**ARIMA Model – Complete Guide to Time Series Forecasting in Python**

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

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 **MultiVariate 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.

**Introduction to ARIMA Models:**

* ARIMA, short for **‘Auto Regressive Integrated Moving Average**’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the **lagged forecast errors**, so that equation can be used to forecast future values.
* Any ‘non-seasonal’ time series that exhibits patterns and is not a random **white noise** can be modeled with ARIMA models.
* An ARIMA model is characterized by 3 terms: **p, d, q,** where,
  + **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.

**What does the p, d and q in ARIMA model mean:**

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

Why? The 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**

* So what are AR and MA models? What is the actual mathematical formula for the 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’.

**Formula 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.

where the error terms are the errors of the autoregressive models of the respective lags.

* So what does the equation of an ARIMA model look like?

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the **AR** and the **MA** terms. So the equation becomes:

**Equation ARIMA**

* ARIMA model in words:

**Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)**

The objective, therefore, is to identify the values of p, d and q. But how?

Let’s start with finding the ‘d’.

**How to find the order of differencing (d) in ARIMA model**

* The purpose of differencing it to make the time series stationary.

But you need to be careful to not over-difference the series. Because, an over differenced series may still be stationary, which in turn will affect the model parameters.

* So how to determine the right order of differencing?

The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick.

If the autocorrelations are positive for many number of lags (10 or more), then the series needs **further differencing**. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably **over-differenced**.

In the event, you can’t really decide between two orders of differencing, then go with the order that gives the **least standard deviation** in the differenced series.

**Autocorrelation** and **partial autocorrelation** plots are heavily used in time series analysis and forecasting.

There are two primary way to determine whether a given time series is stationary:

* Rolling Statistics : Plot the rolling mean and rolling standard deviation. The time series is stationary if they remain constant with time.
* Augmented Dickey-Fuller Test : the time series is considered stationary if the p-value is low (according to null hypothesis) and the critical values are 1%, 5%, 10% confidence intervals are close as possible to the ADF Statistics.

The **AR(p)** models tend to capture the mean reversion effect whereas **MA(q)** models tend to capture the shock effect in error ,which are not normal or unpredicted events. Thus, the **ARMA** model combines the power of **AR** and **MA** components together. An **ARMA(p, q)** time series forecasting model incorporates the **pth** order AR and **qth** order MA model, respectively.



Here, **Φ and θ** represent AR and MA coefficients. The α and εt captures the intercept and error at time t. The form gets very complicated as p and q increase; thus, lag operators are utilized for a concise representation of ARMA models.

There are multiple scenarios to select p and q; some of the thumb rules that can be used to determine the order of **ARMA** components are as follows:

* **Autocorrelation** is exponentially decreasing and **PACF** has significant correlation at lag 1, then use the p parameter.
* **Autocorrelation** is forming a sine-wave and **PACF** has significant correlation at lags 1 and 2, then use second-order value for p.
* **Autocorrelation** has significant autocorrelation and PACF has exponential decay, then moving average is present and the q parameter needs to be set up.
* **Autocorrelation** shows significant **serial correlation** and the **PACF** shows sinewave pattern, then set up a moving average q parameter.