Data Science Report -10Perals

Project: 10Pearls AQI Predictor Data Science
Name: Hasrat Nazar

1. Summary

This project implements a fully automated Data Science project pipeline designed to predict the AQI for Karachi for the next three days. The system collects air pollution and weather data every hour, cleans and preprocesses it, and retrains five machine learning models daily via GitHub Actions.

All trained models and their performance metrics are stored in the Hopsworks Model Registry. A Streamlit dashboard connects to this registry to fetch the best models and displays live AQI forecasts using OpenWeatherMap data. Users can interactively compare model predictions to identify the most accurate forecasts.

2. Exploratory Data Analysis

EDA was conducted to understand the relationship between air pollutants, weather parameters, and AQI. The insights gained were also incorporated into a dedicated EDA page within the Streamlit dashboard.

Key Observations:

- The AQI provided by OpenWeatherMap was on a simplified scale (1–5). A continuous AQI was calculated using ***U.S. EPA standards*** for better granularity.
- **Primary pollutants influencing AQI:** PM2.5 and PM10.
- **Correlation insights:**
    - AQI increased with PM2.5 and PM10 levels.
    - AQI decreased with higher wind speeds.

Final Features Selected for Modeling:

1. PM2.5
2. PM10
3. O_3
4. Temperature

3. Model Evaluation & Selection

Models Used:

The automated pipeline trains the following regression models daily using the most recent clean dataset:

1. **Ridge Regression:** Linear model
2. **KNeighbors Regressor (KNN):** Distance-based regression
3. **Support Vector Regressor (SVR):** Non-linear regression
4. **Random Forest Regressor:** Ensemble model using bagging
5. **Gradient Boosting Regressor:** Ensemble model using boosting

Evaluation Methodology:

- **Data split:** 80% training, 20% testing
- **Scaling:** StandardScaler applied to all models uniformly
- **Performance Metrics:**
    - **R^2 (R-squared):** Indicates variance explained by the model (1.0 = perfect prediction).
    - **MAE (Mean Absolute Error):** Average absolute error in AQI points.

Model Selection:

All models' results were logged in the Hopsworks Model Registry and visualized in the Streamlit dashboard. *Gradient Boosting Regressor* outperformed the others, showing the highest R^2 and the lowest MAE. This model was selected as the primary predictor, although the dashboard allows users to compare predictions from all five models.

---

4. Final Evaluation & Limitations

Project Achievements:

The project successfully delivered a fully automated MLOps workflow that:

- Collects real-time air pollution and weather data
- Creates predictive features

- Retrains multiple machine learning models daily
- Stores models and metrics in Hopsworks
- Displays an interactive Streamlit dashboard with forecasts, model comparison, and EDA analysis

Frontend:

- **3-Day Forecast:** Displays predicted AQI for each day
- **Model:** 72-hour forecast for both top 2 performance models
- **Performance Tab:** Shows R^2, MAE, and feature importance for top-performing models
- **EDA Page:** Provides interactive analysis of historical AQI and pollutant data

Limits:

1. **Dependence on OpenWeather Data:** Model performance is contingent on the quality and reliability of pollutant and weather data provided by OpenWeather.
2. **Data Model:** Current predictions use pollutant data from the same hour, which can inflate R^2 scores. A realistic forecast should rely on past data to predict future AQI
3. **City-Specific Training:** The model was trained exclusively on Karachi data and may not generalize to other cities without retraining on local datasets.

Conclusion

The Pearls AQI Predictor project successfully demonstrates an end-to-end Machine learning workflow for real-time air quality forecasting. By automating data collection, preprocessing, feature engineering, and model retraining, the system ensures that AQI predictions remain up-to-date and reliable.

The Gradient Boosting Regressor emerged as the most accurate model, though the dashboard allows users to compare multiple models and monitor their performance. Through interactive visualizations and EDA, the project provides insights into how pollutants and weather factors influence air quality.

While the model currently focuses on Karachi and relies on OpenWeather data, the modular MLOps pipeline can be extended to other cities or integrated with alternative data sources. Overall, this project highlights the practical application of MLOps principles in building robust, scalable, and user-friendly predictive systems for environmental monitoring