Model Evaluation: XGBoost Regressor

# Objective

To evaluate whether a non-linear tree-based model (XGBoost) could improve the accuracy of PM2.5 concentration forecasting compared to the baseline linear regression model.

# Model Configuration

- Model Type: XGBoost Regressor (`xgboost.XGBRegressor`)

- Hyperparameters:

* • n\_estimators = 100
* • max\_depth = 3
* • learning\_rate = 0.1
* • random\_state = 42

- Feature Set:

* • Lag features: pm25\_lag\_1, pm25\_lag\_2, pm25\_lag\_3
* • Time-based features: year, month, sin\_month, cos\_month

# Evaluation Metrics

Mean Absolute Error (MAE): 1.0782 µg/m³

Root Mean Squared Error (RMSE): 1.4092 µg/m³

# Test Set Predictions

|  |  |  |
| --- | --- | --- |
| Date | Actual (PM2.5 µg/m³) | Predicted (PM2.5 µg/m³) |
| 2021-12-01 | 7.40 | 7.45 |
| 2022-06-01 | 6.99 | 8.98 |
| 2022-12-01 | 6.92 | 8.90 |
| 2023-06-01 | 9.27 | 8.97 |

# Conclusion

Despite XGBoost's reputation for strong performance in structured data tasks, it did not outperform the linear regression baseline in this case. The likely causes include the small dataset size and limited feature complexity, which may have led to overfitting. While predictions remained close to true values, overall error metrics increased. The results suggest that linear models may suffice for small, clean, and seasonal datasets, and further gains may require either more data, domain-specific features, or alternative model classes.