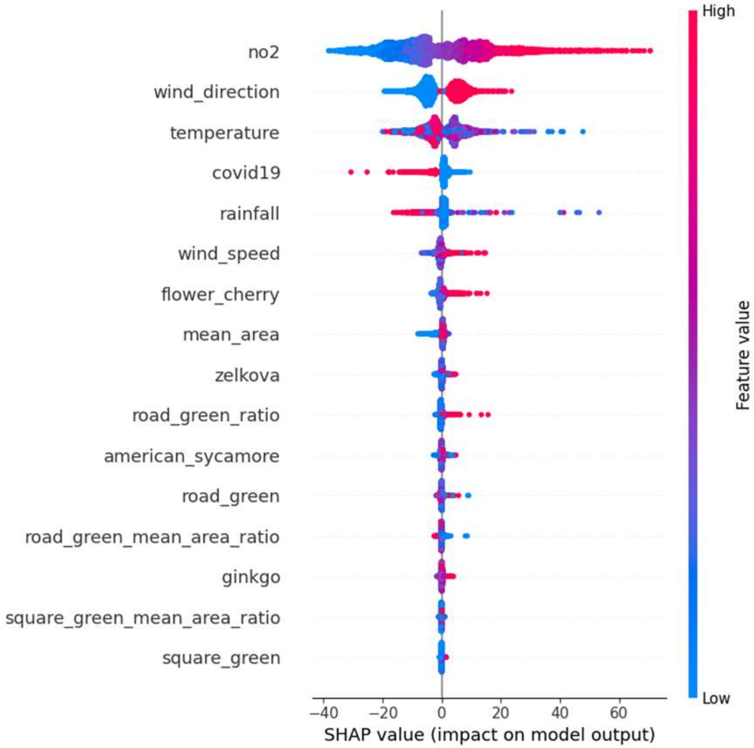
Modeling – Regression Model

Linear Regression	RandomForest Regressor	XGBoost Regressor	LightGBM Regressor
	• Ensemble • feature importance	• Ensemble • feature importance	• Ensemble • feature importance
• p-value	• SHAP value (Shapely Additive Explanations)	• SHAP value	• SHAP value
Assumes a straight-line relationship between input and output, finds the best-fit line to make predictions.	Builds multiple decision trees and averages their predictions for better accuracy and resistance of overfitting.	Combines weak predictive models (usually trees) to create a strong model, optimizing for both speed and accuracy.	Similarly with XGBoost, it uses gradient boosting with a focus on efficiency and speed, especially for large datasets.

Empirical analysis / Results (RF)

Results Interpretation 1 – RandomForest Regressor

SHAP Value Check



SHAP Value Interpretation

* It is shown that the more red dot concentrated toward +(positive) and blue dot toward –(negative), the more the variable affects to y value.

* The higher the index, the higher fine dust had a positive correlation.

1) Representative variables which shows positive correlation

NO2 / wind_direction / wind_speed / flower_cherry

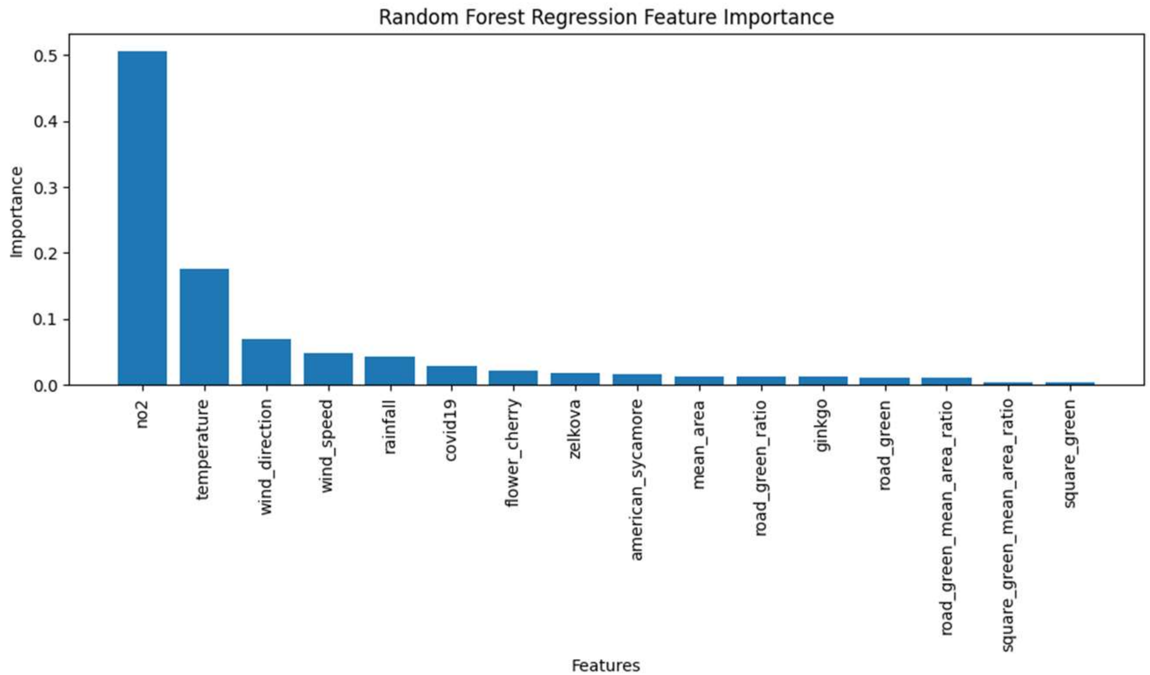
2) Representative variables which shows negative correlation

temperature / covid19 / rainfall

Empirical analysis / Results (RF)

Results Interpretation 1 – RandomForest Regressor

Importance of features



Important Features

NO2 > temperature > wind_direction

(0.1761) (0.0692) (0.04778)

*** Air pollution and atmospheric conditions seems important for prediction.**

Unimportant Features

road_green > square_green
(includes mean area ratio)

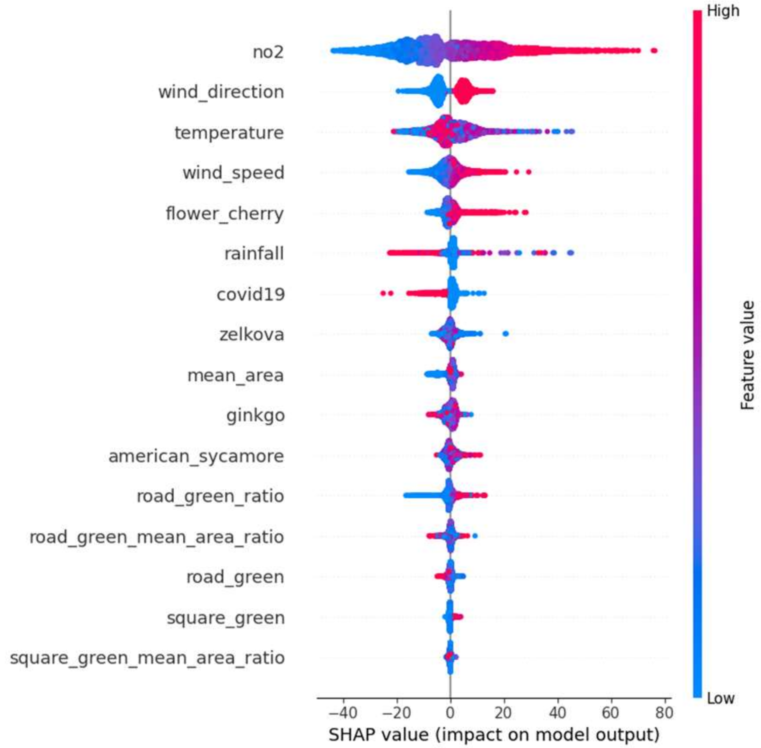
(0.0038) (0.0036)

*** Greenspace data did not made strong relationship between fine dust in this model.**

Empirical analysis / Results (XGB)

Results Interpretation 2 – XGB Regressor

SHAP Value Check



SHAP Value Interpretation

* It is shown that the more red dot concentrated toward +(positive) and blue dot toward –(negative), the more the variable affects to y value.

* The higher the index, the higher fine dust had a positive correlation.

1) Representative variables which shows positive correlation

NO2 / wind_direction / wind_speed / flower_cherry

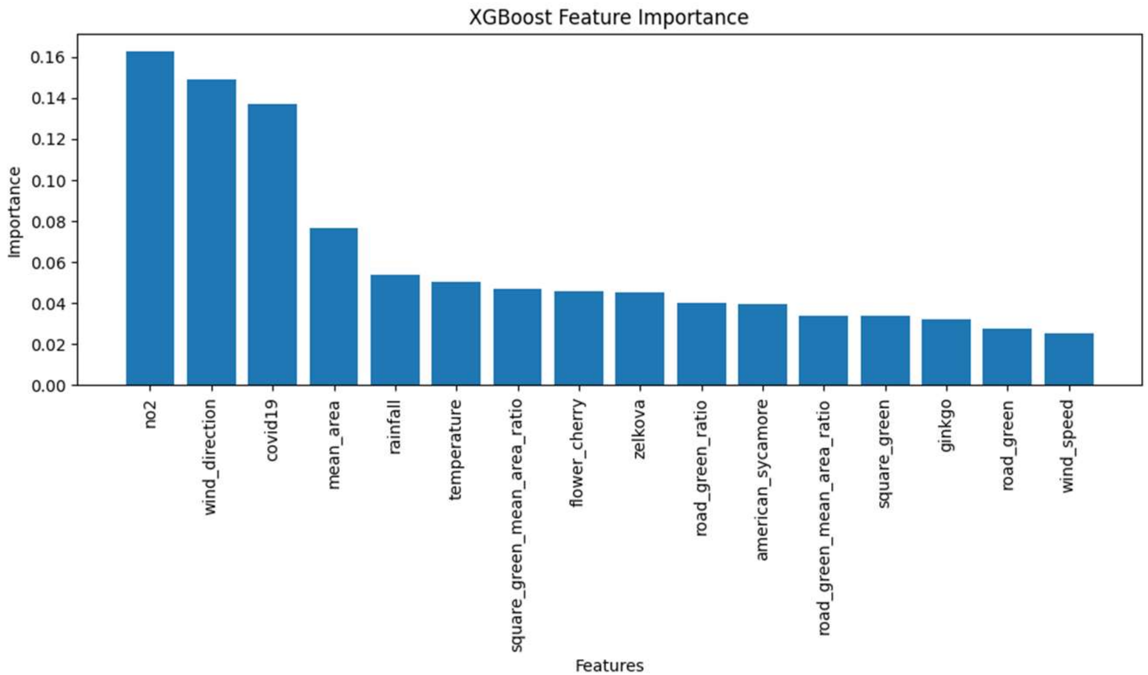
2) Representative variables which shows negative correlation

temperature / rainfall / covid19

Empirical analysis / Results (XGB)

Results Interpretation 2 – XGB Regressor

Importance of features



Important Features

NO2 > wind_direction > covid19

(0.1488) (0.1372) (0.0766)

*** Quite similar with RandomForest, newly revealed that the COVID-19 variable was also highly relevant with fine dust.**

Unimportant Features

road_green > wind_speed

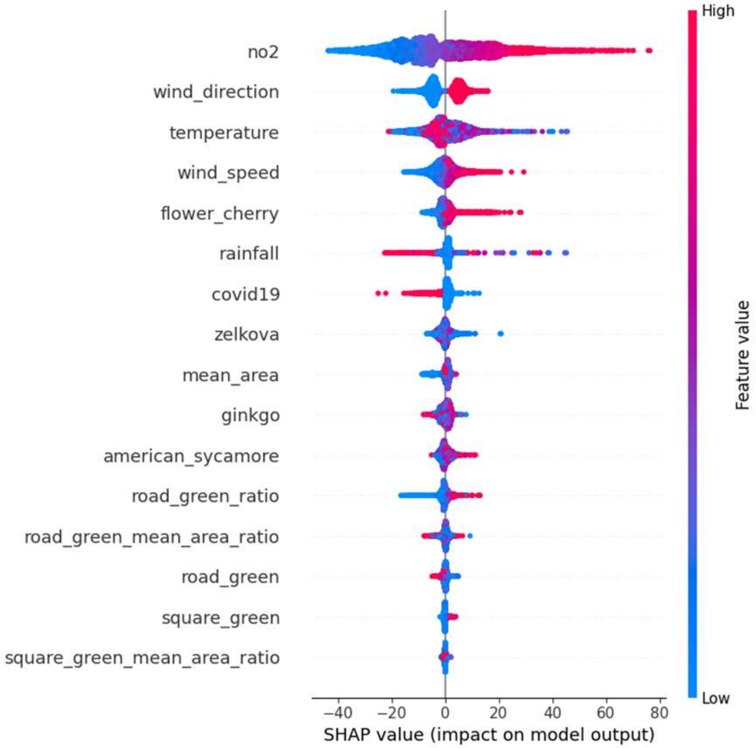
(0.0278) (0.0250)

*** Turns out that speed of wind is not significantly related unlike the wind direction (shown in RF model)**

Empirical analysis / Results (LGBM)

Results Interpretation 3 – LGBM Regressor

SHAP Value Check



SHAP Value Interpretation

* It is shown that the more red dot concentrated toward +(positive) and blue dot toward –(negative), the more the variable affects to y value.

* The higher the index, the higher fine dust had a positive correlation.

1) Representative variables showing a positive correlation

NO2 / wind_direction / wind_speed / flower_cherry

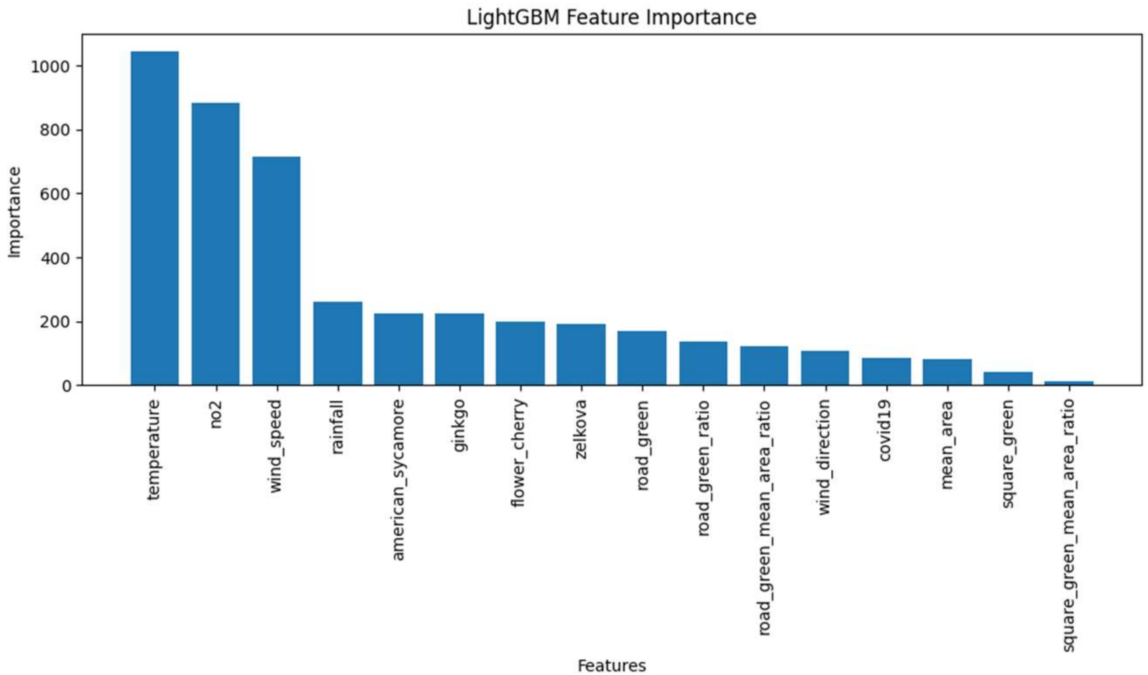
2) Representative variables showing a negative correlation

temperature / covid19 / rainfall

Empirical analysis / Results (LGBM)

Results Interpretation 3 – LGBM Regressor

Importance of features



Important Features

temperature > NO_2 > wind_speed

(884) (714) (263)

*** Wind speed has risen again as an important variable, which seems to require further consideration.**

***In the SHAP analysis, wind_speed always showed a high correlation: considerable as important one**

Unimportant Features

Square_green > Square_green_mean_area_ratio

(42) (13)

*** Same result with RF model, in which the greenspace variables were not important as weather variables**

Conclusion

Model	MSE	Important Variables
RandomForest	994.927	NO2 / Temperature / Wind direction
XGBoost	923.103	NO2 / Wind direction / COVID-19
LightGBM	921.886	Temperature / NO2 / Wind speed

- **By integrating 3 types of data, we were able to analyze the variables that affect find dust.**
- **1) NO_2 is the most influential variable that shows a positive correlation with fine dust. The remaining positive correlation variables are wind_direction, wind_speed.**
- **2) Temperature is the most influential variable that shows negative correlation with fine dust. The remaining negative correlation variable is COVID-19.**
- **3) Weather-related variables mainly affected the results, but some regional green-related variables such as flower_cherry also had a significant positive correlation.**
- **If we conduct analysis with more data sources, it would help to improve climate predictions.**