Application and Performance Comparison of Tree Model in PM2.5 Concentration Prediction

Pengzhou Xu

School of Mathematical Sciences, Inner Mongolia University, Hohhot, 010000, China
a136194@correo.umm.edu.mx

Abstract. With the acceleration of urbanization and the increase of industrial emissions, air pollutants have posed an increasingly serious threat to human health and environmental safety. This study uses an air pollution dataset collected from Southeast Asian countries, which contains multiple pollutant concentration indicators. Three tree planting models were developed to perform regression prediction on PM2.5 concentration: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). The relationships among variables were explored by cleaning the original data set and combining visualization. In the model results, all three models achieved good predictive capabilities, but the RF performed the best. The findings demonstrate that the tree model performs well when the data scale is medium and the feature correlation is poor. This paper further analyzes the possible reasons for the performance differences of the model and points out the limitations of the current research, such as insufficient feature dimensions and inadequate parameter tuning of the model. Finally, this paper puts forward improvement suggestions, including introducing larger-scale data, adopting deep learning methods and combining with spatial visualization platforms, to enhance the accuracy and practical application value of air pollution prediction.

Keywords: Air Pollution; PM2.5; RF; XGBoost; LightGBM.

1 Introduction

Due to the rapid growth of urbanization and industrialization, the issue of air quality has received increased attention in recent years. Human health is closely related to air quality. Air pollution may pose significant health risks to humans. Some studies have shown that prolonged exposure to polluted air may lead to premature death [1]. PM2.5 makes up the majority of air pollution. When it comes to air pollutants that have particles smaller than or equal to 2.5 micrometers, PM2.5 can directly enter the lungs and circulatory system of the human body [2]. Prolonged exposure to elevated levels of PM2.5 can lead to cardiovascular diseases and increase the mortality rate of cardiovascular diseases [3]. To control the harm caused by PM2.5 to the environment and human

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_32

health, accurately predicting the concentration of PM2.5 is an important research direction combining environmental and data science, thereby controlling environmental pollution and reducing the risk of illness.

The prediction of PM2.5 belongs to the problem of time series. Currently, the most widely used methods for predicting PM2.5 concentration are statistical methods, machine learning and hybrid models [4, 5]. The statistical methods for predicting PM2.5 concentration mainly include principal component regression model (PCR) and multiple linear regression model (MLR) [6, 7]. Some scholars have also established mixed statistical models to improve the prediction accuracy [8]. However, these methods have certain limitations. Whether it is time series or multiple linear regression, they often rely on strict model assumptions and have the problem of a single input parameter. Their effects are limited when dealing with nonlinear relationships and high-dimensional data. As computer processing capability enhances and data properties become increasingly diverse, machine learning algorithms have demonstrated high flexibility and stability, and have great potential in the regression and classification problems of air pollution. In this field, some researchers use Support Vector Machine (SVM) to predict air quality, and another widely used algorithm is Artificial Neural Network (ANN) [9, 10]. According to Leng et al.'s research findings, SVM produces good results when used to forecast the quantities of two heavy metals, Fe and Pp. More than 0.7 is the correlation coefficient (R2). According to Guo et al.'s research, the model outcomes of utilizing ANN to predict PM2.5 concentration include mean absolute error (4.6 μg/m) and R2 (0.9570). The findings are likewise excellent. However, SVN and ANN do not perform well for datasets with missing data, and feature standardization is usually required first. In comparison, the tree model has the advantages of strong robustness, insensitivity to feature distribution, and high interpretability, and is very suitable for the prediction problem of air pollutants.

To sum up, different scholars have adopted different methods and models regarding the prediction of air pollutants. The data used in this paper was collected in Southeast Asian countries. Three different tree models were used to conduct regression prediction on PM2.5 concentration, and indicators such as R^2, MSE and MAE were selected for comprehensive evaluation, thereby putting forward targeted suggestions to help optimize urban air quality management strategies.

2 Method

2.1 Data Sources

All of the data used in this article came from Southeast Asian nations, including Bangladesh, Sri Lanka, India, and Pakistan, and it was gathered from the Kaggle website [11]. Dataset contains 5,000 samples and 10 features and was updated in December 2024. Kaggle datasets may vary in source and quality, this dataset has been widely cited and includes detailed pollutant indicators with consistent data formats, making it suitable for academic research and predictive modeling.

2.2 Data Description

This study's dataset contains variables related to pollutant concentrations, meteorological variables, and social environmental variables. The dataset has no missing values. To better present the data and improve the model performance, logarithmic transformation was carried out on some variables. Except for a few variables, most of the variables tended to be symmetrically distributed after the transformation.

Table 1. List of Variables.

Variable	Logogram
Ambient temperature	x_1
Moisture content in air	x_2
PM2.5	x_3
PM10	x_4
NO2	x_5
SO2	x_6
CO	x_7
Proximity to Industrial Areas	x_8
Population Density	x_9
Air Quality	x_{10}

Table 1 lists the feature names used in the dataset, providing a basis for subsequent modeling.

Fig. 1 illustrates how each independent variable and the target variable are related. Except for x_4 and x_3, which show a significant positive correlation, the linear correlation between the remaining variables and PM2.5 is relatively weak.

Fig. 2 shows the box plots of each feature after logarithmic transformation. There are many right-biased distributions and extreme values in the original variables. After logarithmic transformation, the data tend to be symmetrical, but x_3, x_4, and x6 still retain a large number of outliers, which may indicate that air pollutants significantly increased at a certain moment.

2.3 Model Principle

Using reentry, RF randomly selects the initial data to create several decision trees, and makes predictions through majority voting (classification) or averaging (regression) methods. This learning method can effectively reduce the risk of overfitting, is insensitive to outliers and missing values, and has strong generalization ability and robustness. In addition, RFs can still maintain good performance when dealing with high-dimensional data and can assist in feature selection by evaluating the importance of features.

XGBoost is a model based on the Boosting Tree, which learns the residuals of the previous round's prediction in each round and continuously optimizes the objective function, this can reduce the possibility of overfitting. XGBoost supports parallel processing and missing value handling, and performs particularly well on small and medium-sized datasets.

LightGBM has a significant improvement over XGBoost in terms of running speed and memory efficiency. It adopts a leaf-based growth strategy (Leaf-Wise), giving priority to Leaf nodes that can bring the maximum loss reduction when splitting nodes, thereby improving the model accuracy. Meanwhile, LightGBM uses the histogram algorithm to discretize continuous features, significantly reducing the computational load.

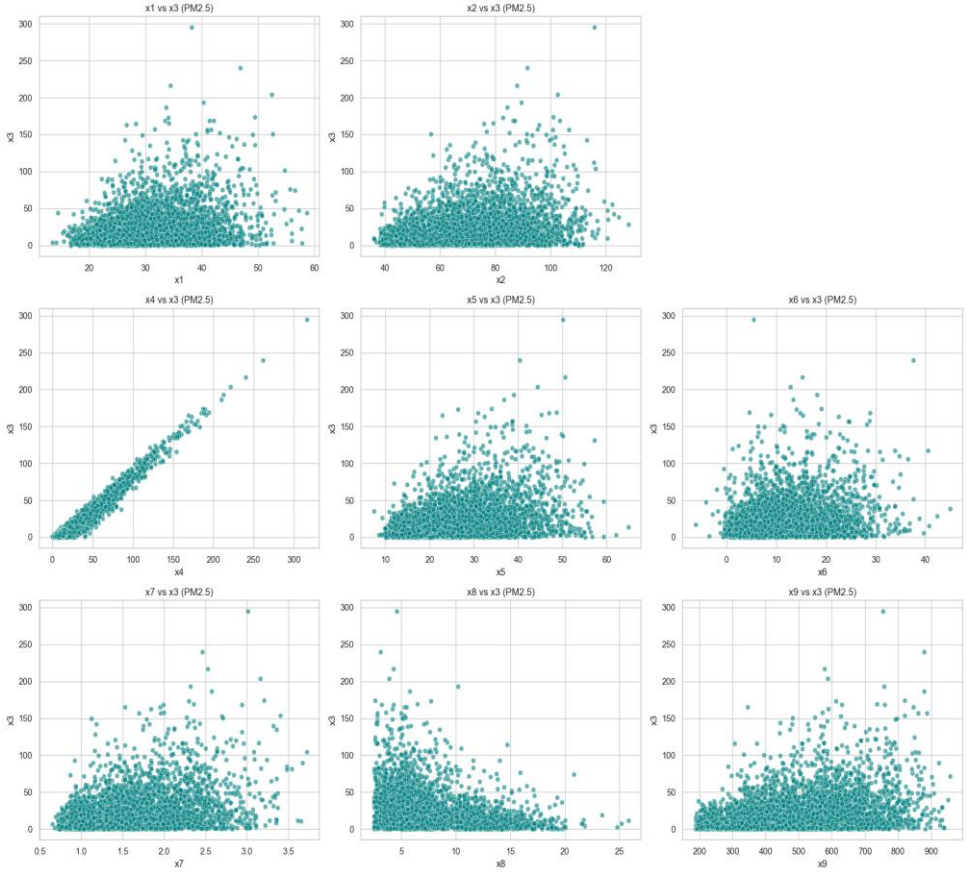

Fig. 1. Scatter Plots (Picture credit: Original).

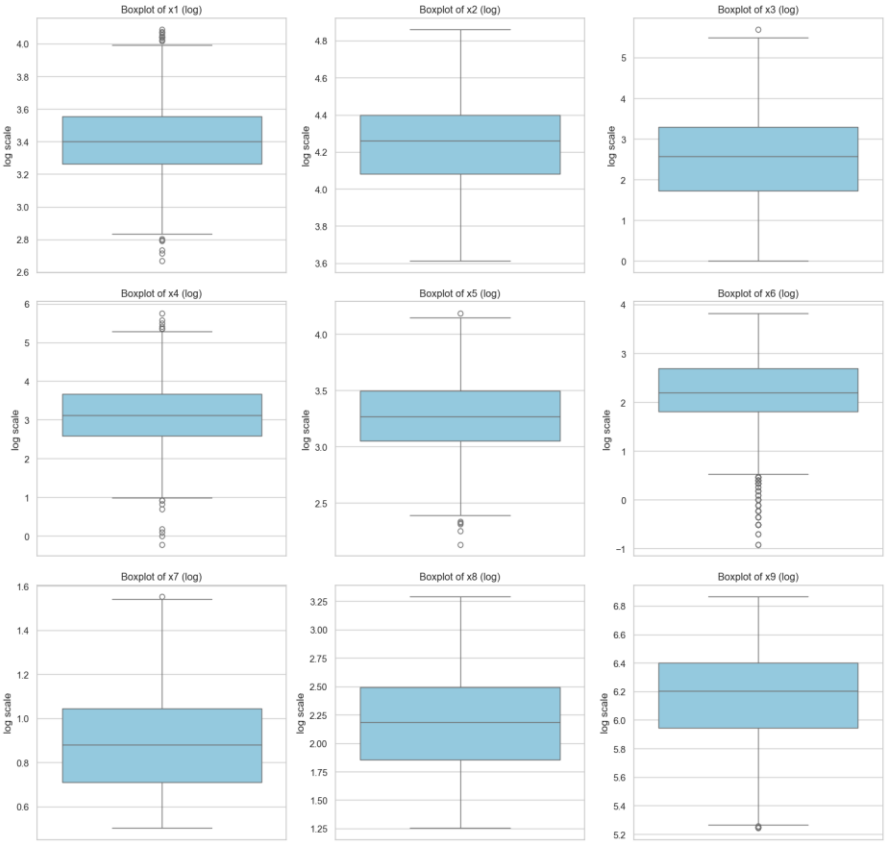

Fig. 2. Boxplots (Picture credit: Original).

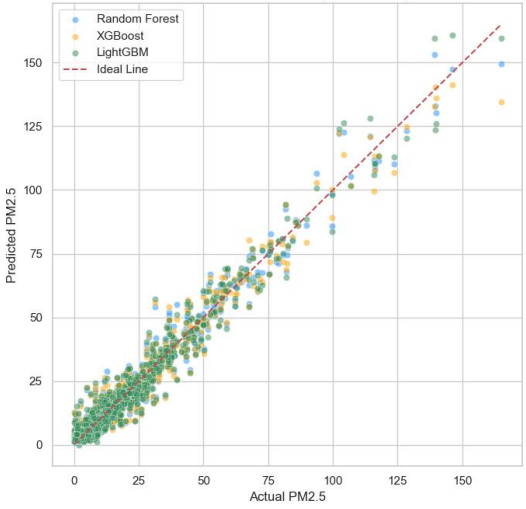

Fig. 3. Model Results (Picture credit: Original)

3 Findings and Analysis

3.1 Model Results

Fig. 3 shows the comparison between actual and predicted PM2.5 values using three tree-based models: RF (blue dots), XGBoost (orange dots), and LightGBM (green dots). The x-axis represents the actual PM2.5 concentration (μg/m³), while the y-axis represents the predicted PM2.5 concentration (μg/m³). The ideal line, when expected and actual values are equal, is shown by the red dashed line. Most data points from the three models are densely clustered around the ideal line, indicating good overall prediction performance. Overall, the predicted value distributions are relatively concentrated and closely distributed around the reference line. In the low concentration section, the RF model outperforms the other two models by a small margin and is comparatively robust. XGBoost performs well in the medium and low concentration range. LigthGBM performs well in the high-concentration range, which may better reproduce the changing trend of extreme points.

3.2 Discussion

The model performance of this study was evaluated by three regression evaluation indicators: MAE, RMSE and R^2.

Table 2. Evaluation indicators

Model	MAE	RMSE	R^2
RF	3.0659	4.4340	0.9611
XGBoost	3.1134	4.5465	0.9591
LightGBM	3.1335	4.5706	0.9587

Table 2 presents the evaluation index results of each model. Judging from the results, the best results were obtained by the RF, while the performances of the other two models were close but slightly lower. The reason for this phenomenon may be related to the dataset itself used. The dataset used in this paper has a moderate amount of data and the correlation among various features is relatively weak. Therefore, the performance of the RF will be better because the Bagging-based ensemble strategy of the RF has stronger robustness and generalization ability on small and medium-sized datasets. Although XGBoost and LightGBM theoretically have higher efficiency and stability, their complex splitting mechanisms may instead reduce the model stability when the feature expressiveness is insufficient, especially when the data has not undergone deep processing. The advantages of their theoretical construction are difficult to fully exert [12]. Overall, the dataset used in this paper has a small sample size and does not construct complex feature relationships. Under this premise, the RF achieves the optimal prediction effect, while the other two tree models are slightly inferior.

Although all three models show good predictive ability, there are also certain limitations. Firstly, the data's sample size is comparatively tiny, which leads to the difficulty in fully exerting the capabilities of the model. Secondly, the model training mainly

adopted basic parameter adjustment and default parameters, and did not use automatic parameter adjustment. This might also have limited the upper limit of the model. Finally, the related work of feature engineering needs to be improved, such as constructing composite variables or introducing time series features, etc., which might improve the model's performance. In the future, research can be conducted using datasets with larger data scales and broader feature dimensions. More complex models such as deep learning and neural networks can also be used to handle time series prediction tasks. Moreover, a visualization platform can be constructed in combination with geographic information systems.

4 Conclusion

This research focuses on the major theme of air pollution. The dataset used contains pollutant concentration variables, meteorological variables, and social environment variables. Three tree models were constructed for training and predicting PM2.5, a pollutant in the air. After analyzing and visualizing the original data set, it was known that there were no missing values in the data set. However, some pollutant concentration variables had right-biased distributions and extreme values. To make the data distribution more balanced, logarithmic transformation was used. After this, regression analysis was conducted using three tree models (RF, XGBoost, LightGBM). The performance of the three models was good, with R2 all exceeding 0.95, and both the MSE and the RMSE were very low, demonstrating good model fitting ability and prediction accuracy. Among them, the RF performs the best, which could be brought on by the original data set's small sample size and straightforward feature relationship. Although XGBoost and LightGBM theoretically have better nonlinear fitting capabilities, this is also consistent with the conclusions of some scholars' research. The performance ability of the model not only depends on the algorithm itself, but is also affected by data characteristics and parameter optimization strategies. However, during the model's operation, some limitations were also found, such as a relatively small sample size, relatively basic parameter tuning, and incomplete feature engineering. These factors may have restricted the further improvement of the model's performance.

In conclusion, this study has established a basis for future research that is more intricate and accurate while also confirming the viability of the tree model in forecasting environmental contaminants. Although there are certain limitations, the good performance ability of the model still has practical significance and development prospects.

References

1. Yang, S., Hao, C., Lam, L.: The role of socioeconomic status in the relationship between environmental pollution and mortality. Journal of Practical Medicine 41(12), 1913–1921 (2025)
2. Zhang, Q., Meng, X.: Overview of particulate air pollution and human health in China: Evidence, challenges, and opportunities. The Innovation 3(6) (2022)

3. Liu, C., Li, Y., Liang, Y.: Research progress on the influence mechanism of outdoor air pollution on cardiovascular disease. Chinese Journal of Experimental Formulas of Chinese Medicine 30(23), 318–326 (2024)
4. Ausati, S., Amanollahi, J.: Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM2.5. Atmospheric Environment 142, 465–474 (2016)
5. Liu, W., Guo, G.: Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm. Atmospheric Pollution Research 10(5), 1482–1491 (2019)
6. Xie, X., Zheng, W., Kai, X., Xu, Y.: Analysis of PM2.5 Concentration in Urumqi Based on Time Series and Multiple Methods. Journal of Yunnan University (Natural Science Edition) 38(4), 595–601 (2016)
7. Zou, X.: Simulation of PM2.5 mass concentration in Ganzhou City based on multiple linear regression model. Innovation of Science and Technology (8), 133–135 (2024)
8. Liu, B., Jin, Y., Li, C.: Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR–SVR–ARMA combined model. Scientific Reports 11(1), 348 (2021)
9. Leng, X., Qian, X.: Leaf magnetic properties as a method for predicting heavy metal concentrations in PM2.5 using support vector machine: A case study in Nanjing, China. Environmental Pollution 242(Part A), 922–930 (2018)
10. He, Z., Guo, Q., Wang, Z., Li, X.: Prediction of Monthly PM2.5 Concentration in Liaocheng in China Employing Artificial Neural Network. Atmosphere 13, 1221 (2022)
11. Kaggle: Air Quality and Pollution Assessment, https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment/data, last accessed 2025/08/02
12. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD, pp. 785–794. ACM, New York (2016)