Derivation of main influencing factors related to high-concentration ozone generation in Seoul using classification models

Lee Jin Hyo · Kwon HS · Son JS · Mok KR · Sa CH (Seoul Metropolitan Government, Air Quality Policy Division)

Introduction

★ Backgrounds

- Unlike ozone located in the stratosphere, during atmospheric high-concentration ozone exposure, it can induce **chronic respiratory diseases**, **coughing**, **reduced lung function**, and more, highlighting the need for efficient management.
- Monitoring the recent **upward trend in ozone levels**, often accompanied by heatwave, and tracking changes in ozone precursors and byproducts alone may not suffice for proactive measures to mitigate ozone pollution.
- While public awareness regarding ozone pollution is increasing, **understanding the comprehensive dynamics of ozone is challenging** due to seasonal influences and intricate chemical reactions involved in its generation process. ***(even experts)***

★ Aim

- In this study, we utilize atmospheric big data, including existing urban air monitoring networks, meteorological data, **to derive the key factors that significantly influence the issuance of ozone alerts**. To present these findings in an easily understandable format, we apply **machine learning-based classification models**.

Material and Method

✓ Target & Period : Gangseo borough / 2022 year

✓ Item : urban air monitoring networks, meteorological data

* 23 items including O_3, PM2.5, PM10, NO_X, CO, SO_2, VOC_S, wind speed, humidity

✓ Algorithm : **Decision Tree**, **Random Forest**, **XGBoost**

* Data preprocessing : **Kalman Filter**, **SMOTE**

(Synthetic Minority Over-sampling Technique)



< Research Area>

Results & Discussion

★ Data preprocessing

- Before / After applying SMOTE, the categorization ratios of N(no alert) and Y(alert)

→ **before N : Y = 8,743 : 17 / after N : Y = 8,743 : 8,738**

★ Decision Tree

- **Decision Tree - Confusion Matrix**

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 1706   10
         Y   24 1757

               Accuracy : 0.9903
                 95% CI : (0.9864, 0.9933)
    No Information Rate : 0.5053
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9806

 Mcnemar's Test P-Value : 0.02578

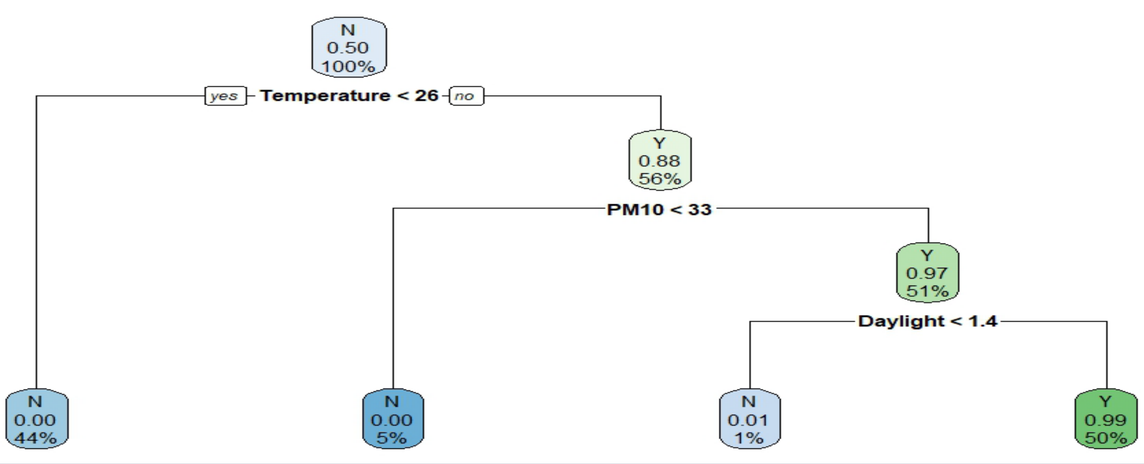
            Sensitivity : 0.9861
            Specificity : 0.9943
         Pos Pred Value : 0.9942
         Neg Pred Value : 0.9865
             Prevalence : 0.4947
         Detection Rate : 0.4878
   Detection Prevalence : 0.4907
      Balanced Accuracy : 0.9902

       'Positive' Class : N
```

★ Accuracy 0.9903, Sensitivity 0.9861, Specificity 0.9943 We have confirmed that the classification of categories 0 (No ozone alert issued) and 1 (Ozone alert issued) has been appropriately performed overall.

★ Generally, decision trees initiate the first split based on the most explanatory variable, and in this regard, we determined that temperature, PM10, and solar radiation are important factors influencing the issuance of ozone alerts.

- **Decision Tree plot**



★ Random Forest

- **Random Forest - Confusion Matrix**

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 1729    1
         Y    1 1766

               Accuracy : 0.9994
                 95% CI : (0.9979, 0.9999)
    No Information Rate : 0.5053
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9989

 Mcnemar's Test P-Value : 1

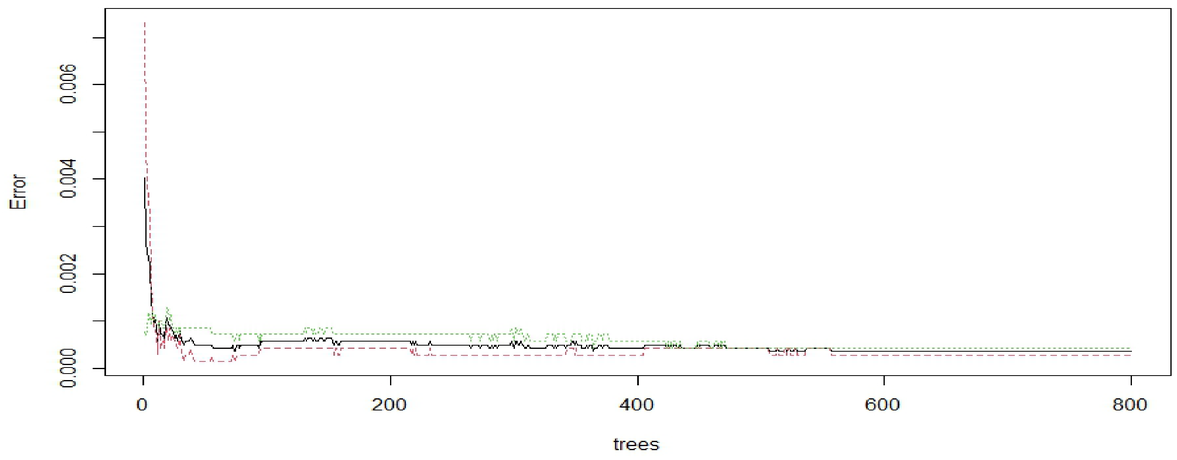
            Sensitivity : 0.9994
            Specificity : 0.9994
         Pos Pred Value : 0.9994
         Neg Pred Value : 0.9994
             Prevalence : 0.4947
         Detection Rate : 0.4944
   Detection Prevalence : 0.4947
      Balanced Accuracy : 0.9994

       'Positive' Class : N
```

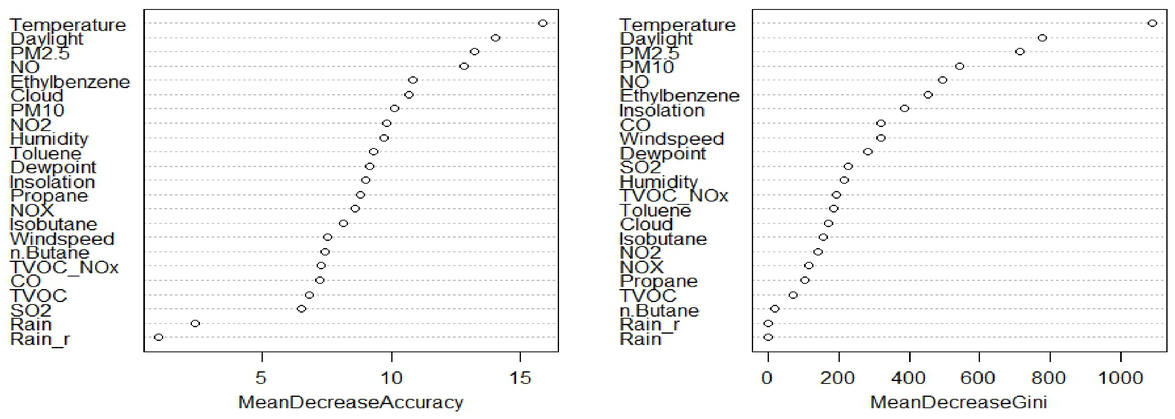
★ Accuracy 0.9994, Sensitivity 0.9994, Specificity 0.9994, Classification was performed more accurately compared to the decision tree model.

★ In terms of measuring variable importance results for MDA (Mean Decrease Accuracy) and MGD (Mean Decrease Gini) revealed that variables such as temperature, solar radiation, and PM2.5 were positioned in the top ranks, indicating their significant impact on the issuance of ozone alerts.

- **Random Forest OOB(out of bags) error rate and plot(influence factor ranking)**



→ **Determine the number of trees (600) to achieve the lowest OOB error rate**



★ XGBoost

- **XGBoost - Confusion Matrix**

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 1753    0
         Y    2 1742

               Accuracy : 0.9994
                 95% CI : (0.9979, 0.9999)
    No Information Rate : 0.5019
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9989

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.9989
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9989
             Prevalence : 0.5019
         Detection Rate : 0.5013
   Detection Prevalence : 0.5013
      Balanced Accuracy : 0.9994

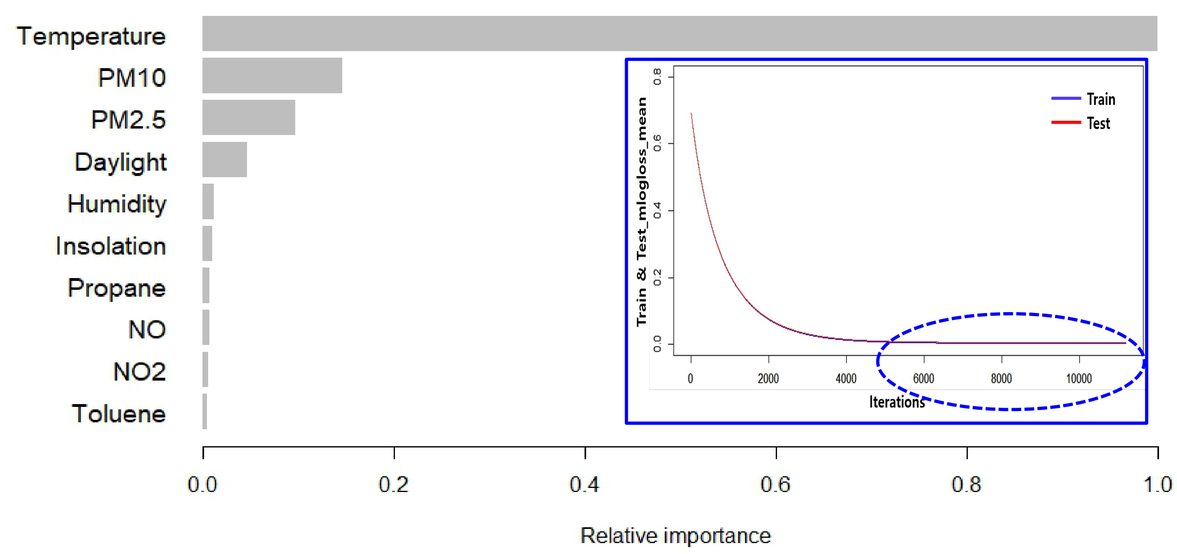
       'Positive' Class : N
```

★ Accuracy 0.9994, Sensitivity 0.9989, Specificity 1.000, Classification was achieved with high accuracy, similar to the Random Forest model, and like both models, temperature, PM2.5, PM10, and solar radiation were identified as crucial factors for the issuance of ozone alerts.

★ Even when the early_stopping_rounds were set to a low value of 10, the model demonstrated robust performance, indicating that unnecessary training was not observed and saving modeling time.

- **XGBoost loss rate and plot(main influencing factors)**

→ We observed that the evaluation metric mlogloss(multiclass log loss) reached a stable value and confirmed the end of training, with **the best iteration at 8,491** out of a total of 15,000 iterations.



Conclusions

- Through various classification model algorithms, we have identified reliable **key factors influencing ozone concentrations**, such as **temperature**, **fine dust**, and **solar radiation**. In the future, when high ozone concentrations occur, we anticipate the possibility of conducting scientific root cause analyses as well as efficient ozone management.
- We plan to integrate various datasets, including personal information and healthcare data, to identify regions vulnerable to high ozone concentrations. We also intend to develop **a system for predicting the issuance of ozone alerts for the following day using factors that can be forecasted in advance**.