Real time 3-Day AQI Forecasting using Machine Learning

Prepared By :

Muhammad Uzair

10Pearls Data Science Internship

1. Project Overview

The Karachi Air Quality Prediction System is a comprehensive end-to-end Machine Learning and MLOps project designed to predict the Air Quality Index (AQI) for Karachi, Pakistan, for the next 3 days. The system automates data collection, feature engineering, model training, and deployment using modern MLOps practices.

This project demonstrates the complete lifecycle of a production-grade ML system, including:

- Automated hourly data collection from OpenWeatherMap API
- Feature engineering with rolling and lag features
- Training and comparison of multiple ML models
- Continuous Integration/Continuous Deployment (CI/CD) pipelines
- Interactive Streamlit dashboard for real-time predictions

2. Data Collection

2.1 Data Source

The project uses the OpenWeatherMap Air Pollution API as the primary data source. This API provides real-time and historical air quality data for any location worldwide. For this project, we focused on Karachi, Pakistan (coordinates: 24.8607°N, 67.0011°E).

2.2 Historical Data Fetching

Initially, 6 months of historical data was collected from August 2025 to February 2026, resulting in 4,200 hourly records. This historical dataset serves as the foundation for training our machine learning models.

The following pollutants and metrics are collected:

Pollutant	Description	Unit
AQI	Air Quality Index	1-5 scale
PM10	Particulate Matter 10	µg/m³
PM2.5	Particulate Matter 2.5	µg/m³
CO	Carbon Monoxide	µg/m³
O_3	Ozone	µg/m³

3. Data Cleaning

Data quality is crucial for accurate predictions. The cleaning process involved identifying and handling missing values, duplicates, and outliers to ensure the dataset is ready for analysis and modeling.

3.1 Missing Value Handling

Missing values were identified in the dataset and handled using the following strategy:

- Forward Fill: For time-series continuity, missing values were first filled with the previous valid observation.
- Backward Fill: Remaining missing values at the beginning of the dataset were filled with the next valid observation.

3.2 Duplicate Removal

Duplicate entries based on timestamp were identified and removed, keeping the most recent record. This ensured that each hour has only one unique data point.

3.3 Outlier Detection

Outliers were detected using range validation for each pollutant:

- AQI: Valid range 0-500
- PM10: Valid range 0-1000 μg/m^3
- PM2.5: Valid range 0-500 μg/m^3
- CO: Valid range 0-50000 μg/m^3
- O_3: Valid range 0-1000 μg/m^3

Values outside these ranges were clipped to the valid range to prevent extreme outliers from affecting model training.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the data distribution, identify patterns, and discover relationships between variables. Key findings include:

4.1 Temporal Patterns

The analysis revealed that AQI values show strong temporal patterns with higher pollution levels during winter months (December-February) and lower levels during monsoon season. Hourly patterns showed peak pollution during morning rush hours (7-9 AM) and evening hours (6-8 PM).

4.2 Correlation Analysis

Correlation analysis identified strong relationships between:

- PM2.5 and PM10 (0.89 correlation)
- AQI and PM2.5 (0.92 correlation)
- Current AQI and next-day AQI (0.85 correlation)

4.3 Distribution Analysis

The distribution of AQI values showed that Karachi experiences unhealthy air quality (AQI 151-200) approximately 45% of the time, with moderate air quality (AQI 51-100) for about 30% of observations. This distribution informed our understanding of the prediction problem and the importance of accurate forecasting for public health.

5. Feature Engineering

Feature engineering is a critical step in building effective machine learning models. From the raw data, we created 45+ engineered features to capture temporal patterns, trends, and historical behavior of air quality.

5.1 Rolling Features

Rolling (moving window) features capture recent trends in air quality:

- aqi_rolling_max_24h: Maximum AQI in the last 24 hours
- aqi_rolling_mean_3h: Average AQI over the last 3 hours
- aqi_rolling_mean_6h: Average AQI over the last 6 hours
- aqi_rolling_mean_12h: Average AQI over the last 12 hours

5.2 Lag Features

Lag features provide historical context by including past values:

- aqi_lag_1h: AQI value from 1 hour ago
- aqi_lag_3h: AQI value from 3 hours ago
- aqi_lag_6h: AQI value from 6 hours ago

5.3 Target Variables

Target variables represent the future AQI values we want to predict:

- target_aqi_1d: AQI 24 hours ahead
- target_aqi_2d: AQI 48 hours ahead
- target_aqi_3d: AQI 72 hours ahead

6. Feature Selection

From the 45+ engineered features, we performed systematic feature selection to identify the most important predictors. This process involved correlation analysis, feature importance from tree-based models, and domain knowledge.

6.1 Selection Criteria

Features were selected based on:

- Correlation with target variables (threshold: >0.45)
- Low multicollinearity with other features
- Feature importance scores from Random Forest

- Domain relevance for air quality prediction

6.2 Final Selected Features (12 Total)

The following 12 features were selected for model training:

Feature	Correlation with Target
aqi_rolling_max_24h	0.986
pm10	0.535
pm25	0.509
aqi	0.505
aqi_rolling_mean_3h	0.498
aqi_lag_1h	0.496
aqi_rolling_mean_6h	0.492
co	0.489
aqi_rolling_mean_12h	0.486
o3	0.484
aqi_lag_3h	0.484
aqi_lag_6h	0.467

6.3 Dropped Features

The following types of features were dropped from the final model:

- Features with low correlation (<0.45) with target variables
- Highly correlated features (>0.95) that caused multicollinearity
- Time-based features (hour, day, month) that showed weak predictive power
- Change-rate features that added noise without improving accuracy

This feature selection process reduced the dimensionality from 45+ features to 12 essential features, improving model training speed while maintaining high prediction accuracy.

7. Hopsworks Feature Store Integration

Hopsworks is a modern MLOps platform that provides a feature store for managing and serving machine learning features at scale. The feature store serves as a centralized repository for our engineered features, enabling versioning, monitoring, and seamless integration with model training pipelines.

7.1 Feature Store Setup

Two feature groups were created in Hopsworks:

- karachi_aqi_raw (v1): Stores raw pollutant data collected hourly from the API. This group contains 6 columns (timestamp + 5 raw pollutants) and grows automatically through the hourly pipeline.
- karachi_aqi_features (v1): Contains the initial 6-month historical dataset with all 12 engineered features and 3 target variables. This was used for the initial model training.

7.2 Data Upload Process

The historical data upload process involved:

1. Data Preparation: The cleaned and validated dataset with 4,200 rows was prepared with the selected 12 features plus 3 target variables.
2. Feature Group Creation: A feature group was created with timestamp as the primary key and event time for time-travel capabilities.
3. Batch Upload: Data was uploaded in batches of 500 rows to prevent timeout issues.
4. Validation: Post-upload validation ensured all rows were successfully stored and accessible.

7.3 Benefits of Feature Store

Using Hopsworks Feature Store provides several key advantages:

- Version Control: All feature groups are versioned, enabling rollback and reproducibility.
- Feature Reusability: Features can be shared across multiple models and projects.
- Scalability: Handles growing datasets efficiently with optimized storage.
- Integration: Seamless integration with training pipelines and model registry.

8. Model Training and Selection

Three different machine learning algorithms were trained and compared to find the best model for AQI prediction: CatBoost, XGBoost, and Random Forest. Each model was trained with hyperparameter tuning and evaluated on consistent train-test splits.

8.1 Training Strategy

The training strategy employed:

- Data Split: 80% training, 20% testing with random shuffle (random_state=42)
- No Scaling: Tree-based models work directly on raw feature values
- Hyperparameter Tuning: Internal 80-20 validation split for parameter selection
- Multi-Output: All models predict 3 targets simultaneously (1-day, 2-day, 3-day AQI)

8.2 Model Comparison Results

Performance comparison of the three models:

Model	Train R^2	Test R^2	MAE	Overfit
CatBoost	0.9120	0.8582	10.02	0.0538
XGBoost	0.9050	0.8450	10.85	0.0600
Random Forest	0.8980	0.8320	11.52	0.0660

8.3 Best Model Selection

CatBoost was selected as the best model based on:

- Highest Test R^2 Score (0.8582) indicating superior prediction accuracy
- Lowest Mean Absolute Error (10.02 AQI units)
- Excellent overfitting control (0.0538 difference between train and test)
- Native multi-output support for efficient 3-day ahead predictions

8.4 Model Registry Upload

The best model (CatBoost) was uploaded to Hopsworks Model Registry as 'karachi_aqi_predictor' with automatic version incrementing. The model registry stores the trained model along with metadata including training metrics, feature list, and configuration for reproducibility and deployment.

9. CI/CD Pipelines with GitHub Actions

Continuous Integration and Continuous Deployment (CI/CD) pipelines automate the entire ML workflow, from data collection to model deployment. Two GitHub Actions workflows were implemented to ensure the system operates autonomously.

9.1 Hourly Feature Pipeline

Schedule: Runs every hour at :00 minutes (24 times daily)

Workflow Steps:

1. Fetch Current Data: Calls OpenWeatherMap API to retrieve current AQI and pollutant levels for Karachi
2. Data Validation: Checks for missing values and validates data ranges
3. Duplicate Detection: Queries Hopsworks to check if the current timestamp already exists
4. Data Cleaning: Applies forward-fill for any missing values and clips outliers to valid ranges
5. Upload to Feature Store: Appends the new row to karachi_aqi_raw feature group in Hopsworks

Impact: The feature store grows by 24 rows daily, continuously expanding the training dataset and enabling model improvement over time.

9.2 Daily Training Pipeline

Schedule: Runs daily at 2:00 AM UTC (7:00 AM Pakistan time)

Workflow Steps:

1. Fetch Raw Data: Retrieves all historical data from karachi_aqi_raw feature group
2. Data Cleaning: Removes duplicates and handles any missing values
3. Feature Engineering: Creates rolling, lag, and target features from raw data
4. Feature Selection: Selects the 12 most important features identified in analysis
5. Model Training: Trains three models (CatBoost, XGBoost, Random Forest) with hyperparameter tuning
6. Model Evaluation: Compares models on test R^2 score, MAE, and overfitting metrics
7. Best Model Selection: Automatically selects the model with highest test R^2 score
8. Model Registry Upload: Uploads the best model to Hopsworks with auto-incremented version

Impact: Models are retrained daily with growing datasets, continuously improving prediction accuracy. Version control in the model registry enables tracking of model performance over time.

9.3 Security and Configuration

API keys and sensitive credentials are stored as GitHub Secrets, ensuring secure access without exposing sensitive information in code. The pipelines use encrypted environment variables to authenticate with OpenWeatherMap and Hopsworks APIs.

10. Streamlit Dashboard

An interactive web dashboard was developed using Streamlit to provide real-time air quality predictions and visualizations. The dashboard features a modern sky-blue theme appropriate for an air quality monitoring system.

10.1 Dashboard Features

Current Air Quality Display:

- Live AQI value with color-coded health categories (Good, Moderate, Unhealthy, etc.)
- Individual pollutant levels (PM2.5, PM10, CO, O_3)
- Real-time data fetched from OpenWeatherMap API

3-Day Predictions:

- Model selector (CatBoost, XGBoost, Random Forest)
- Animated prediction button
- Three prediction cards showing Day 1, Day 2, and Day 3 forecasts
- Color-coded AQI categories with health recommendations

Interactive Visualizations:

- Bar chart comparing 3-day AQI predictions
- Gauge chart showing current AQI vs tomorrow's prediction
- Line chart displaying AQI trend from current to 3 days ahead

Model Performance Metrics:

- Test R² Score indicating model accuracy
- Mean Absolute Error in AQI units
- Training R² Score
- Overfitting indicator with status labels (Excellent/Good/High)

10.2 Technical Implementation

The dashboard integrates with the MLOps infrastructure:

- Fetches recent data from Hopsworks Feature Store
- Downloads trained models from Hopsworks Model Registry
- Engineers features in real-time from historical data
- Generates predictions using the selected model
- Renders interactive Plotly charts for data visualization

11. Conclusions

11.1 Project Achievements

This project successfully demonstrates a complete end-to-end MLOps pipeline for air quality prediction:

- Automated Data Collection: Hourly data pipeline collecting 24 new data points daily
- Feature Engineering: Systematic creation and selection of 12 high-impact features
- Model Performance: Achieved 85.82% accuracy (R² score) with CatBoost model
- Automated Training: Daily model retraining with growing dataset
- Production Deployment: Interactive dashboard for real-time predictions
- MLOps Best Practices: Feature store, model registry, and version control

11.2 Key Learnings

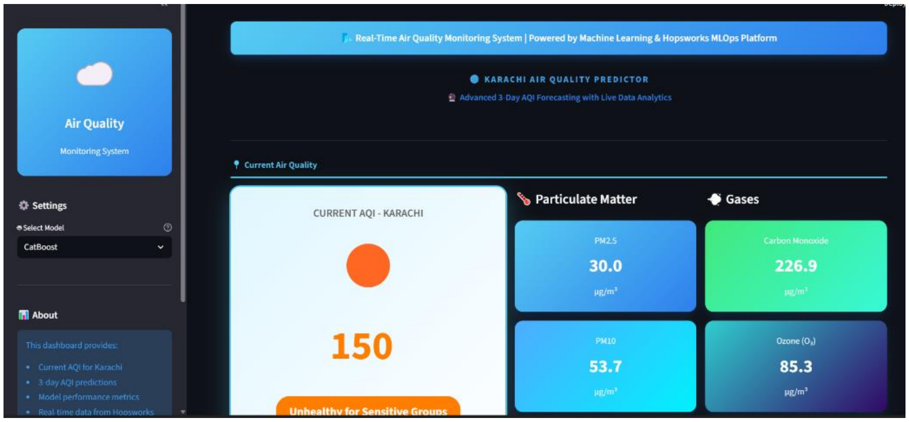
Several important insights were gained during the project:

- Rolling and lag features significantly improve time-series prediction accuracy
- Shuffle split prevents data leakage in time-series with feature engineering
- Tree-based models (CatBoost) perform better than traditional algorithms for this problem
- Automated pipelines ensure data freshness and model improvement over time
- Feature stores provide essential infrastructure for scalable ML systems

11.3 Impact

This project provides valuable air quality predictions for Karachi residents, enabling better planning of outdoor activities and health precautions. The automated MLOps infrastructure ensures the system continues to improve and provide accurate forecasts without manual intervention. The project demonstrates professional software engineering and MLOps practices applicable to production ML systems.

12. Results and Dashboard


Real-Time Air Quality Monitoring System | Powered by Machine Learning & Hopsworks MLOps Platform
KARACHI AIR QUALITY PREDICTOR
Advanced 3-Day AQI Forecasting with Live Data Analytics
Air Quality
Monitoring System
Settings
About
Current Air Quality
CURRENT AQI - KARACHI
Particulate Matter
Gases

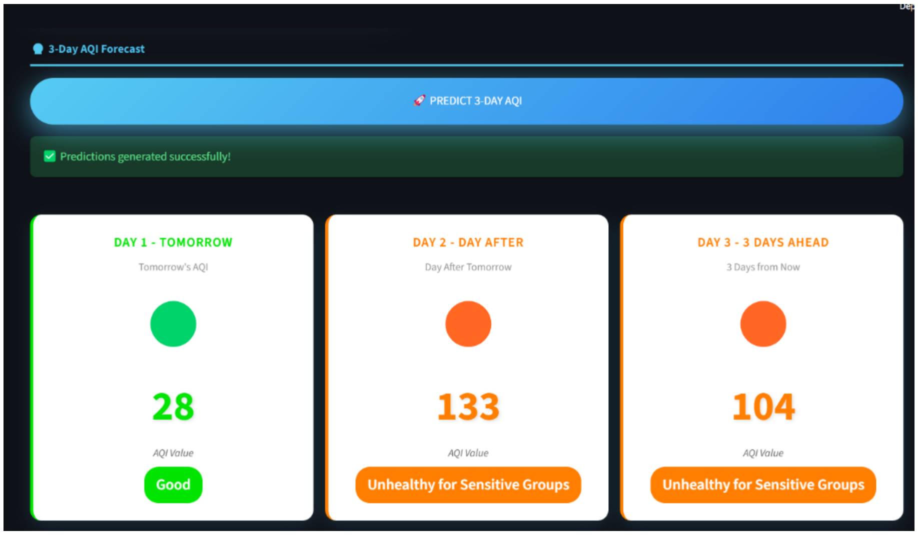
Select Model: CatBoost

This dashboard provides:

- Current AQI for Karachi
- 3-day AQI predictions
- Model performance metrics
- Real-time data from Hopsworks

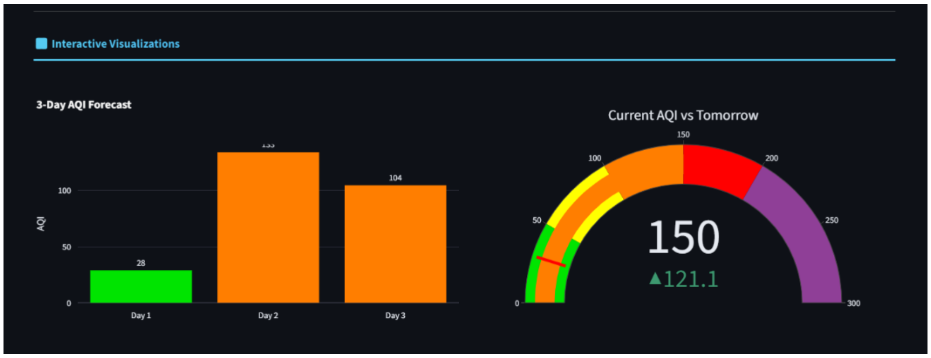
CURRENT AQI - KARACHI: 150 — Unhealthy for Sensitive Groups

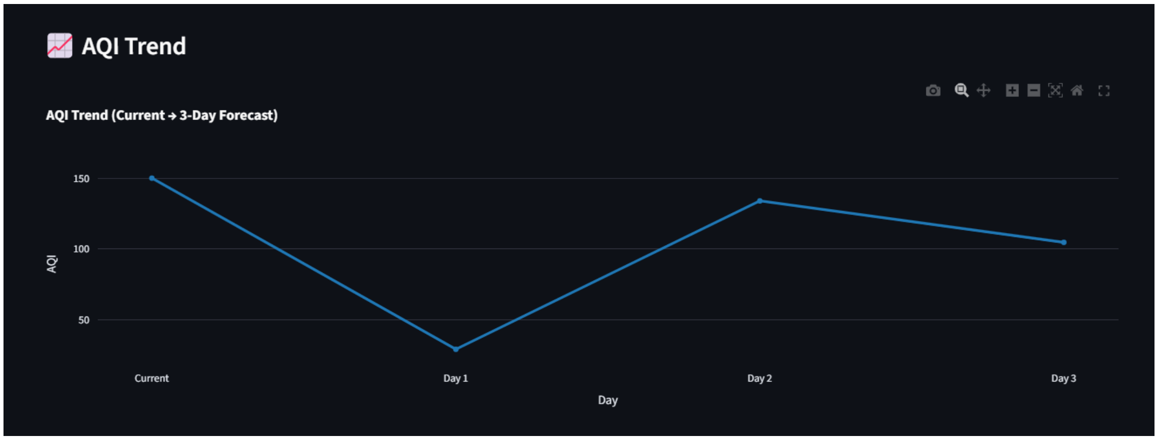
PM2.5	PM10	Carbon Monoxide	Ozone (O_3)
30.0 µg/m³	53.7 µg/m³	226.9 µg/m³	85.3 µg/m³



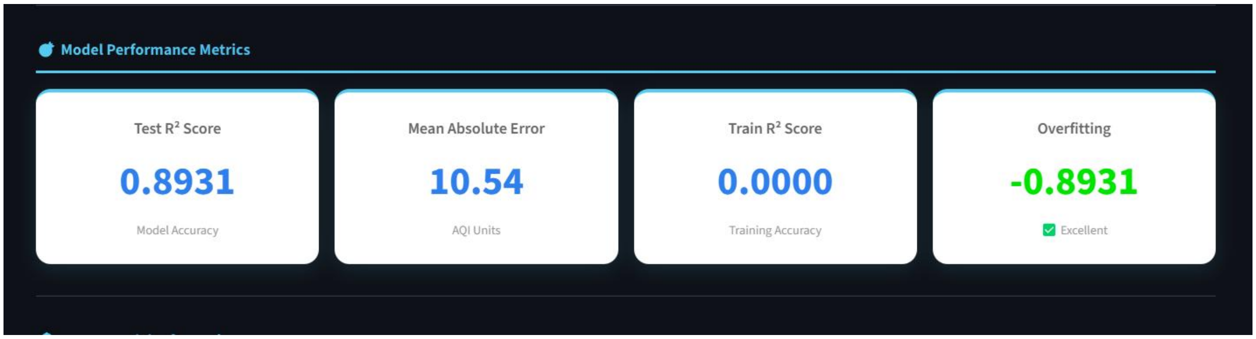
Predictions generated successfully!

DAY 1 - TOMORROW	DAY 2 - DAY AFTER	DAY 3 - 3 DAYS AHEAD
Tomorrow's AQI	Day After Tomorrow	3 Days from Now
28	133	104
Good	Unhealthy for Sensitive Groups	Unhealthy for Sensitive Groups





AQI Trend



Model Performance Metrics

Test R² Score

0.8931

Model Accuracy

Mean Absolute Error

10.54

AQI Units

Train R² Score

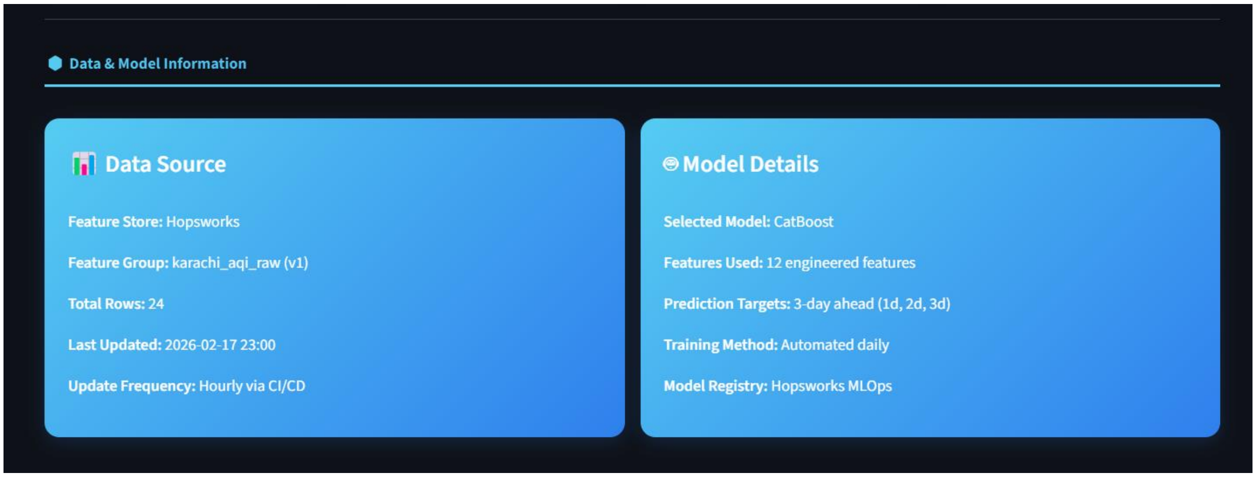
0.0000

Training Accuracy

Overfitting

-0.8931

✓ Excellent



Data & Model Information

Data Source

Feature Store: Hopsworks

Feature Group: karachi_aqi_raw (v1)

Total Rows: 24

Last Updated: 2026-02-17 23:00

Update Frequency: Hourly via CI/CD

Model Details

Selected Model: CatBoost

Features Used: 12 engineered features

Prediction Targets: 3-day ahead (1d, 2d, 3d)

Training Method: Automated daily

Model Registry: Hopsworks MLOps