Time Series Forecasting of PM2.5 Level

Presenter: Hassan Sajjad

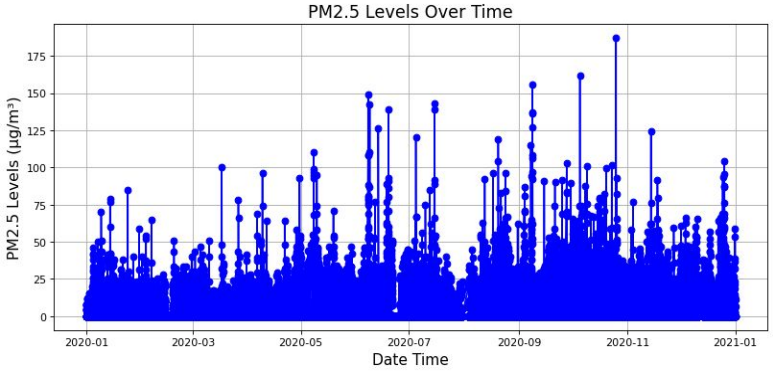
Dataset

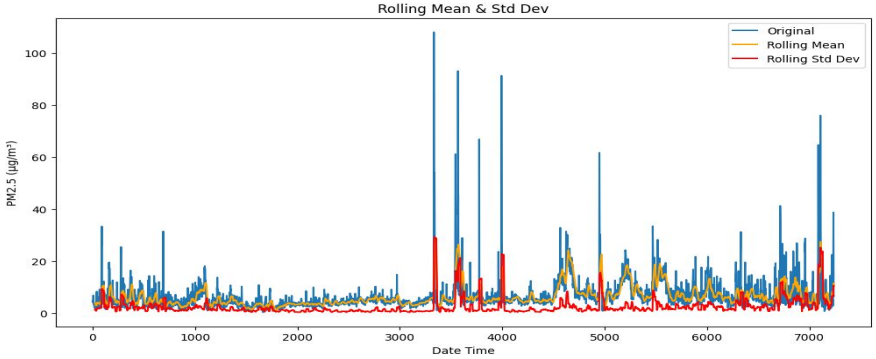
- Dataset downloaded from Openaq site using aws cli
- Dataset download using predefined command *aws s3 cp --no-sign-request s3://openaq-data-archive/records/csv.gz/locationid=2178/year=2020 data --recursive* for year 2020 of Del Norte-2178
- Columns: datetime, location_id, sensors_id, location, lat, lon, units, value
- Task:Predict pm25 level for upcoming year

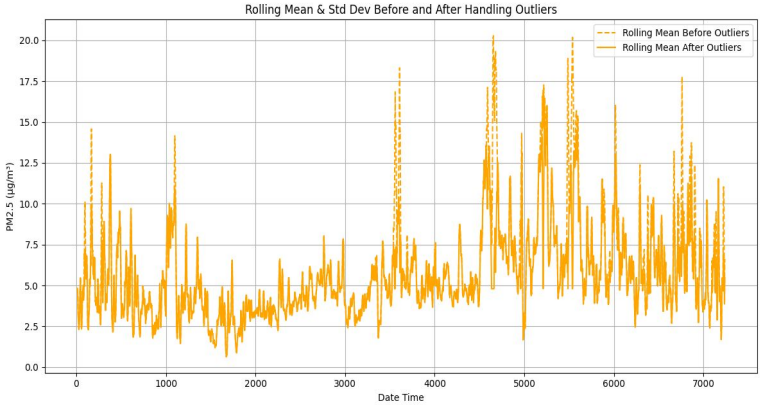
Data preprocessing

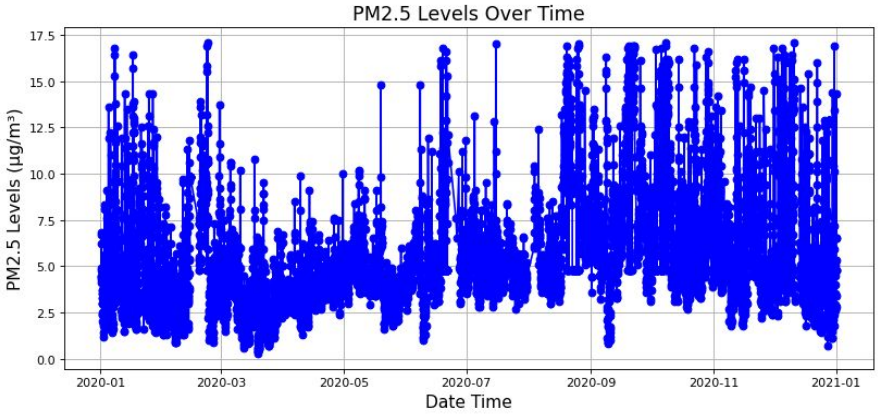
- Downloaded data of year 2020 ,contains 12 month data,combined into a single dataframe.
- Converting ‘datetime’ column and setting it as index.
- Removing other pollutants and keeping only rows for pollutant “**pm25”**
- **Handling Outliers:** Compute rolling mean and standard deviation with a 12-period window.
- **Handle Missing Values:** Interpolate missing values using the nearest method.
- **Detect and Replace Outliers:** Identify values outside 3 standard deviations and replace with the median.
- **After Handling Outliers:** Recalculate rolling mean and standard deviation with the cleaned data.
- Dropping extra columns and selecting columns for training model.
- After preprocessing,7239 rows are selected for pm25 pollutant from 42789 rows.

Data preprocessing









Data preprocessing

	location_id	sensors_id	location	datetime	lat	lon	parameter	units	value
0	2178	3919	Del Norte-2178	2020-01-01T01:00:00-07:00	35.1353	-106.584702	pm10	µg/m³	7.0
1	2178	3919	Del Norte-2178	2020-01-01T02:00:00-07:00	35.1353	-106.584702	pm10	µg/m³	8.0
2	2178	3919	Del Norte-2178	2020-01-01T03:00:00-07:00	35.1353	-106.584702	pm10	µg/m³	7.0
3	2178	3919	Del Norte-2178	2020-01-01T04:00:00-07:00	35.1353	-106.584702	pm10	µg/m³	8.0
4	2178	3919	Del Norte-2178	2020-01-01T05:00:00-07:00	35.1353	-106.584702	pm10	µg/m³	8.0

Raw data

datetime	parameter	value
2020-01-01 08:00:00	pm25	4.4
2020-01-01 09:00:00	pm25	4.1
2020-01-01 10:00:00	pm25	4.3
2020-01-01 11:00:00	pm25	4.8
2020-01-01 12:00:00	pm25	4.5

Data after preprocessing

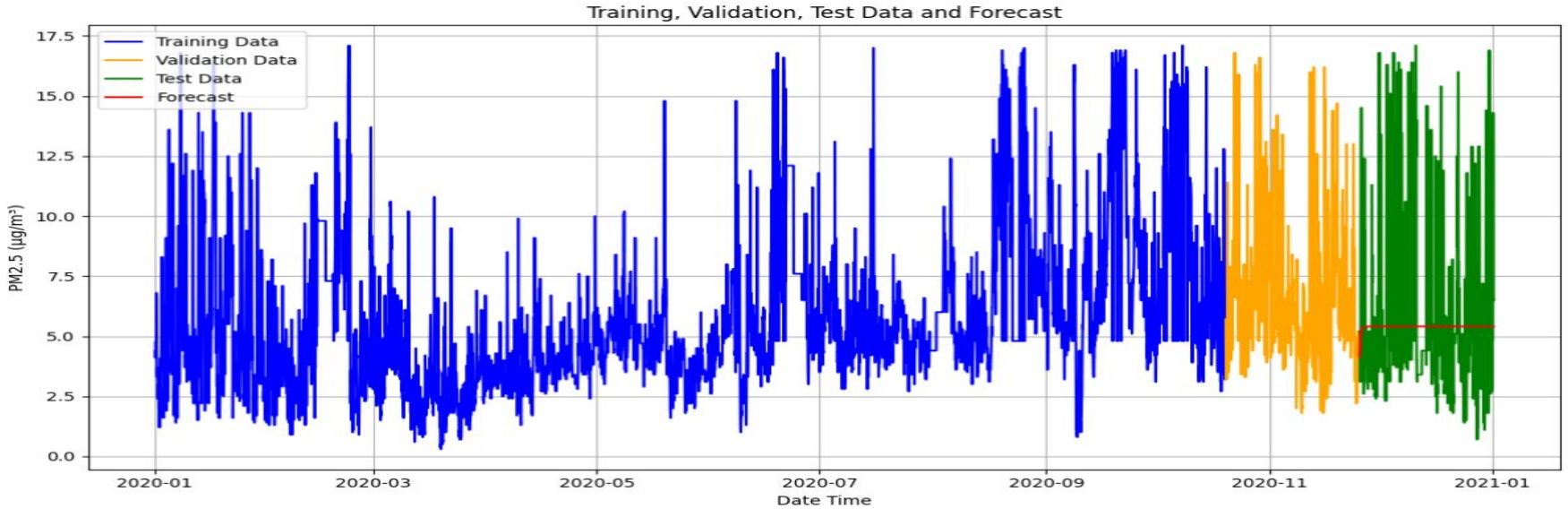
Model Selection

- Various models are analyzed for forecasting the pm25 level.Prophet,LSTM,XGboost,ARIMA,SARIMA
- Chosen Model: Auto ARIMA
- Why chosen : Automatically identifies the best parameters for the ARIMA model.
- It ensures accurate and efficient modeling of non-seasonal data. Its built-in diagnostics and robust performance metrics make it user-friendly and reliable for predicting PM2.5 levels.

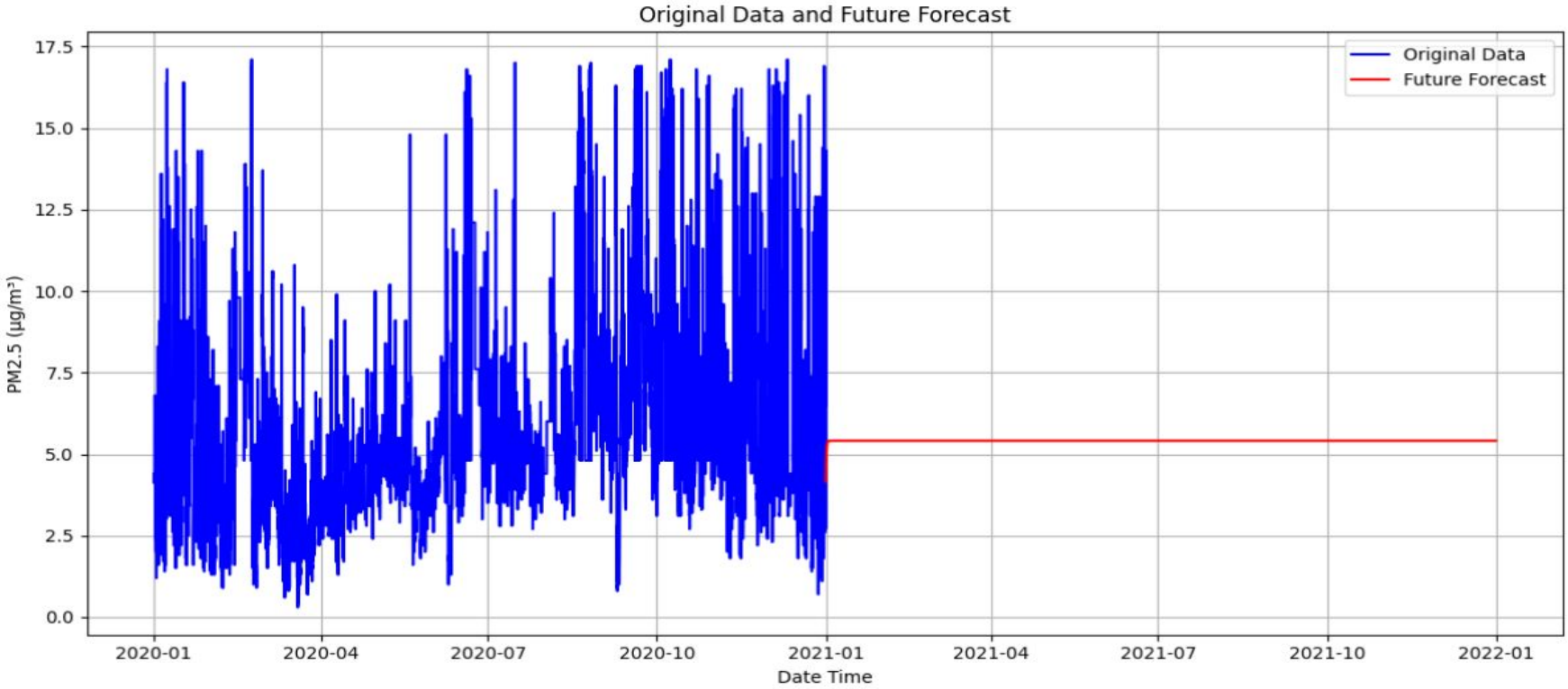
Model training

- Split data into training (80%), validation (10%), and test (10%) sets.
- Metric Used: Mean Squared Error (MSE)

```
Best model:  ARIMA(1,1,4)(0,0,0)[0]
Total fit time: 111.663 seconds
Validation Mean Squared Error: 9.479715332800003
Test Mean Squared Error: 11.38777979188266
```



Future forecasting



LSTM for forecasting

- LSTM are type of neural network model used for time series forecasting as they capture long term dependencies and temporal pattern.Also address issue like vanishing gradient.
- Sequential model -BiLSTM :100 units ,dropout 0.2 LSTM 50 units, dense 25 units
- Activation:relu
- optimizer adam
- Loss function: Mean Square error

```
181/181 [==============================] - 4s 22ms/step - loss: 0.0077 - val_loss: 0.0139
Epoch 45/50
181/181 [==============================] - 4s 23ms/step - loss: 0.0076 - val_loss: 0.0143
Epoch 46/50
181/181 [==============================] - 4s 22ms/step - loss: 0.0077 - val_loss: 0.0142
Epoch 47/50
181/181 [==============================] - 4s 23ms/step - loss: 0.0076 - val_loss: 0.0139
Epoch 48/50
181/181 [==============================] - 4s 22ms/step - loss: 0.0076 - val_loss: 0.0146
Epoch 49/50
181/181 [==============================] - 4s 22ms/step - loss: 0.0076 - val_loss: 0.0144
Epoch 50/50
181/181 [==============================] - 4s 22ms/step - loss: 0.0076 - val_loss: 0.0140
Test Mean Squared Error: 11.707341895696024
```

LSTM for forecasting