Comparing standard Time Series models on Google stock prices data[*]

Abhilash Singh[†] Avik Biswas[‡] Nitika Verma[§] Rasamanjari Nandan[¶] Sanket Agrawal[‖]

M.Sc. Statistics, Second year
Department of Mathematics & Statistics, Indian Institute of Technology - Kanpur

Date of Submission: November 30, 2020

Abstract

This report highlights the working of different Time Series models on the google stock price data. The aim of the project is to identify a model best fitting the google stock price data and to forecast the stock price of google. The data from 2012 to 2016, which have been documented daily , reveal that the ARMA(2,3) model may fit most adequately

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 Google Stock price data. 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),

[*] As part of the course MTH517A - Times Series Analysis

[†] 191003

[‡] 191029

[§] 191085

[¶] 191111

[‖] 191124

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 google stock price data, 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 1 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 $\{X_1,X_2,\ldots X_n\}$
Define

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

$$Q = \sum_{\substack{i \\ i<j}} \sum_{j} q_{ij}$$

Where Q counts the number of decreasing point in the time series and is also the number of discordance
NUll Hypothesis H_0: There is no trend in the time series
against the Alternate Hypothesis H_1: There is a trend in the time series
Under the null hypothesis $E(Q) = \sum_{\substack{i \\ i<j}} \sum_j E(q_{ij}) = \frac{n(n-1)}{4}$.If observed the Q≪E(Q) then it would be an indication of rising trend and if observed Q≫E(Q) the it would be the indication of falling trend. If the observed Q does not differ "significantly" from E(Q)(Under the null hypothesis) then it would indicate no trend. Q is related to the kindall's τ the rank correlation coefficient through the relationship

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

Under the null hypothesis $E(\tau) = 0$ and $Var(\tau) = \frac{2(2n+5)}{9n(n-1)}$
Test Statistics: $Z = \frac{\tau - E(\tau)}{\sqrt{V(\tau)}}$ fallow normal N(0,1) asymptotically (under Null hypothesis)
We would reject the null hypothesis of no trend at level of significant α if observed $|z| > \tau_{\alpha/2}$ where $\tau_{\alpha/2}$ is the $\alpha/2^{th}$ upper cutoff points of a standard normal distribution

2.2 Turning Point Test

Turning point test is a non-parametric test which is used for testing randomness of the time series data set. Let X_1,X_2,X_3,$\dots X_n$ be the data ,then X_i is consider as a turning point if either $X_{i-1} < X_i$ and $X_i > X_{i+1}$ or $X_{i-1} > X_i$ and $X_i < X_{i+1}$.We count the number of turning points in the data.
Define

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

Suppose P is the total number of turning points i.e $P = \sum_{n=2}^{n-1} U_i$
Null Hypothesis H_0: Series is truly random (Does not contain any deterministic components.)
Against the Alternative against H_1: Series is not truly random.
Under the null hypothesis $E(P) = \frac{2(n-2)}{3}$ and $Var(p) = \frac{16n-29}{90}$
Test statistics:

$$Z = \frac{P - E(P)}{\sqrt{\frac{16n-29}{90}}} = \frac{P - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}} \text{ follows } N(0,1) \text{ asymptotically}$$

We would reject null hypothesis Ho at level of significance α if observed $|Z| > \tau_{\alpha/2}$
Where $\tau_{\alpha/2}$ is upper $\alpha/2$ cutoff point of N(0,1).

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

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t$$

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

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

The regression model can be written as

$$Y_t - Y_{t-1} = \phi_0 + (\phi_1 - 1)Y_{t-1} + \epsilon_t$$

$$\triangle Y_t = \phi_0 + \delta Y_{t-1} + \epsilon_t$$

We can calculate the t-statistics on estimated value of δ, 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

2.4 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

$$X_t = c + \sum_1^p \phi_i X_{t-i} + \epsilon_t$$

where $\phi_1, \phi_2, ..., \phi_p$ are parameters with $\phi_p \neq 0$, c is a constant and ϵ_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 $\theta_q \neq 0$, μ is the expectation of X_t (can be assumed to be zero) and $\epsilon_t, \epsilon t - 1, ...$ 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.

2.5 ACF Plot 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.

3 Data & Method

The data was collected from Kaggle. It is available via the following url: `https://www.kaggle.com/medharawat/google-stock-price`. The data was a daily data on Google stock prices recorded for 4 years from January 3, 2012 to December 30, 2016. There were a total of 1258 rows and 5 columns of entries without any missing values. The 5 columns contained data on opening price, closing price, lowest & highest price for the day, and total volume. For this project, we worked with the data on opening 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.	Minimum	1st Quartile	Median	Third Quartile	Maximum
1258	533.71	151.90	279.12	404.12	537.47	654.92	816.68

Table 1: Descriptive Statistics (Std. Dev. means standard deviation)

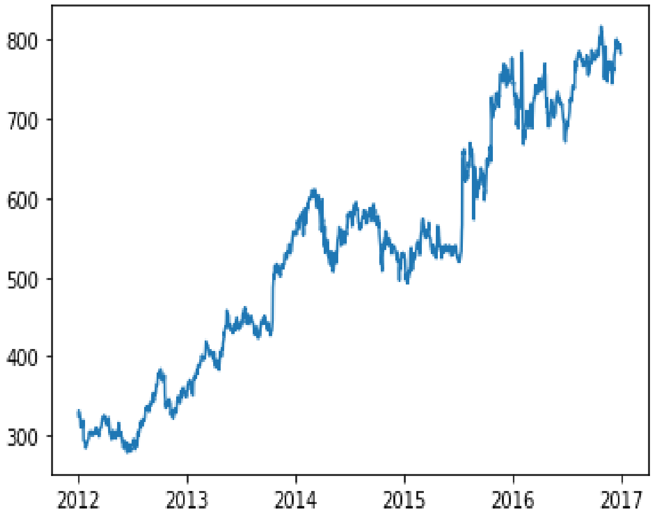


Figure 1: Original Data

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 41.95 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.

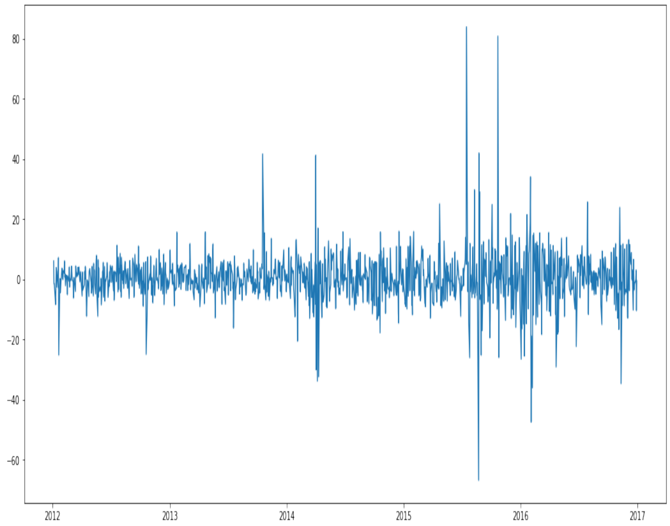


Figure 2: Detrended series after first order deferencing

Then in order to remove trend, we applied a differencing operator of lag 1 on our data. The resultant series Z_t was obtained by the following relation,

$$Z_t = Y_t - Y_{t-1},$$

where Y_t are the values of the original series. The intuition behind choosing the order of differencing to be 1 was that the data is a stock price data and often the stock prices are assumed to be dependent on the immediate previous value. Figure 2 shows the series obtained after differencing. One may notice that the trend is significantly 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 0.2976 with a $p-$value of 0.383. 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 0.2901 with a $p-$value of 0.386. 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 -17.612 against the critical value of -2.863 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

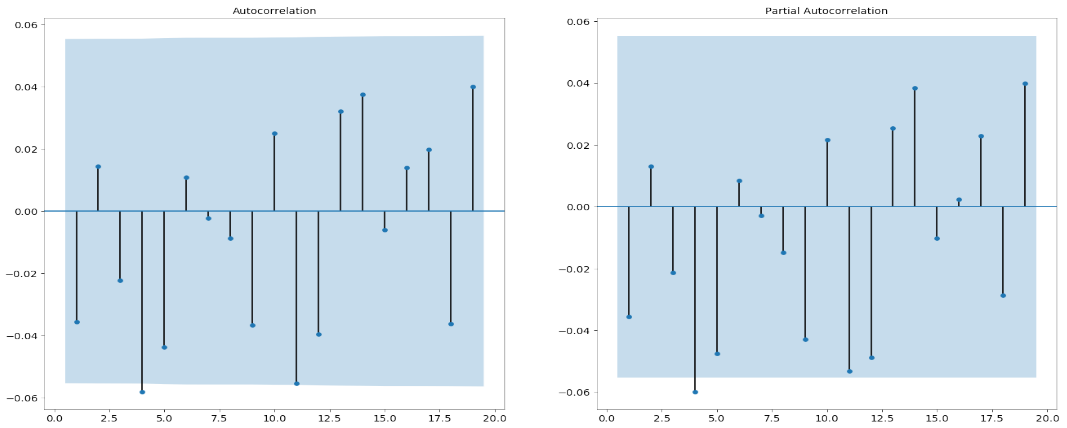


Figure 3: ACF(left) and PACF(right) plots of the data

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 in section 2.5 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 4. Hence we will consider all the models with p and q both less than or equal to 4. This gives us a total of 25 different models. From these 25 models, we will then identify the best model by the criterion of minimizing AIC and BIC. Table 3 summarizes AIC values for these 25 models.

$p\backslash q$	0	1	2	3	4
0	8923.483	8923.930	8925.703	8926.904	8923.852
1	8923.890	8925.630	8919.574	8923.444	8923.637
2	8925.676	8927.607	8921.038	8914.709	8925.623
3	8927.106	8923.546	8924.988	8918.410	8922.383
4	8924.607	8924.277	8926.241	8921.056	8928.230

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(2,1,3) on the original data we forecast the 20 future values of the google stock price .

4 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, 1, 3) model to the original time series data. A plot of the fitted values superimposed over original values is shown in Figure 4.

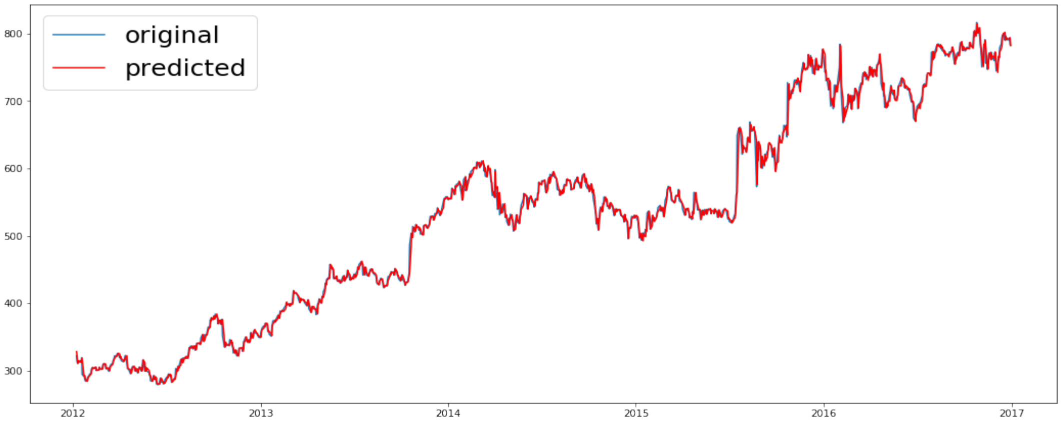
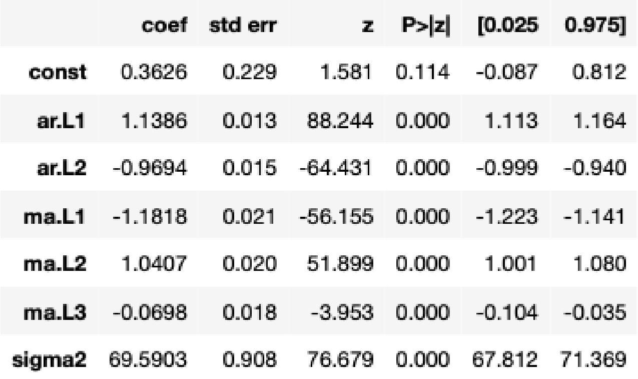


Figure 4: Fitted values over Original data



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3626	0.229	1.581	0.114	-0.087	0.812
ar.L1	1.1386	0.013	88.244	0.000	1.113	1.164
ar.L2	-0.9694	0.015	-64.431	0.000	-0.999	-0.940
ma.L1	-1.1818	0.021	-56.155	0.000	-1.223	-1.141
ma.L2	1.0407	0.020	51.899	0.000	1.001	1.080
ma.L3	-0.0698	0.018	-3.953	0.000	-0.104	-0.035
sigma2	69.5903	0.908	76.679	0.000	67.812	71.369

Figure 5: Summary of model fit

Clearly, we can see that the fitted values match perfectly with the original data. The AIC of this model came out to be around 8915. 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,

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

$$\begin{aligned}
\mu &= 0.3626 \\
\phi_1 &= 1.1386, \\
\phi_2 &= -0.9694, \\
\theta_1 &= -1.1818, \\
\theta_2 &= 1.0407, \\
\theta_3 &= -0.0698,
\end{aligned}$$

where $Z_t = X_t - X_{t-1}$ and $\epsilon_t \sim N(0, \sigma^2); \sigma^2 = 69.5903$ 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.

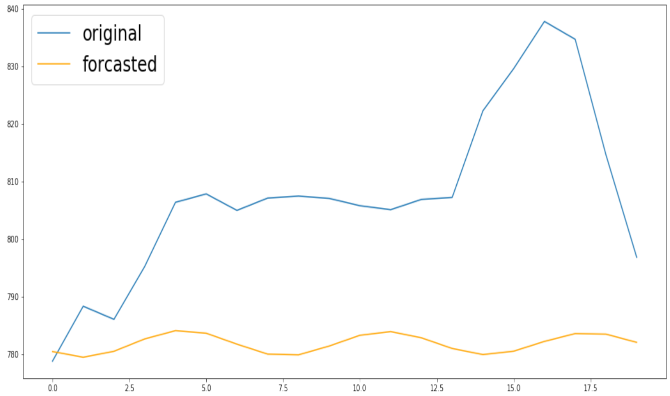


Figure 6: Original and Forecasted values

	Original	Forcasted
1	778.81	780.492154
2	788.36	779.496095
3	786.08	780.533173
4	795.26	782.679152
5	806.4	784.116241
6	807.86	783.671485
7	805	781.772113
8	807.14	781.020353
9	807.48	779.960741
10	807.08	780.549959
11	805.81	782.247812
12	805.12	783.953276
13	806.91	782.872175
14	807.25	781.020353
15	822.3	779.960741
16	829.62	780.549959
17	837.81	782.247812
18	834.71	783.609032
19	814.66	783.512358
20	796.86	782.082707

Table 3: 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 3.1681. Since the MAPE value is very low so we can consider our ARIMA(2,1,3) to be a good model for forecasting google stock prices.

5 Conclusions

To conclude, we found that ARMA(2, 3) fits best on the series obtained after first order differencing. That is, ARIMA(2, 1, 3) fits best for the original data on Google Stock prices. The model equation is given by (1). We employed the standard approach of first eliminating the trend from the data. We used first order differencing for this. After eliminating the trend, we found that the resultant series was purely random and stationary. We plotted the ACF and PACF plots for this resultant series and identified the candidate models for this data. Finally, we fitted all the models and based on the criterion of minimizing AIC, we found the best model. As can be seen from the trace plot in Figure 4, the best model was able to fit the data quite well.

Acknowledgement: We take this opportunity to thank Prof. Amit Mitra for providing us the wonderful opportunity of working on this interesting project as part of the course MTH517A - Time Series Analysis.