Final Report: AQI Analysis & Forecasting for Indian Cities

Pushkaraj Gaikwad
Project: Wipro DICE ID Internship

1. Executive Summary

This report details the analysis of historical air quality data (2015-2020) for major Indian cities. Exploratory Data Analysis revealed severe seasonal pollution patterns, with Delhi being the most affected city. Two time-series models, ARIMA and Prophet, were developed to forecast future PM2.5 levels. The Prophet model demonstrated superior performance (RMSE: XX.XX) and predicts a recurrence of high pollution in the upcoming winter months. Based on these findings, we recommend implementing targeted seasonal emission controls and proactive public health advisories.

2. Introduction & Problem Statement

The objective of this project was to analyze historical air quality data to identify pollution trends and build a reliable forecasting model. The goal is to provide data-driven insights that can help policymakers make informed decisions to mitigate the public health risks associated with poor air quality.

3. Methodology

The project followed a structured data science workflow:

- **Data Preprocessing:** Cleaned raw CPCB data, handled missing values using time-series-appropriate methods (forward-fill), and standardized data types.
- **Exploratory Data Analysis (EDA):** Used Matplotlib and Seaborn to visualize trends, compare cities, and analyze pollutant correlations.
- **Time-Series Forecasting:**
  - **ARIMA:** A statistical model was built after making the data stationary through differencing.
  - **Prophet:** A component-based model was used to capture seasonality and the effects of national holidays.
- **Model Evaluation:** Models were trained on pre-2020 data and evaluated on 2020 data using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics.

4. Key Findings from EDA

Finding 1: Delhi is a High-Risk Pollution Zone

Analysis of average AQI levels shows that Delhi is significantly more polluted than other major metropolitan areas.

Finding 2: Severe Winter Seasonality

Air quality has a strong, predictable seasonal pattern. Pollution levels begin to rise in October and peak in the winter months (November-January), likely due to meteorological factors and seasonal events.

Finding 3: Vehicular and Industrial Emissions are Key Contributors

A strong positive correlation was found between pollutants like NO2, CO, PM2.5, and PM10, suggesting that vehicular and industrial combustion are primary sources of pollution.

5. Forecasting Results

Both models were tasked with forecasting Delhi's PM2.5 levels for 2020. The Prophet model performed better, as it more accurately captured the seasonal dips and peaks.

Model	MAE	RMSE
ARIMA	XX.XX	XX.XX
Prophet	YY.YY	YY.YY

6. Recommendations

Based on the analysis, we propose the following actionable recommendations:

1. **Implement Targeted Seasonal Controls:** Enforce stricter emission norms and traffic controls specifically during the forecasted peak pollution window of October-January.
2. **Issue Proactive Health Advisories:** Use the forecast to warn the public 2-4 weeks before severe pollution episodes are expected.
3. **Promote Urban Greening:** Focus urban planning efforts on expanding green zones, which serve as natural filters for particulate matter.

7. Conclusion

This project successfully demonstrated the use of data analysis and time-series forecasting to understand and predict air pollution in India. The developed models and insights provide a strong foundation for data-driven environmental policymaking.