Price Forecasting and Analysis on the stocks of Exchange Traded Funds

Anupam SInha
MSBA (Fall-2018)
UIC

Abstract:

The time series methods and its various tool have played a vital role in exploring the data from different ware houses. Using data mining tools and analytical technologies we do a quantifiable amount of research to explore new approach for the investment decisions .The market with huge volume of investor with good enough knowledge have a prediction as well as control over their investments. The stock market some time fails to attract new investors. The reason states people are unaware and do not want to fall into the risk. An approach with adequate expertise is designed to help investors to ascertain veiled patterns from the historic data that have feasible predictive ability in their investment decisions. In this paper the Dow Jones Industrial Average dataset from Kaggle has been selected for analysis. The historical data has a significant role in, helping the investing people to get an overview about the market behavior during the past decade. The stationarity of the data has been formally confirmed using Augmented Dickey Fuller Test while the stock data has been collected and trained using ARIMA model with different parameters. The performance of the trained model is analyzed and it also tested to find the trend and the market behavior for future forecast.

Keywords: ARIMA, ACF, PACF, ADF, MA, AR

I. INTRODUCTION

An ETF, or exchange-traded fund, is a marketable security that tracks an index, a commodity, bonds, or a basket of assets like an index fund. Unlike mutual funds, an ETF trades like a common stock on a stock exchange. ETFs experience price changes throughout the day as they are bought and sold. ETFs typically have higher daily liquidity and lower fees than mutual fund shares, making them an attractive alternative for individual investors.

An ETF holds assets such as stocks, commodities, or bonds and generally operates with an arbitrage mechanism designed to keep its trading close to its net asset value, although deviations can occasionally occur.
Most ETFs track an index, such as a stock index or bond index. ETFs may be attractive as investments because of their low costs, tax efficiency, and stock-like features.

In this paper we focus on the real world problem in the stock market. The seasonal trend and flow is the highlight of the stock market. Eventually investors as well the stock broking company will also observe and capture the variations, constant growth of the index. This will aid new investor as well as existing people will make a strategic decision. It can be achieved by experience and the constant observations by the investors. In order to overcome the above said issues, we have suggested ARIMA algorithm in three steps,
Step 1: Model identification
Step 2: Model estimation
Step 3: Forecasting

II. PROBLEM DEFINITION

A. Problem of Stock forecasts

The problems of stock forecasts are indispensable crops up from time to time. In all, the inevitable finish is that no issue what type of endeavor you are in, or what task you perform, there is a need for some kind of future estimate upon which to make a chart. Marketing society need forecasts to conclude to either enter or exit the business

B. Financial Planning

Finance professionals use forecasts to make financial plans. Investors invest their hard earned capital in stocks with the expectation of gaining from their investment through a positive payoff [1]. Since having an excellent knowledge about share price movement in the future serves the significance of fiscal professionals and investors. This familiarity about the future boosts their confidence by way of consulting and investing. But these movements predict the share prices without proper forecasting methods, only for the interest of the financial professional and investors. There are many forecasting methods in projecting price movement of stocks such as the Box Jenkins method.

III. MODELS AND METHODS

ETFs are baskets of securities designed to track the performance of an index. They are designed to provide exposure to broad-based indexes at a lower cost. We first analyzed why ETF should be the choice for an investment. We provide a brief history of this segment, key attributes of ETFs, and investments strategies and implementations with ETFs. The data analysis and the forecast evaluation is to determine the best forecasting model for a single ETF (WCM/BNY Mellon Focused Growth ADR ETF). This study seeks to investigate [2] the best forecasting method under consideration and gives the minimum forecasting error. The objectives are (a) Model specification (or model identification); (b) Model fitting (or model estimation); (c) Model checking (or model verification or model criticism).

Based on the evaluation of a decade of past historical data, we provide a guidance for the price of our ETF (WCM/BNY Mellon Focused Growth ADR ETF) using the ARIMA model, which produced promising results (with low forecast errors of 1% across several forecast metrics)

A. Dataset Description

1. Date: This captures all the Opening, Closing, High and Low ETF that are captured in a day.

2. Open: Open is the price of the ETF at the beginning of the trading day (it need not be the closing price of the previous trading day).
3. High: High is the highest price of the ETF on that trading day.
4. Low: Low the lowest price of the ETF on that trading day.
5. Close: Close the price of the ETF at closing time.
6. Adj_Close: Adjusted close is the closing price of the ETF that adjusts the price of the ETF for corporate actions.
7. Volume: Volume is the number of shares or contracts traded in a security or an entire market during a given period. For every buyer, there is a seller, and each transaction contributes to the count of total volume. That is, when buyers and sellers agree to make a transaction at a certain price, it is considered one transaction. If only five transactions occur in a day, the volume for the day is five.

We were divided on the source of our data and the after a lot of deliberation we came to the conclusion that Kaggle is one of the best and trusted places from where we can get our data. It is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and other users

B. Modelling: Moving Averages

- A simple moving average (SMA) [3] is the simplest type of technique of forecasting. Basically, a simple moving average is calculated by adding up the last 'n' period's values and then dividing that number by 'n'. So the moving average value is considering as the forecast for next period.
- Moving averages can be used to quickly identify whether selling is moving in an uptrend or a downtrend depending on the pattern captured by the moving average.

C. Autoregressive Integrated Moving Average (ARIMA)

- A statistical technique that uses time series data to predict future. ARIMA modeling will take care of trends, seasonality, cycles, errors and non-stationary aspects of a data set when making forecasts.
- ARIMA checks stationarity availability in the data, the data should also show a constant variance in its fluctuations over time. To get the proper information about the parameter used in ARIMA is based on "identification process" which was purposed by Box-Jenkins
- ARIMA is mainly used to project future values using historical time series data. Its main application is in short forecasting with minimum 38-40 historical data points with minimum number of outliers.
- As ARIMA models are applied in cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model can be applied one or more times to eliminate the non-stationarity as we had to do with our dataset
- The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.
- Non-seasonal ARIMA models are generally denoted ARIMA(p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. Seasonal ARIMA models are usually denoted ARIMA(p,d,q)(P,D,Q)m, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.
- When two out of the three terms are zeros, the model may be referred to be based on the non-zero parameter, dropping "AR", "I" or "MA" from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

D. Augmented Dickey Fuller Test

- In statistics and econometrics, an augmented Dickey Fuller test (ADF) [4] tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.
- ARIMA model needs the series to be stationary i.e. the mean, Variance and Auto-covariance should be independent on time. Hence, we will be using Augmented Dickey-Fuller (ADF) test to formally test the stationarity.

IV. PROPOSED APPROACH

The Box-Jenkins methodology is a five-step process for identifying, selecting, and assessing conditional mean models (for discrete, Univariate time series data) [5]

A. Phase 1: Data Preparation and Model Selection

• Transform the data to stabilize the attributes.
• Find the difference if it is not stationary; successively difference
• Series to attain stationary
• Examine data, plot ACF and PACF to identify potential models

B. *Phase 2: Estimation/Testing and Diagnostics*

- Estimate parameters in potential models
- Select best model
- Check ACF/PACF of residuals
- Test the residuals
- Are the residuals are white noise

C. *Phase 3: Forecast the Application*

Forecasting the trend. This model is used to forecast the future

D. *Software Used: R*

1. We used R-Studio for our analysis and forecasting modeling, and for cleaning data. Some of the characteristics of R that we thought we would use to our advantage were:

- "ts" is the basic class for regularly spaced time series using numeric time stamps.
- The zoo package provides infrastructure for regularly and irregularly spaced time series using arbitrary classes for the time stamps (i.e., allowing all classes from the previous section). It is designed to be as consistent as possible with "ts".
- The package xts is based on zoo and provides uniform handling of R's different time-based data classes.
- Various packages implement irregular time series based on "POSIXct" time stamps, intended especially for financial applications. These include "irts" from tseries, and "fts" from fts.
- The class "timeSeries" in timeSeries implements time series with "timeDate" time stamps.
- The class "tis" in tis implements time series with "ti" time stamps.
- The package tframe contains infrastructure for setting time frames in different formats.

Forecasting and Univariate Modeling

- The forecast package provides a class and methods for univariate time series forecasts, and provides many functions implementing different forecasting models.
- Autoregressive models: ar() in stats (with model selection) and FitAR for subset AR models.
- ARIMA models: arima() in stats is the basic function for ARIMA, SARIMA, ARIMAX, and subset ARIMA models. It is enhanced in the forecast package via the function arima() along with auto.arima() for automatic order selection. arima() in the tseries package provides different algorithms for ARMA and subset ARMA models. Other estimation methods including the innovations algorithm are provided by itsmr. FitARMA implements a fast MLE algorithm for ARMA models.

E. *Analysis*

Pre-processing of data

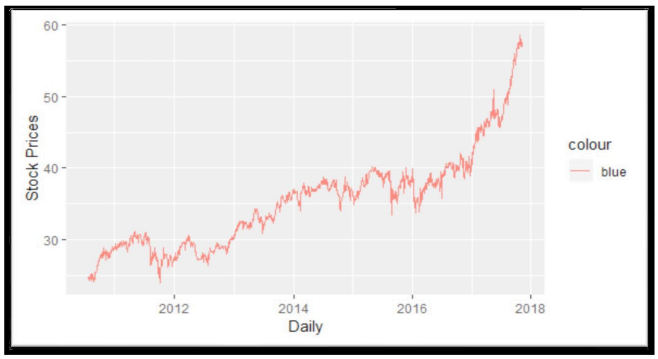
1. Reading Time Series Data

We read the file using read.csv. The data set has these variables and looks like this in general:

```
        Date   Open   High    Low  Close Volume OpenInt
1 2010-07-21 24.333 24.333 23.946 23.946  43321       0
2 2010-07-22 24.644 24.644 24.362 24.487  18031       0
3 2010-07-23 24.759 24.759 24.314 24.507   8897       0
4 2010-07-26 24.624 24.624 24.449 24.595  19443       0
5 2010-07-27 24.477 24.517 24.431 24.517   8456       0
6 2010-07-28 24.477 24.517 24.352 24.431   4967       0
>
```

2. Plotting Time Series

After loading time series data into R, we plot the data using ggplot2.



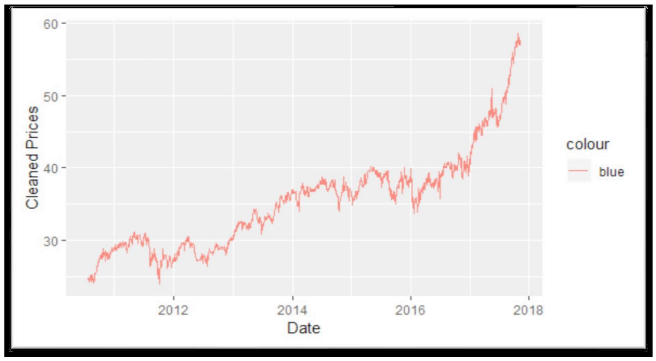
We observed that our dataset has a linear increase in price along with time.

#Removing outliers from the data

count_ts = ts(daily[,c('Open')])

daily$clean_open = tsclean(count_ts)

We counted all the Open Price values and used tsclean() to identify and replace outliers and missing values in our time series data. Here is the plot of Clean Data



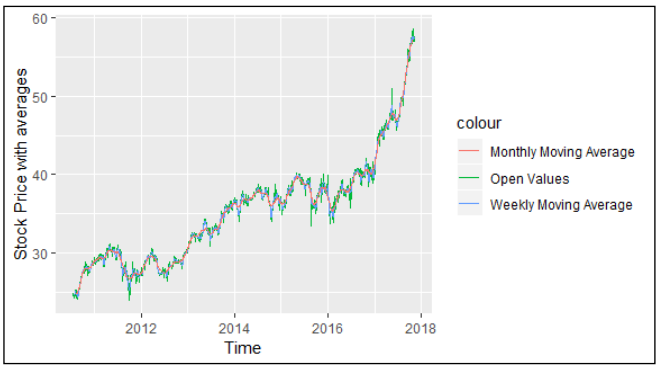
After removing the outliers, we observe that our daily data is pretty stationery. Visually we could draw a line through it i.e. the line would contain average points across several time periods, called the Moving Averages thereby smoothing the observed data into a more stable predictable series. The moving average is extremely useful for forecasting long-term trends. We calculated the Weekly and Monthly Moving Averages.

daily$Open_ma7 = ma(daily$Open,order=7) #Weekly Moving Average

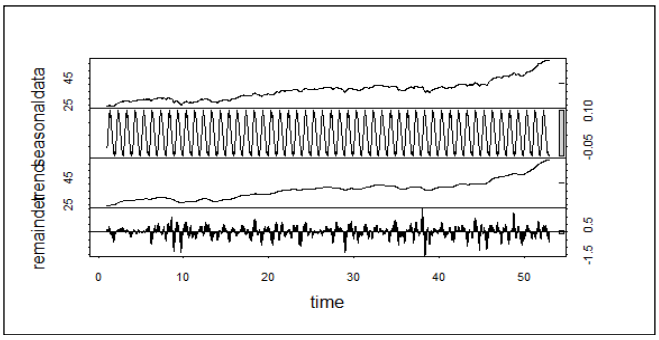
daily$Open_ma30 = ma(daily$Open,order=30) #Monthly Moving Average

3. Adding Moving averages

Here, we calculate the weekly and monthly moving averages. The plot of these moving averages along with the daily open value is plotted as shown in the below chart.



3. Decomposing Time Series



V. COMPUTATIONAL RESULTS

A. *Test Statistics of Open Values*

adf.test(daily$Open, alternative = "stationary")

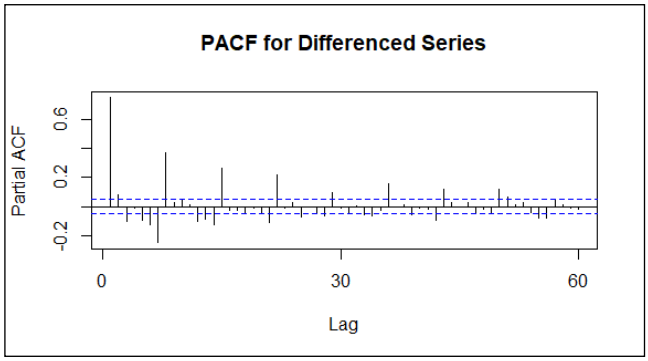
```
        Augmented Dickey-Fuller Test

data:  daily$Open
Dickey-Fuller = -0.62348, Lag order = 11, p-value = 0.9762
alternative hypothesis: stationary
```

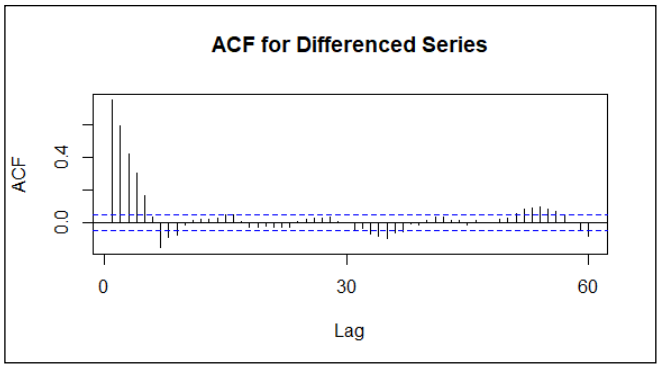
Since, the P-value is huge, we cannot reject the null hypothesis i.e. weekly moving average is not stationary. Hence, we will check for the stationarity for the difference of weekly moving average. We removed the seasonal component by using seasadj() of the decomposed data and performed the ADF test on this decomposed data.

B. *Test Statistics of Differenced Weekly Moving Averages*

```
        Augmented Dickey-Fuller Test

data:  count_d1
Dickey-Fuller = -9.8365, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary
```

Now we observe that the P-value is less than the alpha values, so we reject the null hypothesis, i.e. the difference of weekly moving average is stationary. Hence, we will go ahead with the differenced data for building the ARIMA model. Below are the plots for ACF and PACF for difference of weekly moving averages which supports the ADF test results.





Here we can observe that most of the significant values are within the confidence intervals.

C. *ARIMA Modelling*

ARIMA is combination of Auto Regression and Moving averages. We try to fit an ARIMA model with c (1,1,1) for which below are the statistics

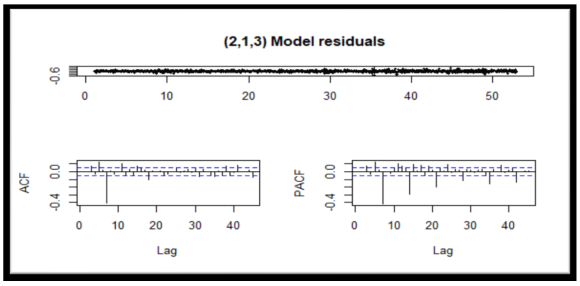
Test Statistics for ARIMA Model order (1,1,1)

```
> summary(try)

Call:
arima(x = deseasonal_cnt, order = c(1, 1, 1))

Coefficients:
          ar1      ma1
      0.7889  -0.0852
s.e.  0.0196   0.0295

sigma^2 estimated as 0.008279:  log likelihood = 1523.38,  aic = -3040.76
```

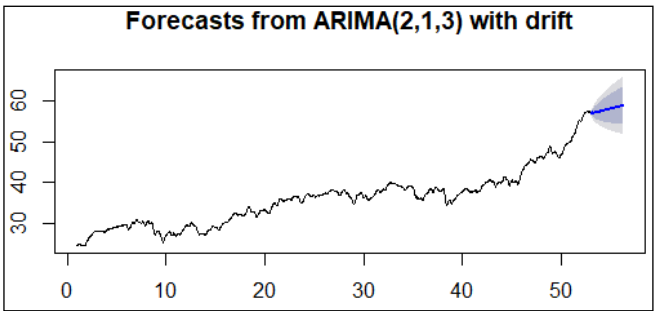
We used auto.arima()that gave us the best model according to sigma^2 value as shown. Hence, ARIMA(2,1,3) with drift is our best model.

```
Series: deseasonal_cnt
ARIMA(2,1,3) with drift

Coefficients:
          ar1     ar2     ma1     ma2     ma3   drift
      -0.1744   0.705  0.8936  0.0367  0.0037  0.0209
s.e.   0.0379   0.029  0.0428  0.0516  0.0372  0.0093

sigma^2 estimated as 0.008031:  log likelihood=1550.05
AIC=-3086.11   AICc=-3086.04   BIC=-3048.65
```

We went ahead and plotted the residuals plot for this.



Forecasting Values: We used the Forecast package to predict the next 100 values for this and this is shown in the graph in the blue line within confidence bounds.



VI. Conclusion

• If the sign of the forecasted return equals the sign of the actual returns we have assigned it a positive accuracy score.

• The accuracy percentage of the model comes to around 68%.

• We can try running the model for other possible combinations of (p,d,q) or instead use the auto.arima function which selects the best optimal parameters to run the model.

VII. Future Scope

Due to time constraints we kept our analysis limited to Open Price values. For better forecasting experience we could also repeat the entire procedure for close, low and high values, observe the trends and results and make wise investment decisions.

VIII. References

[1] Kofi agyarko ababio, June 2012, "Comparative study of stock price forecasting using arima and arimax models". J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Chris Chatfield, "TIME-SERIES FORECASTING"

[3] https://www.bistasolutions.com/resources/blogs/5-statistical-methods-for-forecasting-quantitative-time-series/.

[4] https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test.

[5] Pankratz, Alan New York: John Wiley & Sons, (1983), "Forecasting with Univariate Box–Jenkins models: concepts and cases".M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.