Project Title :

Carbon Monoxide and Temperature Forecasting using ARIMA Model

Team Members :

1. Yuvnish Malhotra, 4th year, IIT Indore
2. Muskan Pardasani, 4th year, IIT Indore
3. Siddhesh Shelke, 3rd year, IIT Indore
4. Tanishq Selot, 2nd year, IIT Indore

The Blue Sky Challenge

About the Problem Statement :

Air quality has a profound impact on personal and societal well-being around the world. There are various advantages to improved air quality, including significant health, environmental, and economic benefits. Air quality in major cities throughout the world is becoming a subject of worry as a result of increased urbanisation and industrialisation. Several countries have tried to develop smart city programmes, in which sensors play an important role in notifying both government officials and the general public about real-time air quality levels via mobile or web-based apps. Traditional sensor monitoring may be made smarter by incorporating cutting-edge machine learning algorithms, allowing for an advance in present air quality monitoring capabilities. In this regard, the hackathon's sub-theme 2 aims to find new and innovative ways to construct smart air quality monitoring systems by combining sensor technology with machine learning algorithms.

A number of factors in the air can have an impact on its quality. Multiple sensors monitoring various parameters are used in air quality monitoring sensing systems, which are available as a whole suite. The role of temperature and carbon monoxide in air quality is vital. The following issue that may be addressed via this hackathon in order to make such systems smarter:

Temporal forecasting of temperature and Carbon Monoxide (CO) sensor data one day ahead: It can assist the general public and government officials in anticipating trends early in order to make timely decisions and take preventative actions. Advanced machine learning algorithms combined with sensor data have the potential to be a leap forward and in addressing the problem listed above. Therefore, the primary emphasis of this sub-theme 2 is on the development of machine learning algorithm to solve the defined problem.

About the Data

The Data consists of:

1. 9357 rows of data points.
2. The data is collected between Mar-2004 and Apr-2005 from an Italian City.
3. CO is measured in mg/m^3. Temperature is measured in degree celsius.
4. -200 is identified as an anomaly in the dataset.
5. The maximum temperature observed was *44.6 degrees celsius*, whereas the maximum CO concentration was *11.9 mg/m^3*.

Contributed by:

Saverio De Vito (saverio.devito '@' enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development.

Data Preprocessing

We carried out the following steps:

1. We used the dropna() function of pandas to remove any NaN values.
2. Finally after getting rid of negative values in CO and Temperature, we are left 7331 values.
3. Normalization was done to bring the CO and temp values to a scale between 0 and 1.
4. Date and Time were made as the index columns for better representation.
5. This data was converted into pandas.series format for training. 80% used for training while 20% used for testing.

Some stats related to data after preprocessing:

Carbon Monoxide

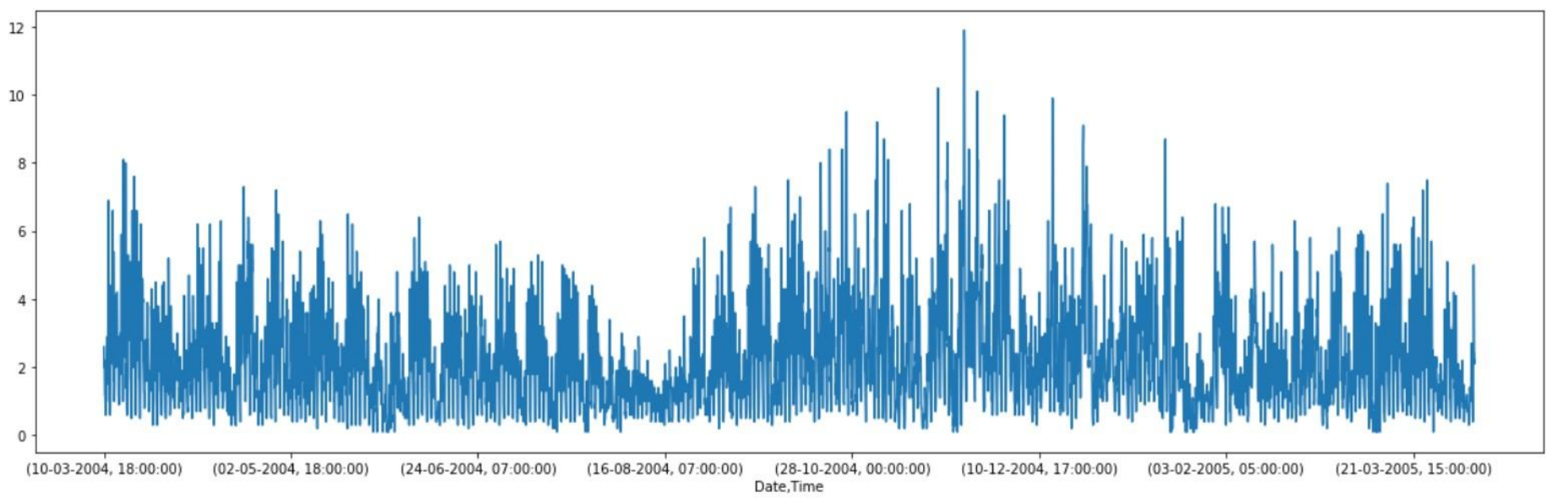
```
count    7331.000000
mean        0.179155
std         0.120702
min         0.008403
25%         0.092437
50%         0.151261
75%         0.235294
max         1.000000
Name: co, dtype: float64
```

Temperature

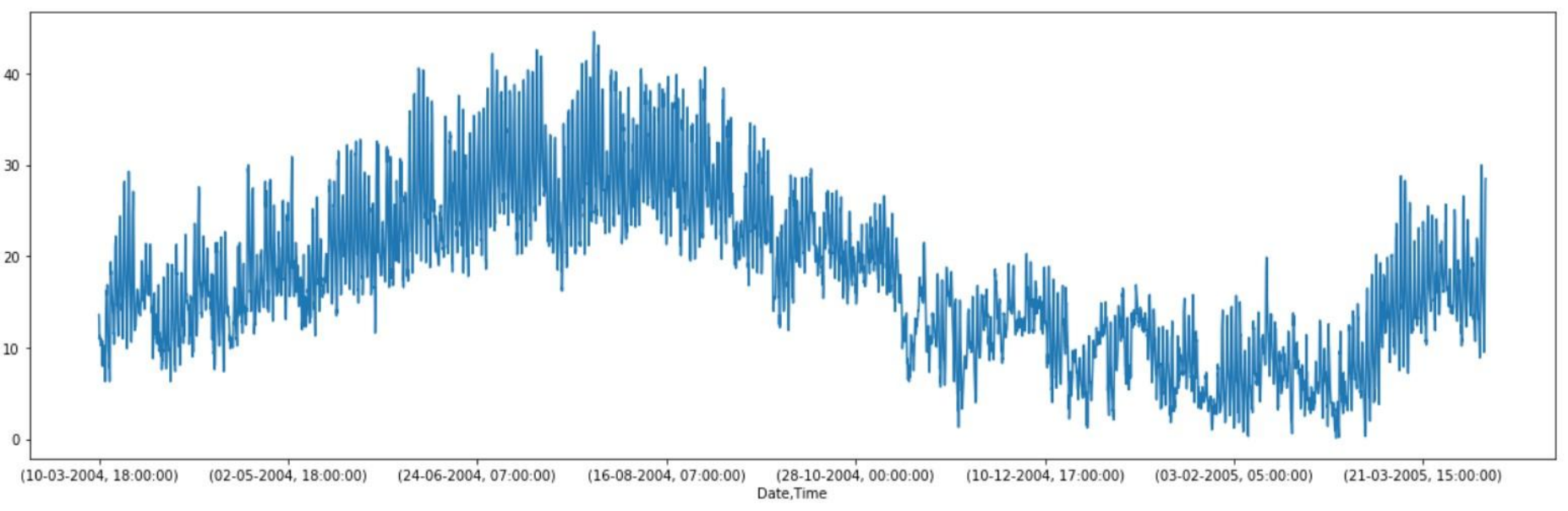
```
count    7331.000000
mean        0.399175
std         0.198122
min         0.002242
25%         0.251121
50%         0.378924
75%         0.533632
max         1.000000
Name: t, dtype: float64
```

Data Visualization:

A very noisy ***Carbon monoxide series***: no trend or seasonality whatsoever.



Temperature series: Some seasonality can be observed due to summer and winter seasons.

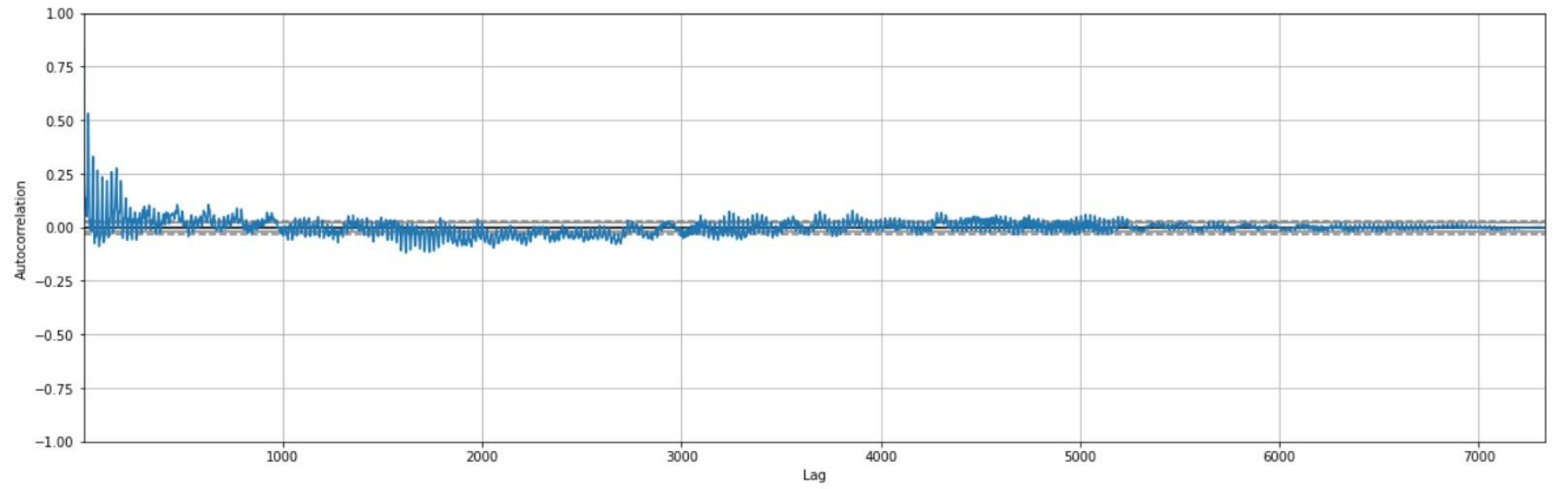


Autocorrelation plot for Carbon Monoxide:

- Autocorrelation represents the **degree of similarity between a given time series** and a lagged version of itself over successive time intervals.

- Positive autocorrelation means that **the increase observed in a time interval leads to a proportionate increase in the lagged time interval**.

- A negative autocorrelation implies that **if a particular value is above average the next value (or for that matter the previous value)** is more likely to be below average. Here, we observe high autocorrelation initially.


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Autocorrelation
1000
2000
3000
4000
5000
6000
7000
Lag

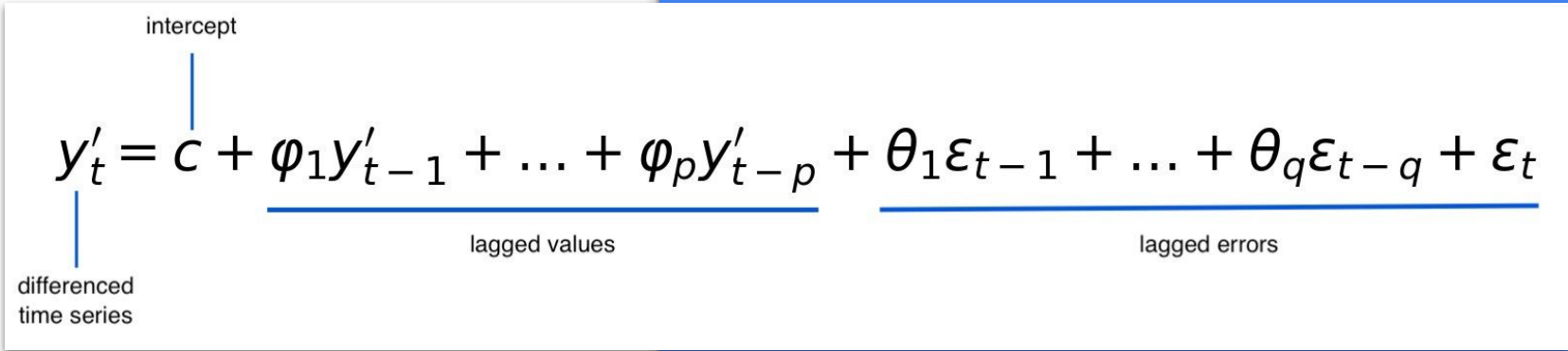
Previously used Models

- We started with an artificial neural network model which performed poorly (high MAPE losses).
- Hence, we used LSTM. They also didn’t perform upto the mark as they are not very good with regression problems.

- We finally implemented the widely used **ARIMA** (*Auto Regressive Integrated Moving Average*) model.

- The motto behind using ARIMA was to get rid of any seasonality or trend as they negatively affect a regression model.
- ARIMA does this by applying *differencing*.

ARIMA Model



More about ARIMA

- ARIMA uses a number of lagged observations of time series to forecast observations. A weight is applied to each of the past term and the weights can vary based on how recent they are. Autoregression is a process of regressing a variable on past values of itself. Autocorrelations gradually decay and estimate the degree to which white noise characterizes a series of data.

- Integrated is a property that reduces seasonality from a time series. ARIMA models have a degree of differencing which eliminates seasonality.

- Moving average (MA) removes non-determinism or random movements from a time series.

- *p*: the number of lag observations in the model; also known as the lag order.
- *d*: the number of times that the raw observations are differenced; also known as the degree of differencing.
- q: the size of the moving average window; also known as the order of the moving average.

ARIMA Parameters

Parameters used and final metrics:

p	7
d	1
q	0
MAPE for CO	0.3061981657100548
MAPE for Temp	0.10167239154329467

Thank you