Time Series Analysis and Forecasting Using ARIMA

DSAI Challenge *Omkar Oak 112103099 Notion notes 23-10-2023*

Dataset: `ForEx Rates 2006-2023`

Model Used: Auto Regressive Integrated Moving Average (`ARIMA`)

- Time Series Analysis is a way of studying the characteristics of the response variable concerning time as the independent variable.
- To estimate the target variable in predicting or forecasting, use the time variable as the reference point

Components of Time Series Analysis

- **Trend**: A trend is a long-term increase or decrease in the data. A positive trend indicates that the data is increasing over time, while a negative trend indicates that the data is decreasing over time.
- **Seasonality**: Seasonality refers to patterns in the data that repeat at regular intervals (e.g., monthly, quarterly, annually). For example, sales of winter jackets may be higher in the winter months and lower in the summer months.
- **Stationarity**: A time series is said to be stationary if its statistical properties (e.g., mean, variance) are constant over time. Many time series forecasting methods assume that the data is stationary, so it is often necessary to transform the data in some way (e.g., by taking differences or logarithms) to make it stationary.
- **Cyclicity:** refers to patterns in the data that repeat at irregular intervals. Cyclicity is often caused by internal factors that have a cyclical influence on the data (e.g., economic cycles). For example, the stock market may experience cycles of bull and bear markets, with periods of strong growth followed by periods of decline.

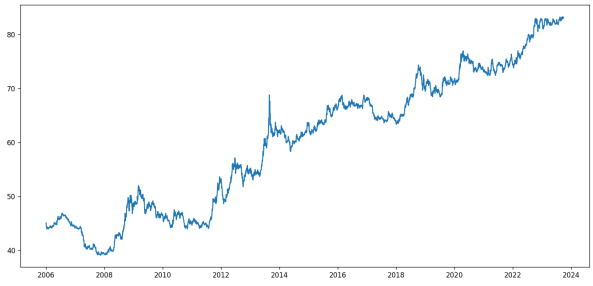
- **Autocorrelation:** Autocorrelation is the degree to which the value of a time series at a given point is correlated with the value of the series at previous points. Autocorrelation can be used to identify patterns in the data and help select an appropriate forecasting method.
- **Forecast horizon**: The forecast horizon is the number of periods ahead that you want to make predictions for.
- **Irregularity**: Unexpected situations/events/scenarios and spikes in a short time span.

How should the data be for TSA?

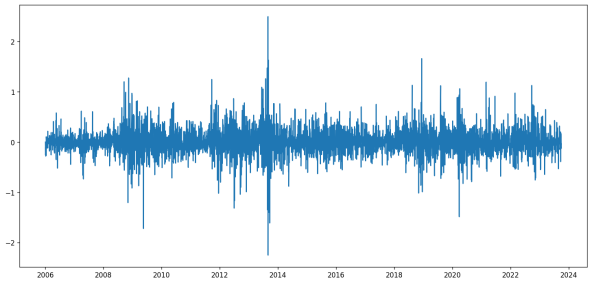
- It should be **stationary** (constant mean, constant variance)
- It should be **not seasonal**

Methods to Check Stationarity

- During the TSA model preparation workflow, we must assess whether the dataset is stationary or not. This is done using **Statistical Tests.**



Non-stationary data (p-value=0.94)



Stationary data (p-value=0.0)

- There are two tests available to test if the dataset is stationary:

1. Augmented Dickey-Fuller (ADF) Test

The ADF test is the most popular statistical test. It is done with the following assumptions:

- Null Hypothesis (H_0): Series is non-stationary
- Alternate Hypothesis (H_A): Series is stationary
  - p-value >0.05 : non-stationary (H_0)
  - p-value <= 0.05 : stationary (H_A)

2. KPSS Test

Converting Non-Stationary Into Stationary:

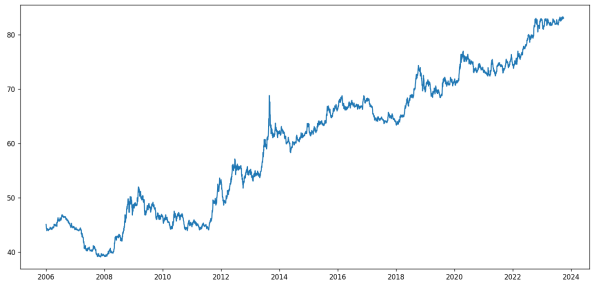
- There are three methods available for this conversion – detrending, differencing, and transformation.

Differencing:

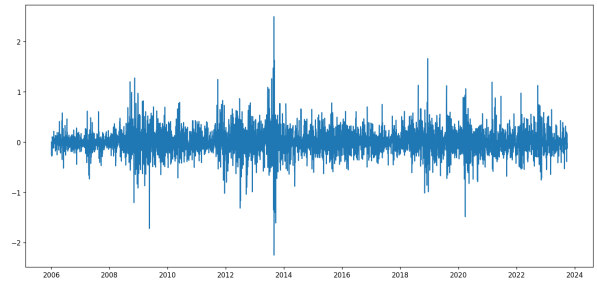
- Differencing involves subtracting the previous value from the current value to remove the trend
- so trend and seasonality are also reduced during this transformation.
- Given by the formula: $Y_t = Y_t - Y_{t-1}$, where Y_t=Value with time

```
#creating a new column `diff` consisting of difference
df['diff'] = df['USD'].diff().fillna(0)
```

- creates a new column called `diff` in `df`
- It calculates the difference between consecutive values in the `USD` column using the `diff()` method.
- The `fillna(0)` part fills any `NaN` values resulting from the differencing operation with zeros.



USD column (p-value=0.94)



diff column (p-value=0.0)

Date	USD	diff
2006-01-02	45.075	0.000
2006-01-03	44.965	-0.110
2006-01-04	44.705	-0.260
2006-01-05	44.600	-0.105
2006-01-06	44.320	-0.280
2006-01-09	44.250	-0.070
2006-01-10	44.185	-0.065
2006-01-11	43.915	-0.270
2006-01-12	44.020	0.105
2006-01-13	44.100	0.080

Transformation: This includes three different methods they are Power Transform, Square Root, and Log Transfer. The most commonly used one is Log Transfer.

Auto Regressive Model (AR)

Regression: Used to predict continuous value of an item based on certain parameters
Auto: Uses its own past values to predict future values

- An auto-regressive model is a simple model that **predicts future performance based on past performance.**
- It is mainly used for forecasting when there is some correlation between values in a given time series and those that precede and succeed (back and forth).
- An AR is a **Linear Regression model** that uses lagged variables as input

- AR(1) 1st order Auto regression: $y_t = C_1 + C_2 \cdot y_{t-1}$
  - So, y_t depends only on 1 previous value (i.e. y_{t-1})
- AR(Q) q^{th} order Auto regression:

$$y_t = C_0 + C_1 \cdot y_{t-1} + C_2 \cdot y_{t-2} + \ ... \ + C_q \cdot y_{t-q}$$

- Example: 12th Marks = C1 + C2(11th Marks) + C3(10th Marks)

Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF)

Correlation: An indicator of relationship between two variables
Auto-Correlation: Relationship of a variable with its own previous time period values (Lags)

Example: Consider the Auto Regression: 12th Marks = C1 + C2(11th Marks) + C3(10th Marks)

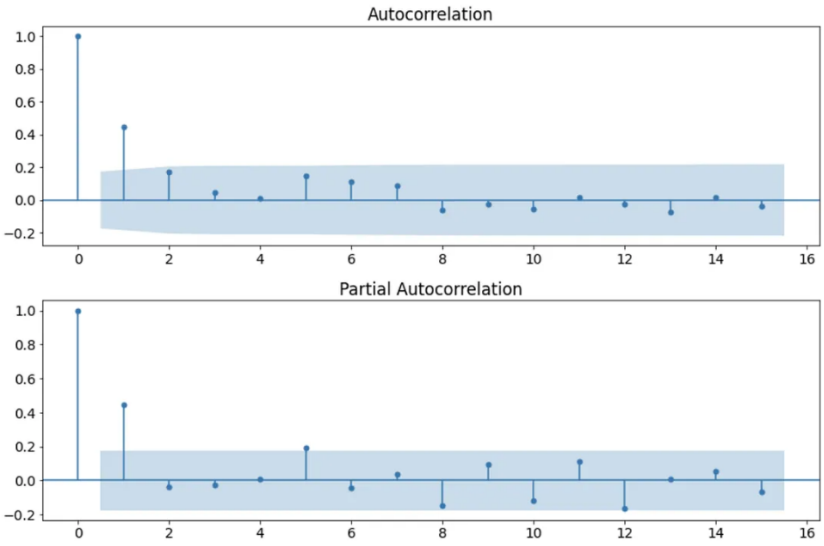
Here, there is a auto correlation between 12th marks and 11th and 10th marks (here time = which std the student is in)

- ACF indicates **how similar a value is with its previous value**. It measures the **degree of the similarity between a given time series and the lagged version of that time series** at the various intervals we observed.
- ACF : Direct and Indirect effect of Values in Previous time Lags
  - E.g. Indirect effect: Consider 12th marks are correlated with 10th marks, so 10th marks can affect 11th marks which further affect 12th marks
- PACF : Only direct effect of Values in Previous time Lags
- Both the ACF and PACF are used to identify patterns and trends in the data and to help select an appropriate model for forecasting future values of the time series.
- Interpreting ACF and PACF plots:

ACF	PACF	Perfect ML -Model
Plot declines gradually	Plot drops instantly	Auto Regressive model.
Plot drops instantly	Plot declines gradually	Moving Average model
Plot decline gradually	Plot Decline gradually	ARMA
Plot drop instantly	Plot drop instantly	You wouldn't perform any model

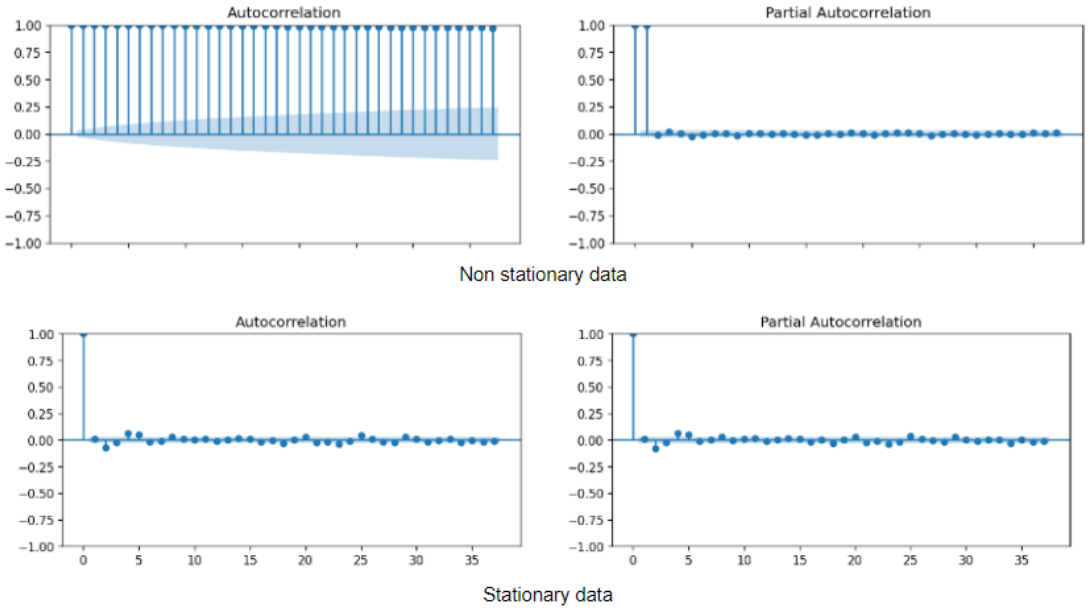
- The ACF plot can provide answers to the following questions:
  - Is the observed time series **white noise/random**?
  - Can the observed time series be modeled with an **MA model**? If yes, what is the order?
- The PACF plot can provide answers to the following question:
  - Can the observed time series be modeled with an **AR model**? If yes, what is the order?

- Both the ACF and PACF start with a **lag of 0**, which is the correlation of the time series with itself and therefore results in a **correlation of 1.** Additionally, you can see a **blue area** in the ACF and PACF plots. This blue area depicts the 95% confidence interval and is an indicator of the **significance threshold**.



Example of ACF and PACF

- Example of ACF and PACF for stationary and non-stationary data



Moving Average Models (MA)

- Models that **predict future values of a time series using the past errors**

- Comparing MA and AR,
  - AR(1) 1st order Auto regression: $y_t = C_1 + C_2 \cdot y_{t-1} + \varepsilon_t$
  - MA(1) 1st order Moving average: $y_t = \mu + C_1\varepsilon_{t-1} + \varepsilon_t$
- So, MA(Q) i.e. q^{th} order Moving Average is

$$y_t = \mu + C_1\varepsilon_{t-1} + C_2\varepsilon_{t-2} + ... + C_q\varepsilon_{t-q} + \varepsilon_t$$

Auto Regressive Moving Average (ARMA)

- The ARMA model **combines** the autoregressive (AR) and moving average (MA) models.

Auto Regression AR: Use past values to make a prediction

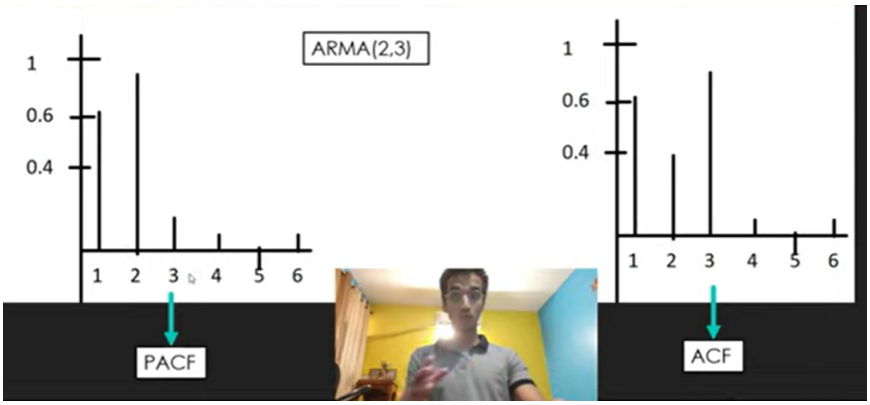
Moving Average MA: Use past errors to make a prediction

Consider,

- AR(1): $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$
- MA(1): $y_t = \theta_0 + \theta_1\varepsilon_{t-1} + \varepsilon_t$
- So, the ARMA(1,1) model will be:

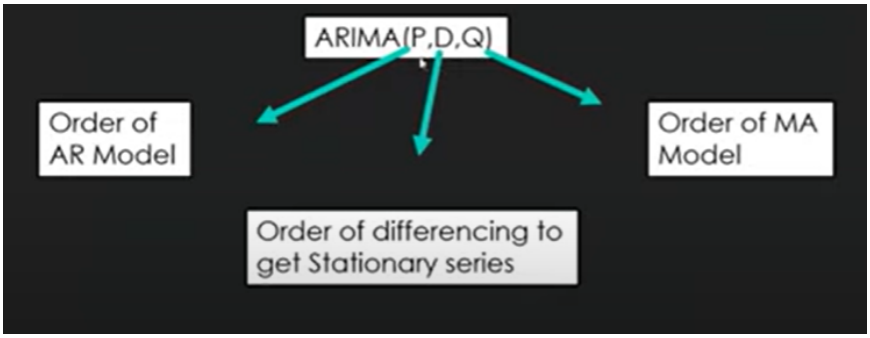
$$y_t = B_0 + B_1 y_{t-1} + \theta_1\varepsilon_{t-1} + \varepsilon_t$$

- So, an $ARMA(p, q)$ model has **AR of order** p i.e $AR(p)$ and **MA of order** q i.e $MA(q)$
- The highest spike in ACF will give the order of MA i.e. q
- The highest spike in PACF will give the order of AR i.e. p



Auto Regressive Integrated Moving Average (ARIMA)

- ARIMA is similar to ARMA, but the only difference is that we have to **convert a Non-stationary series to stationary series** before applying the model
  - **AR** = Auto regression (Use **past values** to make a prediction)
  - **I** = Integrated (**Differencing operation** to convert non-stationary to stationary)
  - **MA** = Moving Average (Use **past errors** to make a prediction)
- So, the order of the model, **ARIMA(p,d,q)** is given by



To implement an ARIMA model, we need to follow these steps:

1. Visualize the time series data to identify any trends, seasonality, or patterns.
2. Determine the order of the ARIMA model (p, d, q). This can be done using techniques such as the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

3. Fit the ARIMA model to the data using the determined values for p, d, and q.
4. Use the fitted model to make forecasts for future values of the time series.

```
from statsmodels.tsa.arima.model import ARIMA
# Fit the ARIMA model
model = ARIMA(data, order=(2,1,2))
results = model.fit()
# Make forecasts
forecasts = results.forecast(steps=12)
```

View file: `DSAI_induction.ipynb` from `112103099_Omkar_Oak_DSAI_induction`