Delhi AQI Forecasting: Stats Models vs Deep Learning

Pranit Ahuja
M.S. in Data Science
Michigan State University

Introduction

Delhi faces critical air pollution levels and the situation is worsening day by day.

The Air Quality Index (AQI) is used for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you

Goal: Compare SARIMA and LSTM (and other DL) models for AQI forecasting (2018–2024)

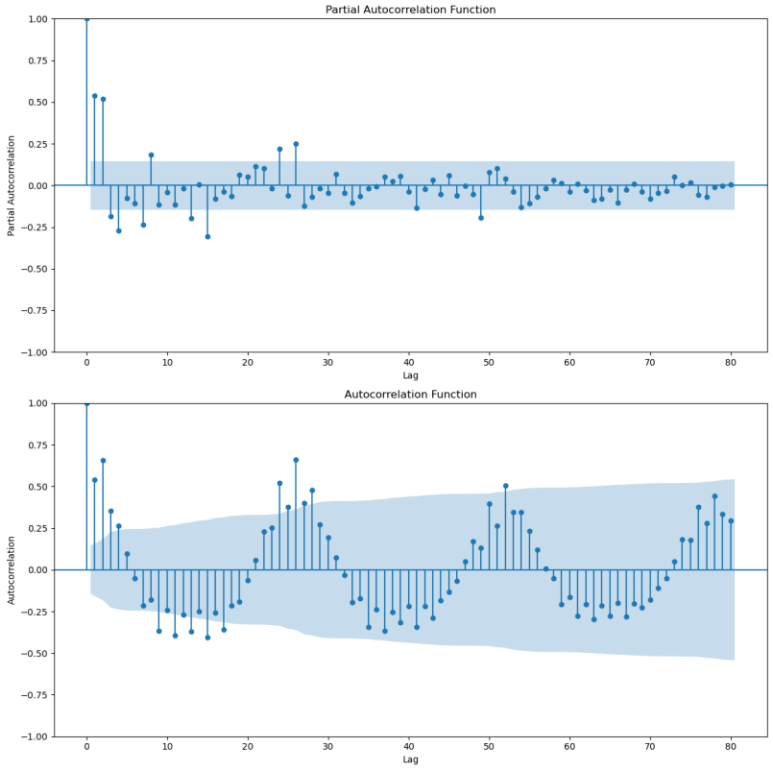
Data Overview

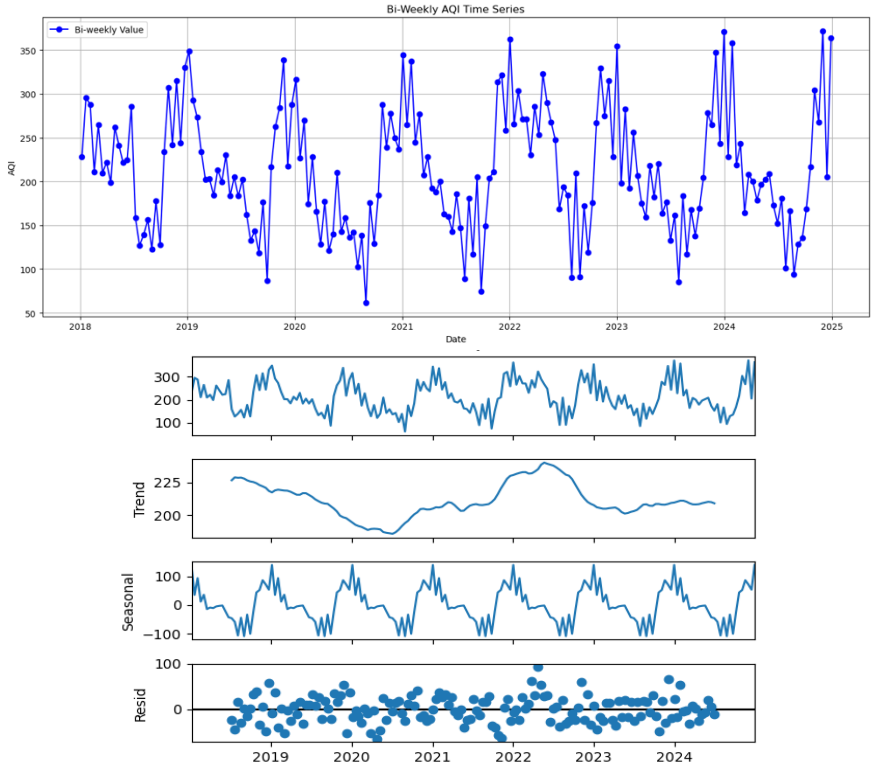
- Columns: AQI, NO2, CO2, O3, PM

- Preprocessing: -
  - Parsing the 'Date' column to `datetime` and setting it as index
  - Dropping two empty unnamed columns
  - First-order differencing to remove long-term trend
  - Creating binary event flags to encode repeating seasonal drivers: Diwali, crop-burning (Punjab-Haryana stubble), winter inversion, monsoon, summer dust storms, New-Year fireworks, Independence-Day traffic.
  - Scaling: for LSTM all features were Min-Max normalised to [0, 1]. No missing values remained after aggregation

Exploratory Analysis

- ACF shows geometrically decay.
- PACF shows 4 significant lags, suggesting AR(4) component.
- Seasonal patterns same every year.
- Seasonal decomposition confirms this pattern.





SARIMA Modeling

1. Applied first-order differencing to remove trend
2. Model: SARIMA(4,1,2)(1,0,0,26)
3. Fitted on data up to end of 2023 and forecasted for 2024.

Model Tuning

- Tried removing MA component: fit worsened
- Seasonal components also necessary for a good fit
- Model selected based on Log-Likelihood and AIC scores

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \omega_t$$

AR(p)

$$X_t = \omega_t + \sum_{j=1}^{q} \theta_j \omega_{t-j}$$

MA(q)

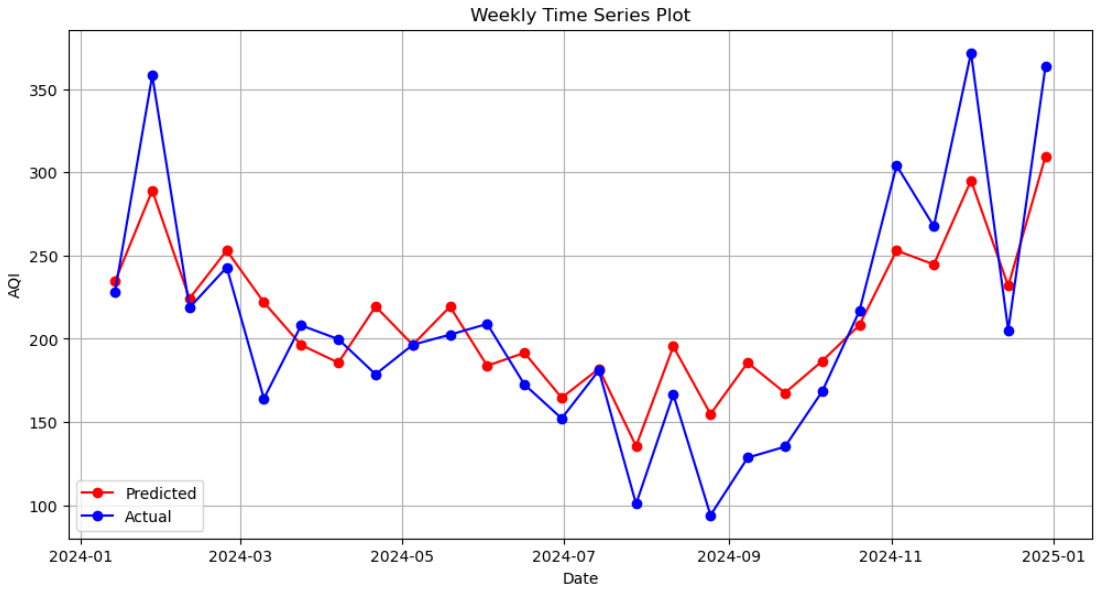
$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \omega_{t-j} + \omega_t$$

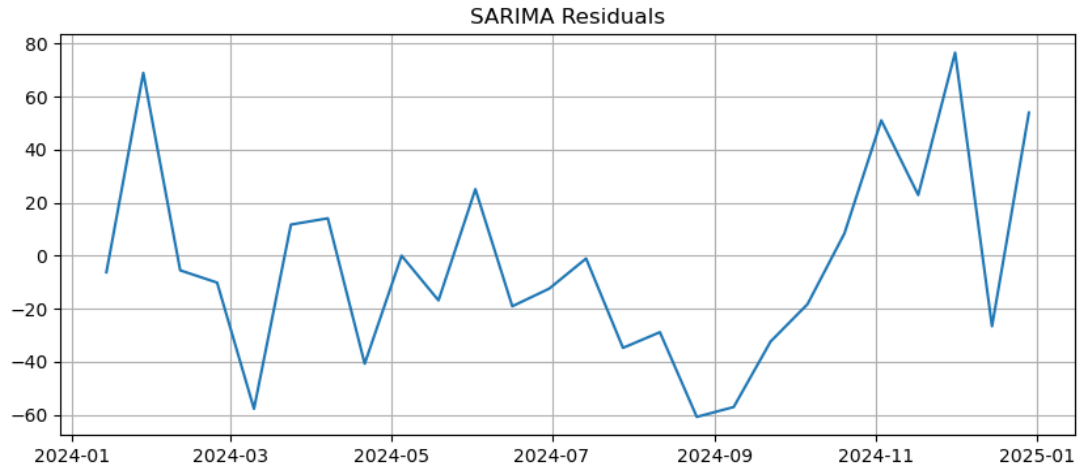
ARMA(p,q)

Evaluation (SARIMA)

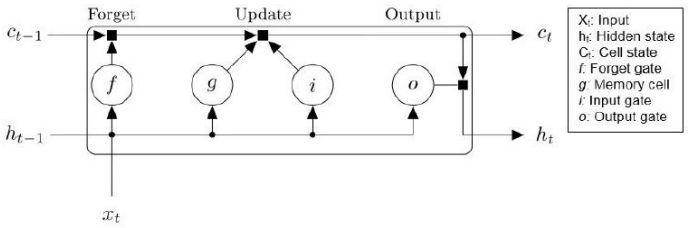
SARIMA is designed to model seasonal patterns and sudden changes explicitly using its parameters:

- **(p,d,q):** captures short-term AR and MA effects
- **(P,D,Q,s):** handles seasonal cycles (like winter smog, Diwali)

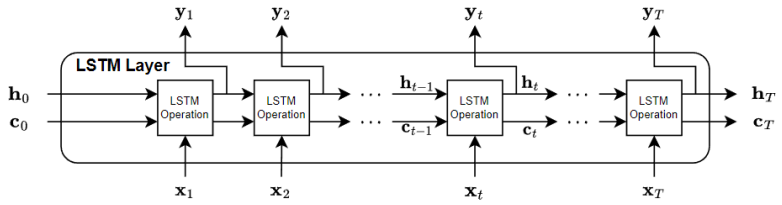




LSTM Modeling



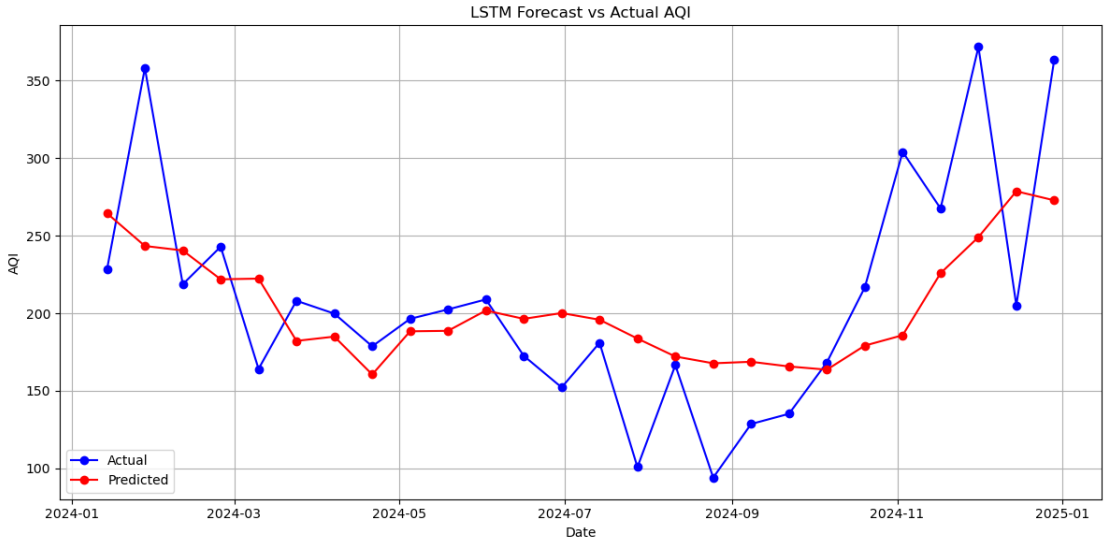
1. Sequential model
2. Three LSTM layers with 128, 64 and 32 units each
3. Dropout layers with rate 0.3 to prevent overfitting
4. Final Dense layer for output
5. Fitted on data up to end of 2023 and forecasted for 2024

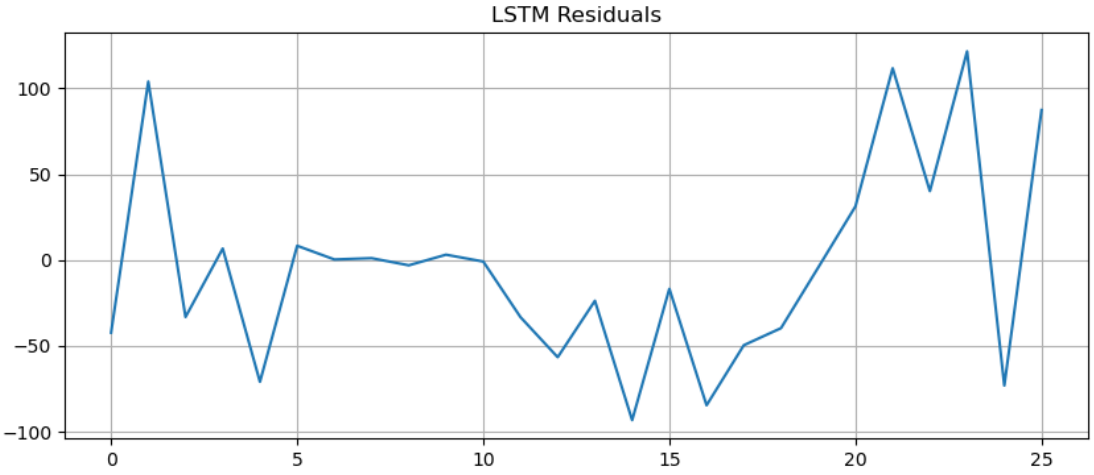


Evaluation (LSTM)

LSTMs are trained to minimize average error so tend to average out noisy fluctuations to minimize loss — especially when:

- The training data has high variance or spikes (like AQI)
- There's no external context (e.g., festivals, weather) to explain sudden changes





Feature Engineering

SARIMA is able to identify the trend and seasonality in the data to better understand the spikes and downfalls in the overall data. Till now LSTM has only been given AQI series without any pattern or event information which it cannot deduce on its own.

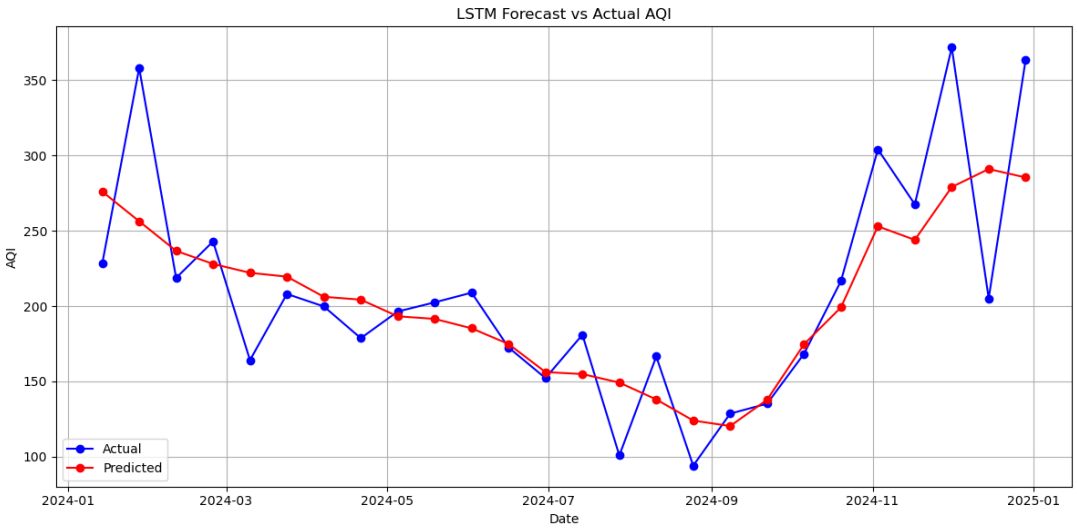
Now, it will have one-hot encoded features of events mapped with their respective 2-week periods.

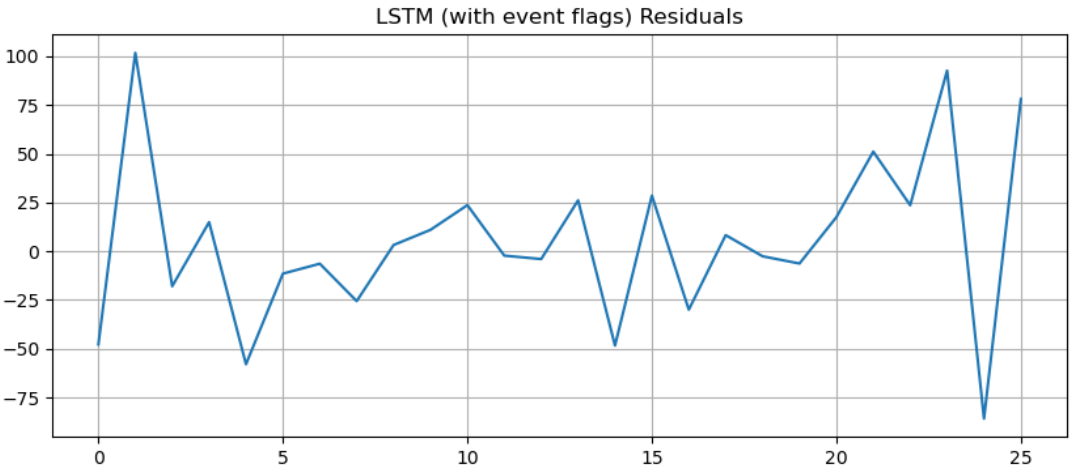
Date	AQI	diwali	crop_burning	winter	monsoon	summer	new_year	independence_day
2018-01-07	0.538107	0	0	1	0	0	1	0
2018-01-21	0.754548	0	0	1	0	0	0	0
2018-02-04	0.730140	0	0	1	0	0	0	0
2018-02-18	0.482155	0	0	1	0	0	0	0
2018-03-04	0.655768	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
2024-11-03	0.781948	1	1	0	0	0	0	0
2024-11-17	0.664287	0	1	0	0	0	0	0
2024-12-01	1.000000	0	0	1	0	0	0	0
2024-12-15	0.463044	0	0	1	0	0	0	0
2024-12-29	0.973521	0	0	1	0	0	0	0

183 rows × 8 columns

Evaluation (LSTM with event flags)

Features that map certain events to their respective dates allow the LSTM to better understand the repeating or seasonal patterns in the data and help it gain a better overall understand of the trend over time.





Evaluation Metrics

Model	Features	R^2	RMSE	MAE
SARIMA (4,1,2)(1,0,0,26)	AQI	0.74	36.55	29.27
LSTM (univariate)	AQI	0.36	57.08	44.28
LSTM (+ event flags)	AQI + 7 flags	0.64	43.45	32.76