Introduction To ARIMA for Time Series Forecasting

Auto Regressive Integrated Moving Average (ARIMA) model is among one of the more popular and widely used statistical methods for time-series forecasting.

Introduction to Time Series Forecasting

A time series is a sequence where a metric is recorded over regular time intervals.

Depending on the frequency, a time series **can be of yearly (ex: annual budget), quarterly (ex: expenses), monthly (ex: air traffic), weekly (ex: sales qty), daily (ex: weather), hourly (ex: stocks price), minutes (ex: inbound calls in a call canter) and even seconds wise (ex: web traffic).**

Now forecasting a time series can be broadly divided into two types.

If you use only the previous values of the time series to predict its future values, it is called **Univariate Time Series Forecasting**.

And if you use predictors other than the series (a.k.a exogenous variables) to forecast it is called **Multi Variate Time Series Forecasting**.

ARIMA, short for '**AutoRegressive Integrated Moving Average',** is a forecasting algorithm based on the idea that **the information in the past values of the time series can alone be used to predict the future values.**

An ARIMA model is characterized by 3 terms: p, d, q

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, **has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'**. More on that once we finish ARIMA.

The first step to build an ARIMA model is to make the time series stationary.

Because, term 'Auto Regressive' in ARIMA means it is a **linear regression model** that uses its **own lags as predictors. Linear regression models, as you know, work best when the predictors are not correlated and are independent of each other.**

So how to make a series stationary?

The most common approach is to difference it. **That is, subtract the previous value from the current value.** Sometimes, depending on the complexity of the series, more than one differencing may be needed.

The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

Next, what are the 'p' and 'q' terms?

'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. And **'q' is the order** of **the 'Moving Average' (MA) term.** It **refers to the number of lagged forecast errors that should go into the ARIMA Model.**

What are AR and MA models?

A pure **Auto Regressive (AR only) model** is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

where**, Y{t-1} is the lag1 of the series,** beta1 is the coefficient of lag1 that the model estimates and `alpha` is the intercept term, also estimated by the model.

Likewise, a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.

Data scientists use an Augmented Dickey-Fuller (ADF) test to determine whether the data is stationary.