

Karachi AQI Prediction Project Report

Submitted By: Waniya Badar

1. Introduction and Objectives:

The project aims to predict the Air Quality Index for Karachi using historical air quality and weather data. The outputs include a training pipeline that prepares features and trains models, a simple model registry for saved artifacts, and a dashboard for exploration. The target is AQI as a continuous value. The approach treats the problem as a time aware regression task where temporal context and careful handling of information flow are essential. A core principle throughout was to avoid information leakage from variables that directly define AQI.

2. Exploratory Analysis:

I began with a comprehensive review of the data to understand its distribution, correlations, and temporal patterns. AQI distributions were examined with histograms and box plots. Time series views showed periods of elevated AQI with recurring patterns by hour and month. Summary statistics were computed for pollutant concentrations and weather features. The analysis included inspection of missing values and the time coverage range.

A correlation matrix confirmed that pollutant concentrations such as pm25 and pm10 are highly correlated with AQI. This is expected because AQI is calculated from those quantities. I also observed strong short term autocorrelation in AQI which supports adding lagged versions of the target and rolling statistics. Category counts by AQI health bracket showed a heavy tail in unhealthy ranges which has implications for model error distribution.

3. Feature Selection Rationale:

Early experiments mistakenly included pollutant concentrations as predictors. This produced unusually high test R squared values that did not hold up under further checks, a clear sign of leakage. Based on this, I excluded pollutant concentrations from the training feature set. Instead, I focused on temporal and derived features that respect causal direction.

The final feature design included calendar features hour, day, month, year, day of week, and a weekend indicator, plus coarse time of day bins. I added lagged versions of AQI at one, three, six, twelve, and twenty four hours. Rolling mean and standard

deviation windows over six, twelve, and twenty four hours were also included. I computed short change rates to capture recent movement. Initial rows without lag coverage were dropped to ensure the model only sees past information. Missing values in continuous variables were handled with forward fill and backward fill where appropriate, then medians for remaining gaps.

4. Models and Training Process:

I trained three classical regression models end to end in the pipeline. Random Forest, Gradient Boosting, and Ridge Regression. Data was split into train and test sets with scaling where appropriate. Evaluation metrics included RMSE, MAE, and R squared on train and test. The best model was selected by test R squared for simplicity and consistency.

Model capacity was chosen to balance generalization and artifact size so that saved models are easy to persist. Tree based models used reduced estimators and depth. Linear Ridge serves as a strong baseline that is robust when the signal is smooth. Feature importance was reviewed after training. For tree models I used feature importances and for linear models absolute coefficients.

Beyond the pipeline, I experimented in the notebook with a multi layer perceptron and an AdaBoost regressor. I used LIME explanations to examine sensitivity to input features. These experiments helped with understanding but did not change the production choice in the pipeline.

5. Evaluation and Selection:

The final selection was ensemble based and chosen by test R squared with checks against MAE and RMSE. Gradient Boosting tended to balance bias and variance well on the available data. Random Forest was competitive but a little noisier with fewer samples. Ridge acted as a reliable baseline and sometimes matched the ensembles when lag features dominated the signal.

This approach captured short term temporal structure and produced stable results when recent AQI values were available. It struggles when lagged predictors are missing or stale, and during rapid regime shifts such as holidays or events. Without additional covariates, a lag heavy design has limited ability to anticipate abrupt changes.

6. Explainability Insights:

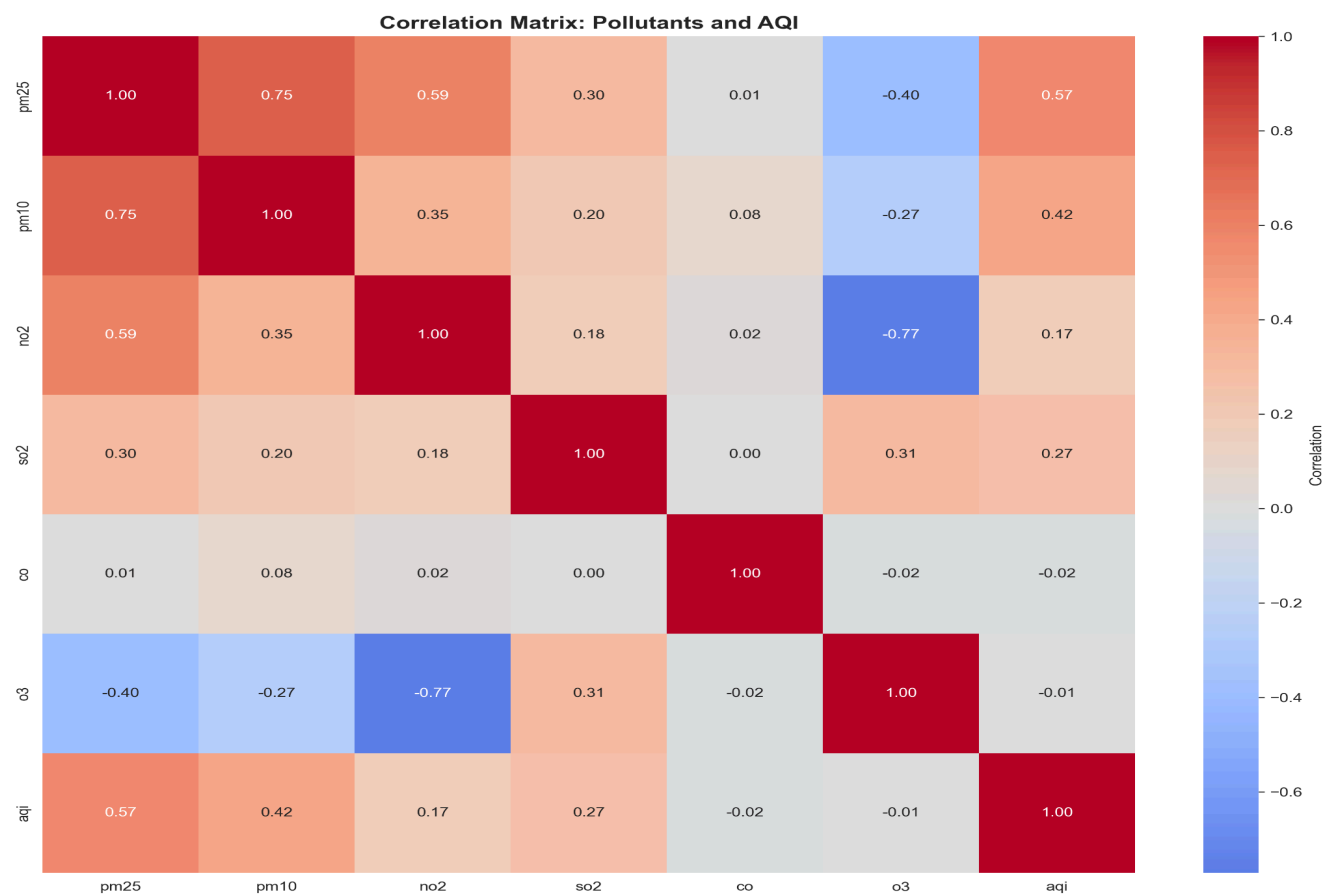
I used both model specific importance and LIME for interpretability. LIME consistently highlighted lagged AQI and rolling statistics as the most influential features, followed by calendar features when coverage was sufficient. This aligns with the feature design and reassured me that training did not silently rely on excluded pollutant columns.

Plots saved from the notebook include an overview of AQI distribution, a correlation heatmap, and a comprehensive trend analysis. These visualizations make clear that AQI has strong autocorrelation and seasonal components.

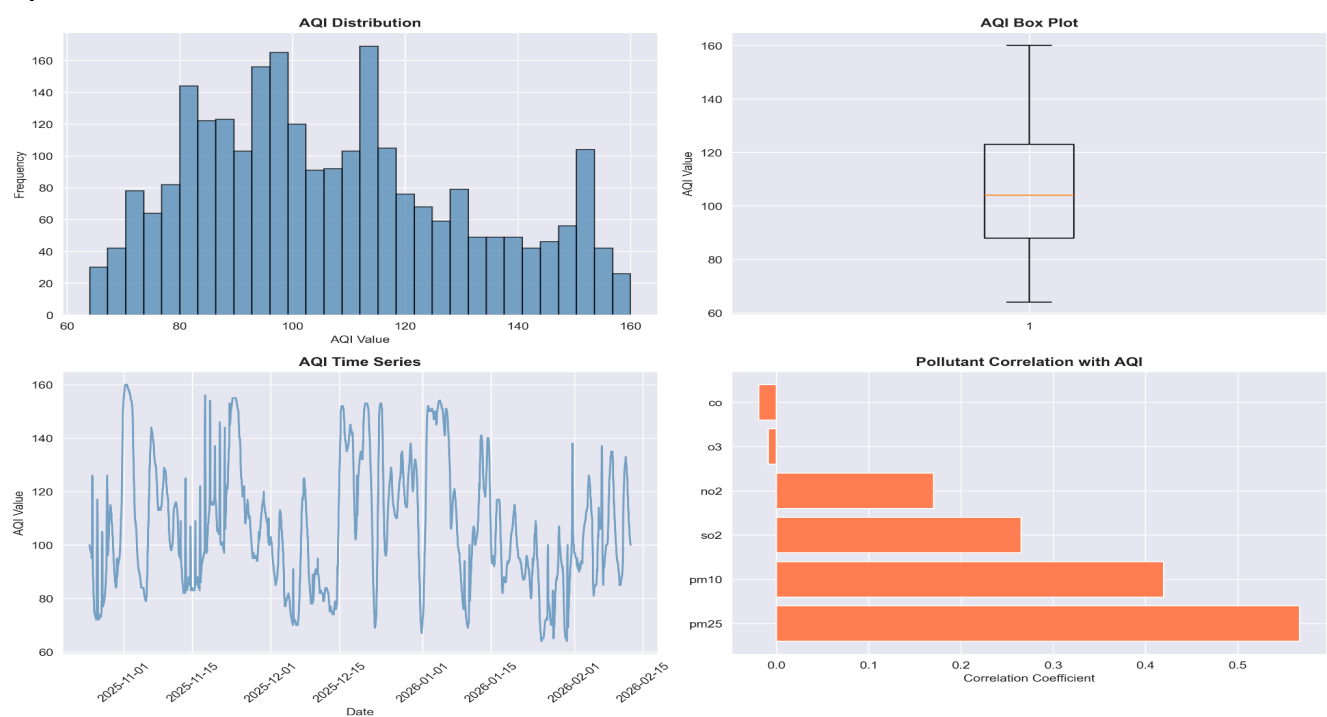
i) Comprehensive Trend Analysis:



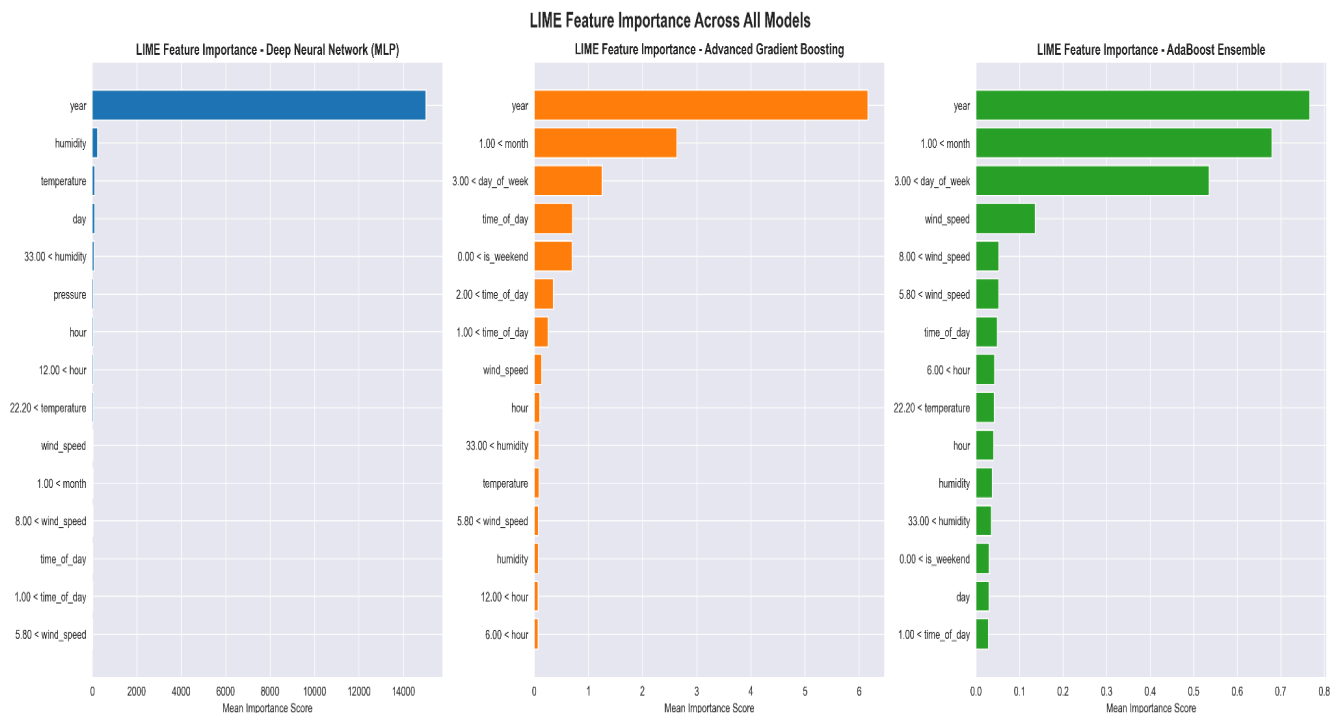
ii) Correlation Matrix:



iii) EDA Overview:



iv) Lime Importance:



7. Limitations and Risks:

The chosen feature set avoids leakage by excluding pollutant concentrations. This prevention comes with a tradeoff because those variables are the most informative instantaneous drivers of AQI. The models lean heavily on recent AQI values. Forecast quality degrades when those values are missing or delayed. Distribution shifts across seasons and events can reduce accuracy unless the pipeline retrains more frequently.

Evaluation uses a single held out split. While sufficient for development, cross validation would provide more robust estimates. Another limitation is the absence of horizon specific targets. Predicting future AQI at fixed horizons would better justify inclusion of lagged pollutant features while still avoiding leakage from current values.

8. Challenges and Lessons Learned:

Leakage was the biggest challenge. Including pollutant concentrations caused inflated metrics and masked real modeling issues. A careful review of correlations and feature definitions resolved this and improved evaluation honesty.

A second challenge was handling missing values and time alignment. Forward and backward fill can smear spikes and flatten variability if applied without care. I adjusted rolling window shifts and dropped initial rows to ensure proper temporal direction.

Small data segments in early periods caused overfitting for larger ensembles. Reducing estimators and depth produced more stable performance and kept saved artifacts compact. I also needed to ensure that feature names were saved and carried through to inference so inputs align with model expectations.

Finally, notebook execution order matters. Several plots depend on computed lists of pollutants and precomputed feature names. I reorganized cells to avoid errors and to ensure charts remain reproducible.