# Time Series Univariate Analysis

Comparing Standard Time Series Models on Nifty Fifty.The Indian Index data set for past 10 years.

# Abstract

This report highlights the working of different Time Series models on the Indian market Index data. The aim of the project is to identify a model best fitting the Nifty Fifty data and to forecast the Indian index. The data from 2010 to 2019, which have been documented daily ,reveal that the ARMA(2,3) model may fit most adequately.


Department of Mathematics,Delhi Technological University

Date of Submission:28 Jan 2021

Submitted by:
Sadhvi Mehra
2k19/MSCMAT/13
M.Sc.Mathematics,Second Year,DTU

Submitted to:
Ms.Sumedha Seniaray

# Contents

# 1 Introduction

Time series is a collection of data points indexed over time. Time series are analyzed in order to understand the underlying structure that produce the observation. Time series data is of two types:

1. **Univariate Time Series** - it is a time series in which observations are sequentially recorded on a single variable over time.

2. **Multivariate Time Series** - it is a time series in which observations are sequentially recorded on more than one variable over time.
In this project, we have done a Time Series Analysis of Nifty Fifty dataset. Stock prices are not randomly generated values rather they can be treated as a discrete time series.Time series data for stock market prediction can be collected on a daily, weekly, monthly or yearly basis. It is important to identify a model to analyze the trends of stock prices for decision making and make predictions for the future. The basic assumption made while forecasting stock data is that future market trends are influenced by the stock prices in the past.

This means,the historical stock data provides an insight into its future behavior. So, we can fit different time series models namely, AutoRegressive(AR) and AutoRegressive Moving Average(ARMA),and AutoRegressive Integrated Moving Average (ARIMA) models and choose the best model in terms of maximum accuracy, for forecasting.In order to determine the best model for forecasting the index of Indian Market, we have converted our non-stationary data to stationary by removing the deterministic components,namely Trend and Seasonality from the data. Then we fitted AutoRegressive model and AutoRegressive(AR) Integrated Moving Average (ARIMA) to this stationary data and concluded AutoRegressive with lag 4 to be the best model as it has high accuracy on the training dataset. We have done all the analysis using Python programming language and used its inbuilt libraries for the project. The remaining part of the project covers the following topics. The Theory section has information about all the different tests and definitions used in the analysis. Next, Data and Method section consists all the necessary tests and mathematical analysis of the time series. Then, we have used AR model for forecasting. In the next section,we have mentioned all the results we have obtained in the previous section.

# 2 Theory

## 2.1 Relative Ordering Test

This is a non parametric test procedure used for testing the existence of trend components.Let the time series be denoted by $\{X_1, X_2, ...X_n\}$
Define

$$q_{ij} = \begin{cases} 1, \text{ if } X_i > X_j \text{ when } i < j \\ 0, \text{ otherwise} \end{cases}$$

$Q = \sum_{i=1} \sum_{j=2} q_{ij}$

where Q counts the number of decreasing points in time series.
Null Hypothesis $H_0$:There is no trend in time series against the Alternate hypothesis $H_1$:There is trend in time series.
Under the null hypothesis $E(Q) = \frac{n(n-1)}{4}$.If observed $Q << E(Q)$,then it would be an indication of rising trend.If observed $E(Q) << Q$,then it would be an indication of falling trend.If observed Q doesn't differ "significantly"from $E(Q)(under H_0)$ then it would indicate no trend in time series.Q is related with kendall's $\tau$,the rank correlation coefficient,through the relationship

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

Under $H_0, E(\tau) = 0$ and $V(\tau) = \frac{2(2n+5}{9n(n-1)}$
Test Statistic :$Z = \frac{\tau - E(\tau)}{\sqrt{Var(\tau)}}$follow N(0,1) asymptotically.We should reject null hypothesis of no trend at level of significance $\alpha$ if $|Z| > \tau_{\frac{\alpha}{2}}$,where $\tau_{\frac{\alpha}{2}}$is the $\alpha/2^{th}$ upper cut off point of standard normalisation.

## 2.2 Turning Point Test

This is a non parametric test procedure for testing randomness of time series. A turning point is defined as peak when the value is greater than its 2 neighbouring values or a trough when a value is less than its 2 neighbouring values.All n-2 time points (2,3,..,n-1) are checked for being declared as turning points.Define

$$U_i = \begin{cases} 1,\text{if i is a turning point} \\ 0\ ,\text{otherwise} \end{cases}$$

P is the total number of turning points.i.e.$\sum_{n=2}^{n-1} U_i$
Null Hypothesis $H_0$ Series is purely random
Alternate Hypothesis $H_1$:Series is non-random.

Under the Null hypothesis
$E(P) = \frac{2(n-2)}{3}$and $Var(p) = \frac{16n-29}{90}$

Test Statistic :$Z = \frac{P - E(P)}{\sqrt{Var(P)}}$follow N(0,1) asymptotically.We should reject null hypothesis of no trend at level of significance $\alpha$ if $|Z| > \tau_{\frac{\alpha}{2}}$,where $\tau_{\frac{\alpha}{2}}$is the $\alpha/2^{th}$ upper cut off point of standard normalisation.

## 2.3 Dickey-Fuller Test

The Dickey–Fuller test, tests the null hypothesis that a unit root is present in an autoregressive model. We are testing as our null hypothesis is that our time series is actually non-stationary, The idea with Dickey Fuller test is that we start off with an AR process

$$X_t = \phi_0 + \phi_1 X_{t-1} + \epsilon_t$$

$$H_0 : \phi_1 = 1(Time\ Series\ is\ non - stationary)$$

$$H_1 : \phi_1 \leq 1(Time\ Series\ is\ Stationary)$$

The Regression model can be written as

$$X_t - X_{t-1} = \phi_0 + (\phi_1 - 1)X_{t-1} + \epsilon_t$$

$$\Delta X_t = \phi_0 + \delta X_{t-1} + \epsilon_t$$

We can calculate the t-statistics on estimated value of $\delta$, Since the test is done over the residual term rather than raw data, it is not possible to use standard t-distribution to provide critical values so we compare the t-statistics with the value of Dickey-fuller distribution, if t <D.F $\rightarrow$ we reject $H_0$

# 3 ARMA Model

Given a time series $X_t$,the ARMA model is a tool to understand and predict future values of the series. The AR part involves regressing variables on its own lag and MA part involves modeling the error term upto suitable lag as a linear combination.
Suppose p is the lag for AR model and q is the lag for MA model. Then AR(p) model will be

$$X_t = c + \sum_{i}^{p} \phi_i X_{t-i} + \epsilon_t$$

where $\phi_0, \phi_1, ..., \phi_p$ are parameters with $\phi_p \neq 0$ c is a constant and $\epsilon_t$ is white noice. Similarly, MA(q) model will be

$$x_t = \mu + \epsilon_t + \sum_{1}^{q} \theta_i \epsilon_{t-i}$$

where $\theta_1, \theta_2, ..., \theta q$ are parameters of the model with $\theta_q \neq 0$, $\mu$ is the expectation of $X_t$ (can be assumed to be zero) and $\epsilon_t, \epsilon_{t1}, ...$ are again white noise terms.Hence ARMA(p,q) model can be written as

$$X_t = c + \epsilon_t + \sum_{1}^{p} \phi_i X_{t-i} + \sum_{1}^{q} \theta_i \epsilon_{t-i}$$

with usual assumptions on parameters.

# 4 ACF and PACF plot

Autocorrelation and partial autocorrelation plots are heavily used in time series analysis and forecasting. These are plots that graphically summarize the strength of a relationship between an observation of a time series and observations at prior time steps. Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) give us different view points of time series.

A partial autocorrelation plot is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed i.e. PACF only describes the direct relationship between an observation and its lag. This would suggest that there would be no correlation for lag values beyond k.

While ACF describes the autocorrelation between an observation and another observation

at a prior time step that includes direct and indirect dependence information.
The lags p and q of AR and MA model can be determined from ACF and PACF plots. For plotting ACF, we compute correlation between $X_t$ and $X_{t-k}$ for different lag values k and plot them in the graph. So it is quite natural to have negative values as well. Now when if the correlation values come within the significance band, then we can assume that the correlations are indifferent from zero. So, generally we take that value of k for MA as q, for which the correlation will cross the significant band for the last time i.e. we can assume current time point $X_t$ is directly and indirectly dependent on previous q many time points. Now for plotting PACF, we regress current data point on previous time series data points. Then we plot the coefficients on the graph where each coefficients indicate the effect of corresponding previous data points. Now similar to ACF plot, if a value goes outside the band for some k, we assume that the observation with lag k has a direct effect on the current observation.

## 5 Data and Method

The data was taken from Yahoo Finance.The data is a daily data recorded for 10 years from january 2010 to december 2019.There were a total of 2469 rows and five coloumns of entries without any missing values.The five coloumns contained data on opening price, closing price, lowest  highest price for the day, and adjusted closing price. For this project, we worked with the data on adjusted closing price.

Before fitting of any time series model, we began our analysis by calculating some descriptives of the data and visualizing the plot for presence of any systematic patterns. Table 1 summarizes the results obtained from this descriptive analysis and Figure 1 shows the plot of the original data.

| Count | Mean | Std.Dev | Min | 1st Quartile | Median | 3rd Quartile | Max |
|---|---|---|---|---|---|---|---|
| 2444.00 | 7751.13 | 2207.52 | 4544.20 | 5677.44 | 7784.02 | 9654.51 | 12271.80 |

Table 1:Descriptive Statistics

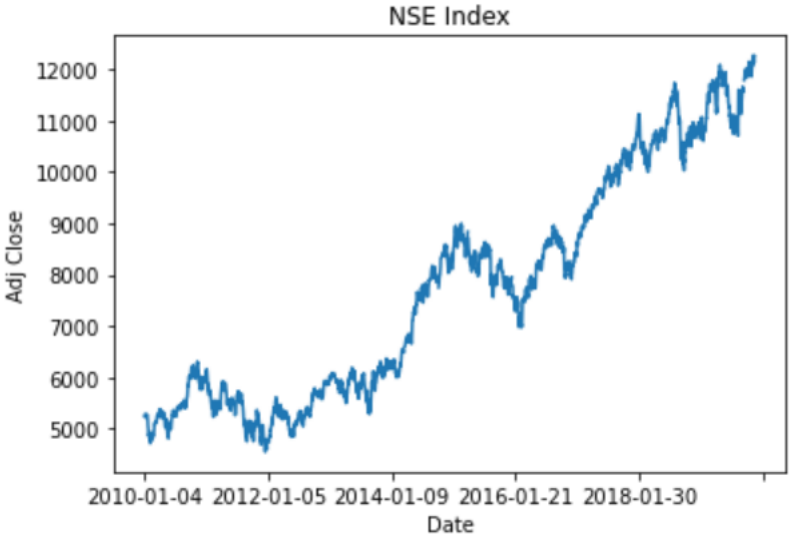


Figure 1: Original Data

Clearly, one can see that there is an increasing trend in the data. So, we applied a non-parametric test for a formal diagnostic of this trend. In particular, we used the Relative ordering test (see Section 2.1 for details) and found that there was a significant trend present. The value of test-statistic came out be 59.1616 which is quite off from the critical z values. Although this was also very obvious from the graph, one does needs a support of sophisticated statistical evidence for making any decision.

Then in order to remove trend, we applied a differencing operator of lag 4 on our data. The resultant series $Z_t$ was obtained by the following relation, $Z_t = X_t$-$X_{t-1}$-$X_{t-2}$-$X_{t-3}$,where $X_t$ are the values of the original series. The intuition behind choosing the order of differencing to be 4 was that the data is a stock price data and often the stock prices are assumed to be dependent on the previous values.As we can conclude from the Figure 2 there is no upward or downward trend here. One may notice that the trend is significantly

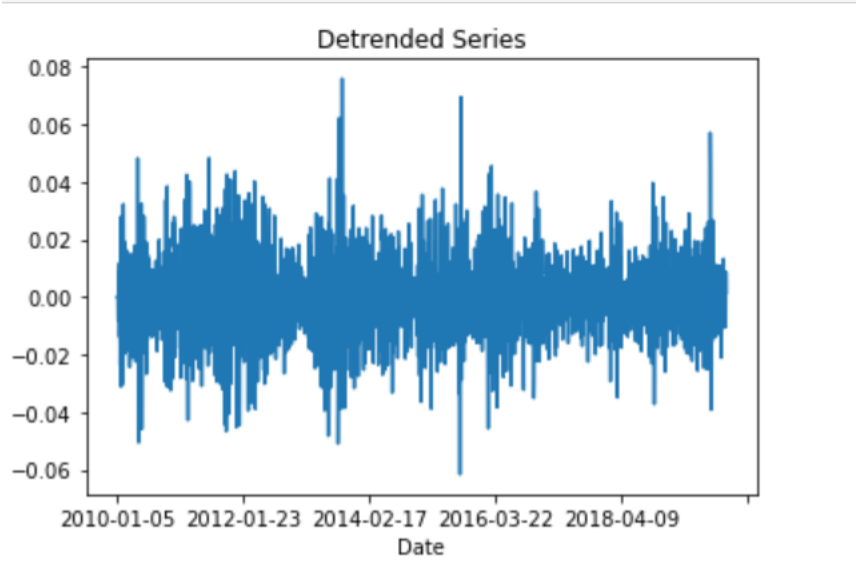


Figure 2: Detrended Data

removed. Nonetheless, we again applied the Relative ordering test of Section 2.1 to back our conclusion with statistical evidence. The value of the test-statistic was obtained to be equal to -0.2112 with a pvalue of 0.4164 . Hence, we concluded that the null hypothesis is true and that the trend is indeed removed from the data.

Looking at the graph, we then hypothesized that our detrended series is purely random and free from any deterministic fluctuations. For testing this hypothesis, we applied the Turning point test of Section 2.2. The test-statistic for the same came out to be -0.2573 with a pvalue of 0.3985. Again, the evidence from the data was insufficient and we failed to reject the Null hypothesis that the series is purely random. So now we have a series which has no deterministic components in it. A natural way to proceed will be to fit standard time series models on it and see which gives the best approximation.

However, before doing that, we need to first verify that the series is stationary. For that, we applied the Dickey-Fuller Test from Section 2.3. As described in Section 2.3, the Null hypothesis for the Dickey-Fuller test is that the series is non-stationary. We used an in-built function from a python library to carry out this test.

The test statistic was obtained to be equal to -15.939155 against the critical value of 0.000000 at 5 % level of significance. Hence, we rejected the null hypothesis that the series is non-stationary and proceeded to fitting different models to this series.

Now there exists a large number of models which can be fitted to any given time series data. However, a good way to proceed is to identify a small number of candidate models and then differentiate them using various model checking criterion. Towards that, we plotted the ACF and PACF plots for our detrended data (see Figure 3). We described

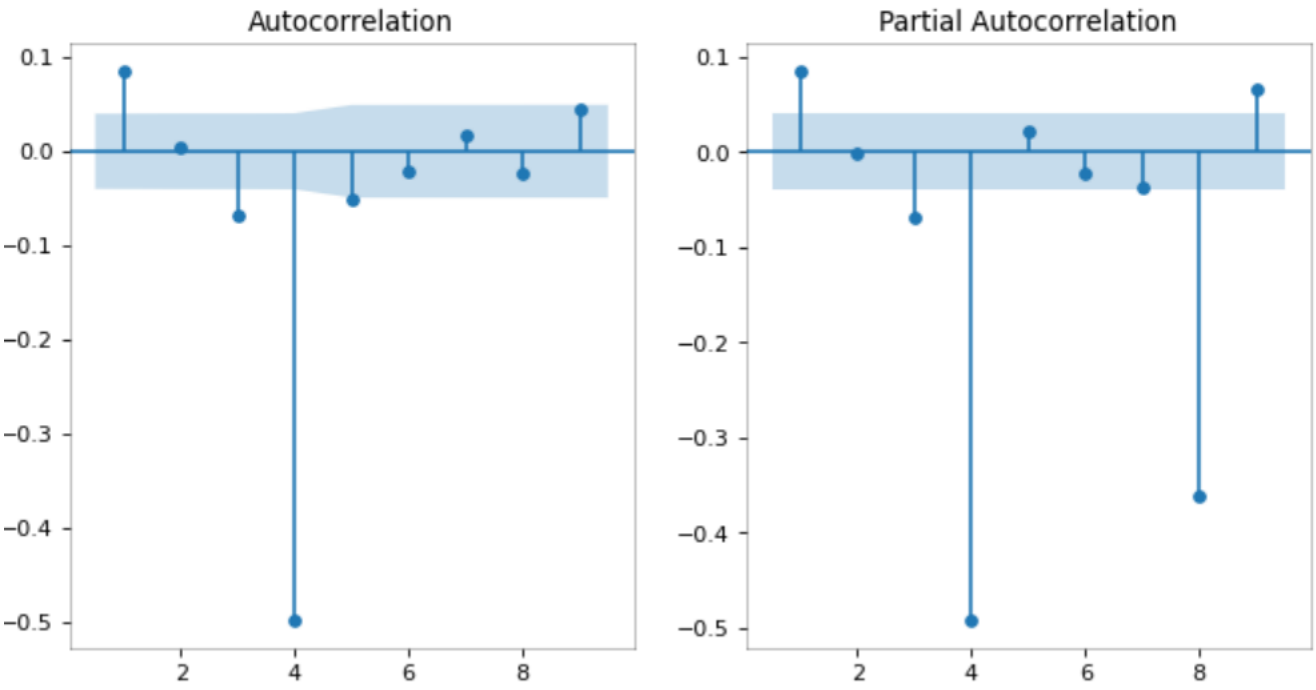


Figure 3: ACF and PACF Plots of the data

in section 3 a method to identify the model parameters for a time-series data. In both the plots we will look for the first lag for which the acf/pacf value lies outside the insignificant band. The first such lag from the ACF plot will be considered for the MA parameter while that from the PACF plot will be considered for the AR parameter.

From Figure 3, we can see that in both the plots, the first lag at which the value goes outside the insignificant band is equal to 3. Hence we will consider all the models with p and q both less than or equal to 3.

This gives us a total of 16 different models. From these 16 models, we will then identify the best model by the criterion of minimizing AIC and BIC. Table 2 summarizes AIC values for these 16 models.

| p\q | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | -13819.59428792 | -13835.06585854 | -13831.78882356 | -14032.979723 |
| 1 | -13835.19856245 | -13833.20138921 | -13873.9234912 | -13945.136341 |
| 2 | -13833.20935225 | -13831.19553248 | -13829.33890944 | -14461.765167 |
| 3 | -13842.69992189 | -14257.63853955 | -14208.20065406 | -14279.464593 |

Table 2: AIC for different values of p and q

From this table, we conclude that the best model is ARMA(2, 3) since it gives us the lowest AIC.
Using ARIMA(2,4,3) on the original data we forecast the 20 future values of the Nifty Fifty data set.

## 6 Results Discussions

From our discussion in the previous section, we saw that ARMA(2, 3) fits best to the de-trended series of our data. For assessing the validity of this model, we further fitted the ARIMA(2, 4, 3) model to the original time series data. A plot of the fitted values superimposed over original values is shown in Figure 4.

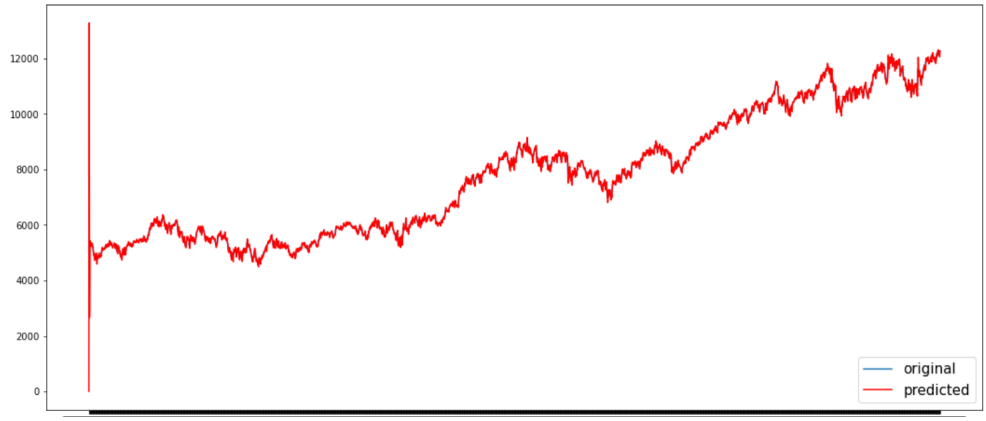


Figure 4: Fitted values over Original Data

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -4.237e-06 | 2.11e-05 | -0.201 | 0.841 | -4.56e-05 | 3.72e-05 |
| **ar.L1** | 0.9244 | 0.032 | 28.881 | 0.000 | 0.862 | 0.987 |
| **ar.L2** | -0.2973 | 0.031 | -9.645 | 0.000 | -0.358 | -0.237 |
| **ma.L1** | -0.9898 | 0.031 | -31.599 | 0.000 | -1.051 | -0.928 |
| **ma.L2** | 0.5482 | 0.033 | 16.481 | 0.000 | 0.483 | 0.613 |
| **ma.L3** | -0.5258 | 0.016 | -33.318 | 0.000 | -0.557 | -0.495 |
| **sigma2** | 0.0001 | 3.17e-06 | 44.005 | 0.000 | 0.000 | 0.000 |

Figure 5: Summary of Model Fit

Clearly, we can see that the fitted values match perfectly with the original data. Figure 5 shows summary of the in-built fit function of a python library. From the table, we see that all the coefficients of the model are significant at 5% level of significance. If the original series is $X_t$, the final model equation is written as follows,

$$Z_t = \mu + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \epsilon_t$$

$$\mu = -4.237e - 06$$

$$\phi_1 = 0.9244$$

$$\phi_2 = -0.2973$$

$$\theta_1 = -0.9898$$

$$\theta_2 = 0.5482$$

$$\theta_3 = -0.5258$$

where $Z_t = X_t$-$X_{t-1}$-$X_{t-2}$-$X_{t-3}$ and $\epsilon_t \sim N(0, \sigma^2); \sigma^2 = 0.0001$ for all t.

After forecasting the future 20 values, we compared them to the actual 20 test values. The plot for the forecasted and the actual values is shown in Figure 6.

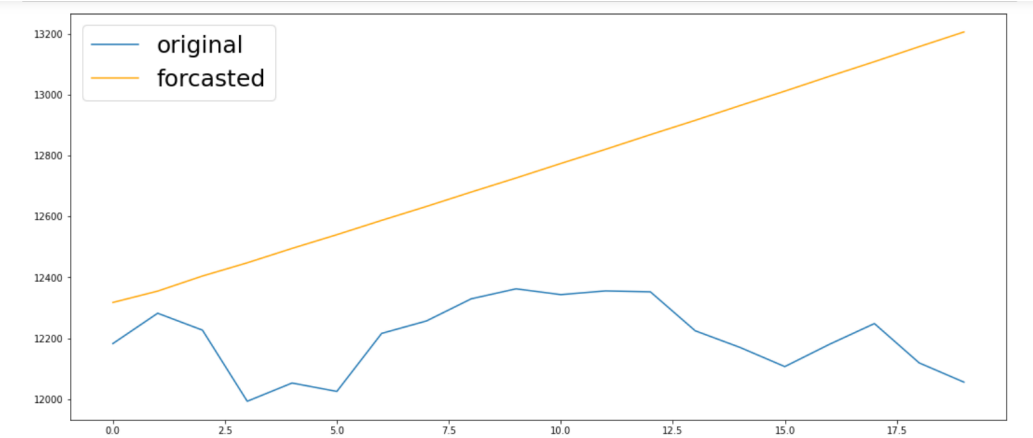


Figure 6: Original and forecasted values

In order to determine the accuracy of our forecasted values, we have calculated Mean Absolute Percentage Error (MAPE) which came out to be 4.15 . Since the MAPE value is very low so we can consider our ARIMA(2,4,3) to be a good model for forecasting Indian Market Index - Nifty Fifty.
The link for the all the files used can be found here
https://drive.google.com/drive/folders/1Gi4_Ne28I4e8LWfcG8rBZ_wZub9xUr-2?usp=sharing