AQI Forecasting System Report

This report summarizes the design, implementation, and performance of a fully automated MLOps pipeline for hourly Air Quality Index (AQI) forecasting in Karachi, Pakistan. The system delivers a **96-hour forecast** every hour and visualizes historic and future data on a live Streamlit dashboard.

Feature Development & Data Engineering

This part handles data acquisition, cleaning, feature engineering, and exploratory analysis:

- **Data Acquisition & US-EPA Conversion:** Hourly pollutant data ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) is fetched from **OpenWeatherMap**. Proprietary AQI is converted to US-EPA AQI using EPA breakpoints; overall AQI is the maximum of sub-AQIs.
- **Data Cleaning:** Duplicates removed, negative values clipped, outliers capped at 99.5th percentile, and missing values imputed via forward/backward/median fill.
- **Feature Engineering:** Temporal/cyclical features (hour, month, day, weekend + sine/cosine transforms), interaction/derived features (pm_ratio, total_pm, total_gases, no2_o3_ratio), and time-series features (3-hour rolling means, lagged AQI values) for historical context.
- **Exploratory Data Analysis (EDA):** Statistical summaries, correlation analysis, and visualizations (distributions, trends, scatter plots, heatmaps) were used to guide feature selection for model training.

Feature Store & Data Engineering (Hopsworks)

This part includes manages clean feature preparation, validation, and integration with Hopsworks Feature Store:

- **Feature Selection:** 17 features selected from prior EDA, including pollutant concentrations, AQI lag values, cyclical temporal features, and interaction/rolling metrics.
- **Data Cleaning & Validation:** Datetime parsing, missing values filled, invalid timestamps dropped, numeric columns cast to float64, NaN/inf replaced with 0. Deduplication ensures unique timestamps.
- **Hopsworks Integration:** Cleaned data uploaded as a versioned feature group (aqi_features v4), including an auto-generated id primary key. Feature group can be listed, deleted, and re-uploaded safely.
- **Data Fetch & Verification:** Features fetched for training or analysis, with optional date filtering. Local CSV and Hopsworks data compared to ensure consistency (max numeric difference < tolerance).
- **Training Data Ready:** Validated dataset is fully prepared for downstream AQI model training.

Model Training, Evaluation, and Checkpointing

The training dataset consisted of 17,521 samples with 17 features, fetched from the Hopsworks Feature Store.

- **Model Suite:** XGBoost, LightGBM, CatBoost, RandomForest, GradientBoosting, Ridge, and Linear Regression.
- **Time-Series Split:** The data was split chronologically (70% Train, 10% Validation, 20% Test) to simulate real-world forecasting and prevent data leakage.
- **Hyperparameter Tuning & CV:** Models were optimized using key hyperparameters (e.g., max_depth, learning_rate, subsample, colsample) and evaluated with TimeSeriesSplit (n_splits=5). The model with the lowest CV-MAE was chosen as the best_model.

1. Training Performance Analysis

Model	Train MAE	Val MAE	Test MAE	Train R^2	Val R^2	Test R^2

XGBoost	2.024	2.943	3.803	0.999	0.934	0.967
LightGBM	1.823	3.010	4.097	0.999	0.924	0.966
CatBoost	1.958	3.833	4.058	0.999	0.913	0.969
RandomForest	1.163	5.768	5.673	0.999	0.883	0.951
GradientBoosting	1.410	2.796	3.798	0.999	0.928	0.963
Ridge	8.999	7.042	10.268	0.986	0.892	0.857
Linear	8.673	6.675	8.867	0.987	0.898	0.887

- Tree-based models (XGBoost, LightGBM, CatBoost, RandomForest, GradientBoosting) achieved very high training accuracy ($R^2 \approx 0.98$-0.99) and strong generalization on validation and test sets.
- Linear models (Ridge, Linear Regression) underperformed, indicating inability to capture the non-linear AQI patterns.
- The moderate increase from train to test MAE suggests **no severe overfitting**, and the models remain generalizable.

2. Data Drift Consideration

The historical best cross-validation MAE (2.042 for XGBoost) was slightly better than the current run (2.260 for RandomForest).

- **Historical Best Test MAE:** 3.700 (XGBoost)
- **Current Run Test MAE:** 5.673 (RandomForest)
  The minor variation indicates minimal data drift, confirming that the feature distributions remain stable over time.

3. Model Selection

- Based on cross-validation MAE and test performance, **XGBoost** remains the **best model**.
- Although the current run's RandomForest achieved a CV MAE of 2.260, it did not outperform XGBoost's historical best CV MAE of 2.042.
- Therefore, the previously saved XGBoost checkpoint was retained as the best-performing model.

4. Best Checkpoint Model

- **Model:** XGBoost
- **Historical Best CV MAE:** 2.042
- **Historical Best Test MAE:** 3.700
- **Current Run CV MAE (RF):** 2.260
- **Current Run Test MAE (RF):** 5.673

Reason for Checkpoint Retention:
Although XGBoost demonstrated superior forecast performance in this single run, the checkpointing system prioritizes cross-validation MAE stability over isolated test or forecast metrics. Since RandomForest's CV MAE (2.260) was not better than XGBoost's historical best (2.042), the older, more consistent XGBoost model was retained as the best checkpoint.

5. Forecasting Performance

- The 96-hour AQI forecast showed that **tree-based models** (XGBoost, LightGBM, GradientBoosting) significantly outperformed linear models.
- **XGBoost** produced a low forecast MAE of **1.208** and high R^2 of **0.953**, confirming its robustness for short-term AQI prediction.
- **Gradient Boosting** achieved the highest R^2 (**0.981**) but had a slightly higher MAE than XGBoost.
- **Linear and Ridge** models performed poorly, with high forecast errors and negative R^2 values.
- Overall, XGBoost maintained consistent performance across both training and forecasting stages, validating its selection as the best checkpoint model.

6. Continuous Data Shift Management
Environmental time-series data can change due to factors like weather anomalies or policy shifts. The system mitigates this by:

- **Daily Retraining:** Automated daily training updates the model with the latest data, ensuring adaptation to new patterns.
- **Short Forecast Horizon:** Limiting predictions to 96 hours keeps forecasts within a period where data distribution is relatively stable.

Automated CI/CD Pipeline (GitHub Actions)

The MLOps workflow is fully automated via three chained GitHub Actions:

1. **fetch_data.yml:** Hourly/manual trigger; runs fetch_data.py, pulls data, applies US-EPA conversion, commits data, and dispatches data-updated event.
2. **eda.yml:** Triggered by data-updated or manually; runs eda.py for cleaning, feature engineering, selection, uploads features to Hopsworks Feature Store, commits artifacts, and dispatches eda-completed event.
3. **train.yml:** Hourly/manual trigger; runs training.py to retrain models, checkpoint the best model, generate 96-hour forecasts, commit models, metrics, forecasts, comparison plots, and upload artifacts.

Web Application Dashboard (Streamlit Cloud) AQI Prediction Dashboard · Streamlit

A dynamic Streamlit dashboard provides transparent access to system outputs.

- **Deployment:** app.py is deployed on Streamlit Cloud, linked to GitHub for automatic containerization and hourly updates via CI/CD commits.
- **Key Features:**
    - **Future Predictions:** Interactive selection of models (default: best checkpoint) with 96-hour Actual vs. Predicted AQI line charts.
    - **Historical Data:** Zoomable 30-day time-series plots for AQI and pollutants for EDA.
    - **Forecast Comparison:** Bar charts of model metrics (MAE, R^2) for evaluating model drift and performance shifts.