Lahore AQI Monitor & Forecaster

End-to-End Machine Learning System for AQI Prediction

Abdullah Fayyaz
Data Science Intern
10Pearls

February 2026

Contents

1 Introduction

Air pollution in Lahore has become a critical public health crisis, with the city frequently ranking among the most polluted globally. The primary objective of this project was to build a data-driven system capable of monitoring current air quality and predicting future smog events to assist residents in making informed health decisions.

1.1 Project Scope

The project scope encompassed:

- **Data Collection:** Automated ingestion of historical and real-time weather/pollution data.
- **Machine Learning:** Training and evaluating regression models (XGBoost, Neural Networks) to forecast AQI.
- **MLOps:** Implementing experiment tracking, model registry, and version control.
- **Deployment:** Developing a web-based dashboard for visualization and an automated alert system.

2 System Architecture

The system follows a modular architecture integrating cloud services, local processing, and a web frontend.

2.1 Technology Stack

Component	**Technology**
Programming Language	Python 3.9+
Frontend	Streamlit, Plotly
Backend API	FastAPI
Database	MongoDB Atlas (NoSQL)
Machine Learning	Scikit-Learn, XGBoost, SHAP
MLOps Platform	MLflow, DagsHub
External APIs	OpenWeatherMap, Brevo (Email)

Table 1: Technology Stack

3 Data Pipeline & Feature Engineering

A robust data pipeline was established to ensure high-quality input for the models.

3.1 Data Sources

- **OpenWeatherMap API:** Provides real-time and forecasted meteorological data (Temperature, Humidity, Pressure, Wind Speed).

3.2 Feature Engineering

To capture the temporal nature of air quality, several advanced features were engineered:

1. **Lag Features:** `aqi_lag_1`, `aqi_lag_6`, `aqi_lag_24` to capture immediate past trends.
2. **Rolling Statistics:** 24-hour rolling mean of AQI to smooth out volatility.
3. **Cyclical Encoding:** Sine and Cosine transformations for `Hour`, `Month`, and `Wind Direction` to preserve cyclical continuity.

4 Machine Learning Models

Multiple regression models were trained and compared using Mean Absolute Error (MAE) as the primary metric. To ensure fair comparison, data scaling was applied where necessary (e.g., for Ridge and Neural Networks).

4.1 Model Selection

The following algorithms were evaluated to determine the best approach for forecasting AQI:

- **Random Forest Regressor:** An ensemble learning method using 100 decision trees to reduce overfitting (`n_estimators=100`).
- **XGBoost Regressor:** A gradient boosting framework optimized for efficiency and performance on tabular data (`learning_rate=0.1`, `n_estimators=100`).
- **Ridge Regression:** A linear model with L2 regularization to handle multicollinearity between weather features. Implemented as a pipeline with `StandardScaler` (`alpha=1.0`).

- **Neural Network (MLP):** A Multi-Layer Perceptron trained via a pipeline with `StandardScaler`. The architecture consists of two hidden layers with 64 and 32 neurons respectively, using the ReLU activation function (`hidden_layer_sizes=(64,32)`, `max_iter=1000`).

4.2 Best Performing Model

The **NeuralNetwork v12** was selected as the production model, achieving the lowest Mean Absolute Error (MAE) among all candidates.

- **MAE:** 12.81
- **Architecture:** Input Layer → Dense(64, ReLU) → Dense(32, ReLU) → Output(1)
- **Optimization:** Adam Optimizer with Early Stopping (max_iter=1000).

4.3 Model Explainability (SHAP)

SHAP (SHapley Additive exPlanations) values were calculated to interpret model decisions. As shown in the summary plot, the model relies heavily on temporal trends and meteorological cycles:

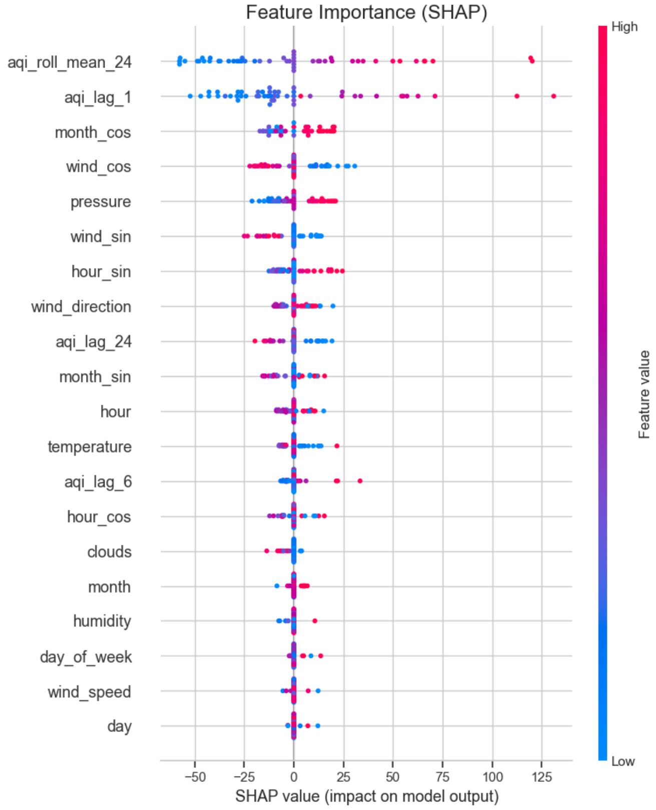

Figure 1: SHAP Feature Importance

- **AQI Rolling Mean (24h):** This is the most dominant feature (`aqi_roll_mean_24`). High values (red points) strongly push the model's prediction upward, indicating that the past 24-hour pollution trend is the best predictor of future smog.
- **Recent History (Lag 1):** The immediate past hour's AQI (`aqi_lag_1`) is the second most important feature, reinforcing the persistence of smog events.
- **Seasonal & Daily Cycles:** Cyclical features like `month_cos` and `hour_sin` play a significant role, confirming that the model has successfully learned the seasonal (winter smog) and diurnal (rush hour) patterns of pollution in Lahore.
- **Meteorological Impact:** Weather features like `pressure` and `wind_direction` (encoded as sin/cos) show a complex but notable impact, with specific wind vectors effectively dispersing or trapping pollutants.

5 Implementation Details

5.1 Real-Time Dashboard

A Streamlit dashboard was developed to visualize data:

- **Live Metrics:** Displays current AQI, Temperature, and Humidity.
- **Gauge Chart:** Visual representation of the AQI category (Good to Hazardous).
- **Forecast Cards:** 3-day prediction cards with dynamic weather icons.
- **Historical Analysis:** Interactive Plotly line charts with smoothing (24h rolling average) and range sliders.

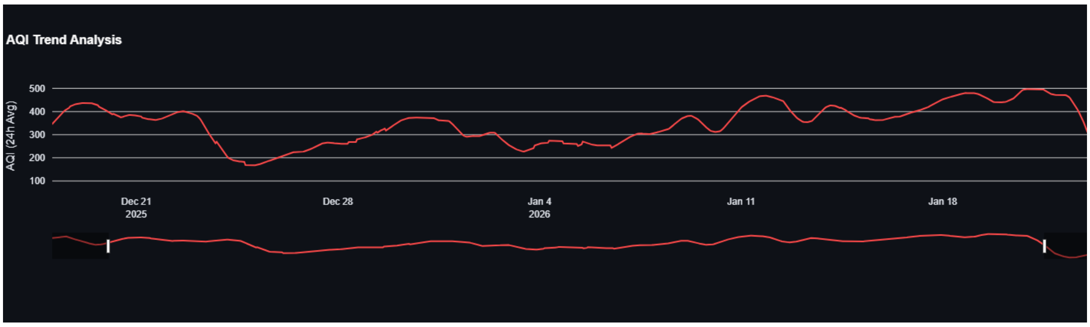

Figure 2: AQI Trend Analysis

5.2 Automated Alert System

A critical safety feature is the automated email alert system powered by the **Brevo API**.

- **Trigger Condition:** AQI predictions 300 (Hazardous).
- **Cooldown Mechanism:** Implemented a 6-hour cooldown using MongoDB logs to prevent spamming users during prolonged smog events.
- **Content:** Emails contain HTML-formatted warnings and actionable health advice (e.g., ”Wear N95 Mask”).

6 Deployment & MLOps

6.1 MLflow & DagsHub

All experiments were tracked using MLflow. DagsHub was used as the remote model registry, allowing for:

- Version control of models.
- Comparison of hyperparameters across different runs.
- One-line loading of the production model via the MLflow Python API.

6.2 FastAPI Integration

The model is served via a FastAPI backend, which exposes a `/predict` endpoint. This decouples the ML logic from the frontend, ensuring scalability.

7 Conclusion & Future Work

The ”Lahore AQI Predictor” successfully demonstrates the application of Machine Learning to solve real-world environmental challenges. The system provides accurate short-term forecasts and actionable alerts.

Future Improvements:

- Integration of satellite imagery data for broader spatial coverage.
- Development of a mobile application using React Native.