Baseline Forecast Model: Linear Regression

# Objective

To establish a simple and interpretable baseline model for forecasting PM2.5 concentration levels based on historical data using standard time series features.

# Model Configuration

- Model Type: Linear Regression (`sklearn.linear\_model.LinearRegression`)

- Feature Set:

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

# Data Preparation

PM2.5 measurements were resampled to uniform 6-month intervals. The time series was augmented with lag features and cyclical time encodings. A chronological train-test split was performed (not random):

* • Training set: First 23 records
* • Test set: Last 4 records

# Evaluation Metrics

Mean Absolute Error (MAE): 0.7581 µg/m³

Root Mean Squared Error (RMSE): 1.0888 µg/m³

# Test Set Predictions

|  |  |  |
| --- | --- | --- |
| Date | Actual (PM2.5 µg/m³) | Predicted (PM2.5 µg/m³) |
| 2021-12-01 | 7.40 | 7.16 |
| 2022-06-01 | 6.99 | 7.08 |
| 2022-12-01 | 6.92 | 6.27 |
| 2023-06-01 | 9.27 | 7.20 |

# Conclusion

The linear regression model demonstrated moderate forecasting capability and served as a reliable benchmark. While results were generally accurate, the model exhibited limitations in capturing non-linear dynamics and underperformed during higher concentration periods. To address these shortcomings and potentially improve accuracy, a non-linear tree-based model (XGBoost) was subsequently explored.