Stock Market Analysis and Time Series Predictions

Vignesh Kumar S
Department of Chemical Engineering
Indian Institute of Technology, Madras
Email address: ch18b118@smail.iitm.ac.in

***Abstract*—The purpose of this article is to analyse a new type of dataset, called the Time Series where the order of data plays a key role using techniques covered through the coursework and it's extension. In this work, we demonstrate the application of the learnings illustrated in the world of finance through Stock Prices of different companies. Due to the complex nature and the abstractness involved in Stock Market, it is often necessary to forego a detailed analysis before investments to make informed choices. Using the concepts learned in the course, we try obtaining insights and analysing different stocks in a large amount of data comprehensively. We then try to forecast our predictions that could help in analysing risks and benefits of an investment.**

I. Introduction

In this paper, we will study Time Series Analysis, a technique used to analyze data where sequence is of importance, usually data involving date and time. We will use this technique to forecast a stock of interest that was analysed by weighing risks and returns. This analysis would aid in obtaining insights among a large dataset and hopefully uncover the trend behind a stock of interest.

The stock market allows numerous buyers and sellers of securities to meet, interact, and transact. Stock analysis is important for any investors and traders to make buying and selling decisions. By studying and evaluating past and current data, one could attempt to gain an edge in the markets by making informed decisions.

The dataset used for this problem comprises opening, high, low and closing prices of six stocks in India. We try to analyse each of those stocks to compare and contrast each of them. We then attempt to forecast the predictions of a stock that could suit a particular candidate using time series analysis and the techniques learnt in the course.

This work represents implementing the concepts learnt in the course and analysis of Time Series data in the finance world. This is purely written in academic interest and should not be used for personal decisions without professional advice.

II. Time Series Analysis

A time series is a sequence of data points that occur in successive order over some period of time. Since Time series data have a natural temporal ordering, this can be contrasted with cross-sectional data which captures a point-in-time and has no natural ordering of the observations. In particular, a time series allows one to see what factors influence certain variables from period to period.Time series analysis can be useful to see how a given asset, security, or economic variable changes over time. This helps in judging the future in the context of its past performance.

A data $v[1], v[2], .., v[k-1]$ is likely to be a time series if $v[k]$ depends on the past data. Though this gives an intuition that the memory is long term, i.e, $v[k]$, for any k, strongly depends on $v[1]$ too, this need not always be the case. Some times data could also depend on current and past residuals. The effect could also be dying, in other words, could get weaker and weaker with the more and more past data. An example of this would be weather, where the present weather is strongly dependent on yesterday's but weakly on the weather two months before.

A. Stationarity

A process is said to be stationary if the statistical properties of the process are invariant with time.

Strictly,

$$f(v[1], .., v[k]) = f(v[T+1], .., v[T+N]) \; \forall N, T \in \mathcal{Z}^{+}$$

where f is probability density function. Often strict stationary is difficult in practice, so a weaker sense of stationary could be defined as invariance upto second order moments. Mean of the process is independent of time and the process has finite variance and is independent of time. There are few additional conditions such as Auto-Covariance should be a function of only Time Difference (Lag) which we will discuss below.

Autocovariance is nothing but a covariance matrix of the data itself considering each sample instance. For example, consider two instances, $v[k]$ and $v[k+l]$, autocovariance $\sigma_{k,k+l}$ is $var(v[k], v[k+l])$. In order to calculate auto-covariance in strict sense, we would need the distribution of each $v[k]$. But when we assume weak stationarity condition, we consider $\sigma_{k,k+l}$ to be independent of sampling time k and hence this allows to compute the variance by considering all possible pairs with time difference(lag) l.

Since the above considers variance, it is also common and convenient to use correlation in the above place. This is hence known as Auto-Correlation Function.

B. Stationary Process

Any signal $v[k]$ can be expanded as $\hat{v}[k] + e[k]$, where $\hat{v}[k]$ is the predictable component and $e[k]$ is noise. In an ideal model, $e[k]$ should be unpredictable or in other words, uncorrelated to other sampling instants. So the ACF (Auto

Correlation Function) is expected to be 1 only at lag 0 (i.e with itself) and is 0 at any non-zero lag.

$$\rho_{ee}[l] = \begin{cases} 1 & \text{if l=0} \\ 0 & \text{if l} \neq 0 \end{cases}$$

1) MA Process: When the current prediction depends on the noise/unpredictable component of current instance and of the past instance, the process is said to follow MA (Moving Average) Process assuming (weak) stationarity is held true. An MA process of order 1 depends on the noise component of previous instance, while on order M depends on M such instances.

$$\text{MA(M)} \quad v[k] = \sum_{i=1}^{M} c_i e[k-i] + e[k]$$

We have already seen that the ACF of $e[k]$, the noise component is 1 only at lag 0 and 0 otherwise. Here, since a $v[k]$ depends on past M noise components, the ACF of $v[k]$ zeroes after M instances. (i.e) $\rho[l] = 0$, $\forall l > M$ (abrupt zeroing of ACF after M instances).

2) AR Process: When the present data depends on purely on the past instance, the process when (weak) stationarity is obeyed, is said to be a AR (Auto-Regressive) Process.

An AR process of order 1 depends only on its immediate past, while an AR process of order P depends upto p instances.

$$\text{AR(P)} \quad v[k] = \sum_{j=1}^{P} d_j v[k-j] + e[k]$$

where $e[k]$ is noise/unpredictable component.

Unlike MA process,since it depends on past data, ACF no longer zeroes abruptly but rather decays exponentially. We have seen that ACF is nothing but correlation with the same variable at different sampling instants. In order to establish a similar measure, we try to take the conditional correlation so that link between $v[k]$ and $v[k-l]$ is broken. This is called Partial Auto-Correlation Function (PACF). PACF goes to zero after l=p instances and for a AR Process, the ACF decays exponentially.

The important aspect of this ACF and PACF graph is that they help up in understanding the type of process and the mathematical model behind it which is not known to us before hand in practice. By analysing the nature of ACF and PACF plots, it is possible to come up with a linear time series model.

3) ARMA Process: In practice, a time series model could be a combination of both MA and AR process , (i.e) current predictions could depend on both past data and past residuals. This is collectively called an ARMA process.

$$v[k] = \sum_{i=1}^{M} c_i e[k-i] + \sum_{j=1}^{P} d_j v[k-j] + e[k]$$

Here, both ACF and PACF plots decays exponentially and it's difficult to judge the order of the process without further analysis.

C. Non-Stationary Process

Often, the (weak) stationarity condition may seldom hold. Mean and Variance could change with time. An example of this would be $v[k] = v[k-1] + e[k] + c$, where $e[k]$ is noise. These type tend to accumulate with time and grow upwards. Hence they can not be classified under ARMA process. However, the difference $v[k] - v[k-1]$ is clearly a MA/AR process of order 0. Here, on taking the difference a single time, the process reduces to a stationary process and hence is said to Integrating of order 1. If D differences are needed to reduce a process to be stationary, the process is said to be integrating of order D.

1) ARIMA Process: Let us represent one-time differencing $v[k] - v[k-1]$ as $\nabla v[k]$

$$\nabla^D v[k] = w[k]$$

where w[k] is ARMA process of order (P,M) after D- such differencing. Such v[k]'s are said to in ARIMA of order (P,D,M) . In other words, if repeated D differences reduce a sample to an ARMA process of order (P,M), the sample itself is said to be in ARIMA of (P,D,M).

The presence of integrating effects could also be statistically verified by tests such as *adf, pp, kpss* whose aspects we will not delve into in this work.

D. Relationship with Machine Learning

After all these analysis, this could make one think what this has got to do with concepts learnt in this course. Once a model is determined, the problem reduces to a simple Linear Regression problem of finding the coefficients. For example, let's say a given process is an AR process of order 2. (i.e) AR(P) $v[k] = d_1 v[k-1] + d_2 v[k-2] + e[k]$. Once this is known (by visualizing ACF and PACF plots), the problem reduces to finding the coefficients d_1 and d_2 which could be easily found out by setting the feature matrix X appropriately and using the methods learnt in the first paper to determine the coefficients.

III. The Problem

In this section, We will analyse the stockmarket dataset of different stocks and try to find a low risk stock with comparable returns. We then try to get a forecast of the desired stock that could help making a better informed choice.

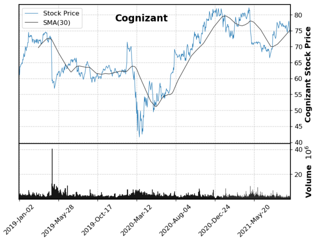
The dataset consists of Open, Low, High and Close prices of six different stocks along with Volume traded. We are also presented with the information of how INR values with respect to USD which could be used to account for inflation and judging if it's worthwhile to invest in stocks in long term at all. Let's start by understanding few finance terms.

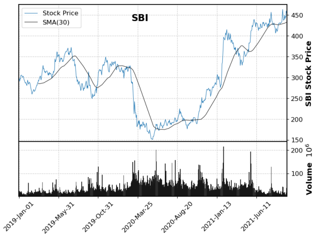
Opening Price is the price of a stock at the time of open. This need not be identical to the previous day's closing price.

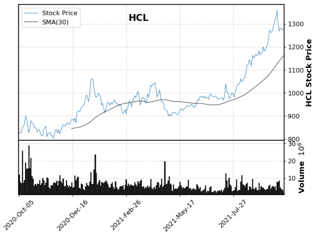
High Price is the highest selling price of a particular stock in that day. *Low Price* is the lowest selling price of a particular stock in that day. *Closing Price* is the price of a stock at the closing time of market hours. The Normal opening and closing time are from 9.15 a.m. – 3.30 p.m in India. *Adjusted Closing Price* is the adjusted stock's closing price to reflect that stock's value after accounting for any corporate actions post market time.

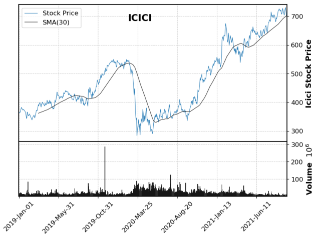
In a short term trading strategy, all these metrics would be important as these tell the traders about the volatility and the nature of stock from time to time throughout the day. Various plots such as Candle-stick plots, OHLC bar charts helps in visualization to gain insights using such strategies. But this strategy comes with its down-folds such as higher risks and extended real time analysis throughout the day.

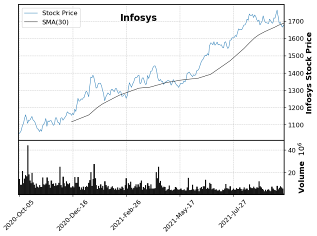
In long term keep and hold strategy, investors would like to spend in a low risk and high returns stock. Typically analysing volatility throughout a day is not necessary for such investments and one would like to keep it simple by tracking only the closing prices over time with the help of line charts.Because line charts usually only show closing prices, they reduce noise from less critical times in the trading day, such as the open, high, and low prices. In this discussion, we'll analyse stocks in a long term perspective and will account Closing Price more into account than other prices.

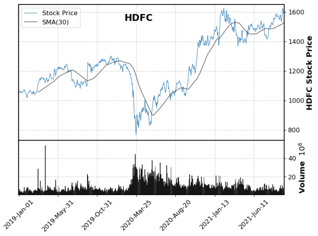








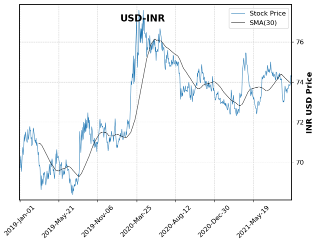




The blue line represents the stock prices while the black line is the simple moving average (SMA) of 30 days of the stock data. SMA is simply the mean of stock prices for the past 30 days. This helps in filtering random price fluctuations and smoothen it out in order to see the average value. These are used to identify trends and confirm reversals.

1) When price is above the moving average line, we consider the instrument/stock to be in an uptrend and the converse.
2) Breaking of moving average line usually implies trend reversal

For example, one could see a trend reversal around May followed by an uptrend on HCL stocks since May. On all the bank based stocks, SBI, ICICI, HDFC, one could see a downtrend on the March 2020 period which on further investigation reveals that's due to the 'synchronised slowdown' in the World Economy. We could also see a trend reversal post August period implying revertion to normal trends.

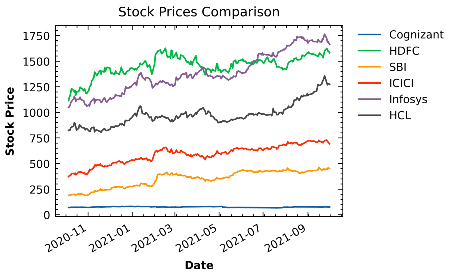


The INR plot confirms our observation that stock prices indeed went down due to decrease in INR value during that time. A trend restoration was seen at similar instances between these plots.

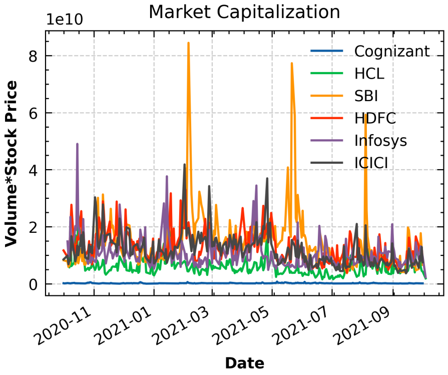
Note that since SMA is based on past prices, they are not ideal for future prices but rather confirms when a trend change has taken place. When the price crosses up and over the SMA, traders take this as a signal to validate their buying.

A. *Comparison between stocks*

Note that the time frame for few stocks are only from 2020-Oct while others are from 2019-Jan. It is worthwhile to look into the common timeframes in the same scale to get a clear idea on which stock performs better.



Most stocks except Cognizant seems to have significant uptrend and in first glance, HCL and Infosys seems to have a steady growth. HDFC seems to have uptrend but shows little stagnation in the recent times. Though these are representative of how the companies does, we should also consider Volume into account to get an estimate of amount obtained through shares.



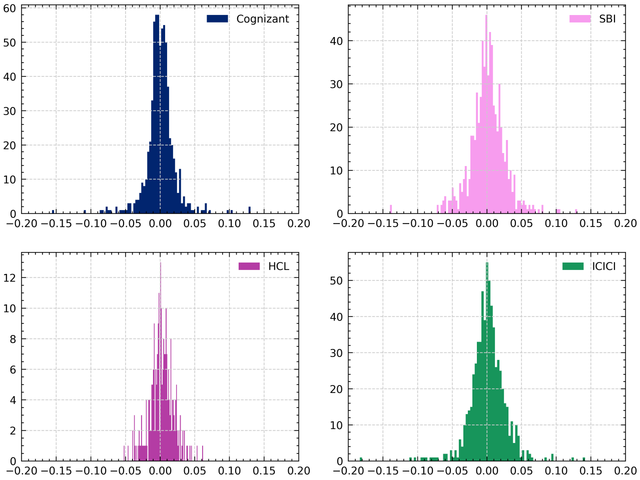
Though in the previous plot the price of SBI shares weren't the highest, SBI has the highest market capitalization. It is then followed closely by Infosys and the rest. One could see that SBI gains the most due to more stock volumes traded which wasn't considered in the previous analysis. Cognizant seems to be the least among these too

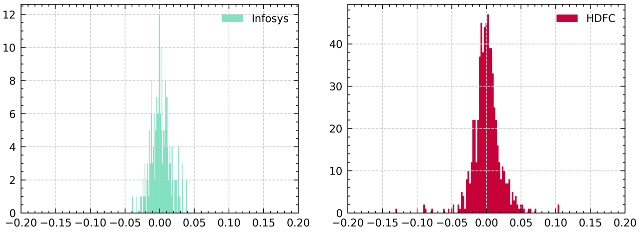
The main factor of importance to any investors is the Rate of return. One would like to earn significant profits per rupee invested, considering for most individuals investments is limited by their financial potential. If $v[k]$ is the cost of stock at k^{th} day, Rate of Return per rupee invested could be expressed as:

$$\text{ROR} = \frac{v[k] - v[k-1]}{v[k-1]} = \frac{v[k]}{v[k-1]} - 1$$

But since safer options such as Fixed deposits (which are almost equal to the inflation rates) are potential options, it is useful to analyse rate of return considering changing value of rupees. If $r[k]$ is the value of INR at k^{th} day,

$$\text{ROR} = \frac{\frac{v[k]}{r[k]}}{\frac{v[k-1]}{r[k-1]}} - 1$$



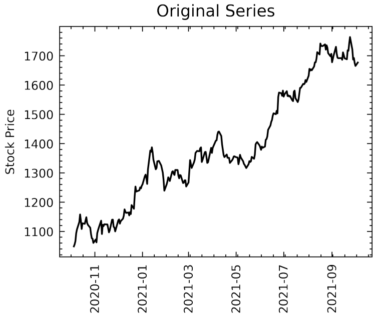


The following charts show the Return of Investments per rupee invested accounting for change in INR value. The plot seems to closely follow normal distribution centred closely to zero. The width of the distribution is representative of the volatility of the stock. Assuming we are looking for a low risk stock, Infosys seems a good option. The below table mentions the Mean returns per day per rupee invested and the associated risk (standard deviation) of the stock analysed in a year period.

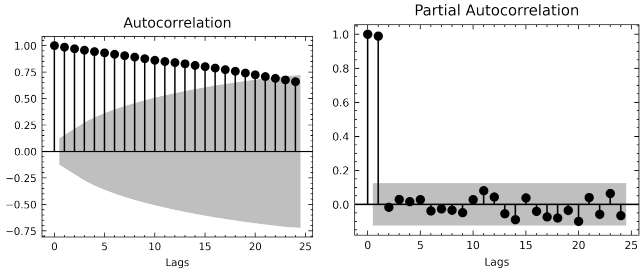
Stock Analysis (Return of Investments)		
Stock	Mean Returns(10^{-3})	Standard Deviation
Infosys	1.9056	0.0138
Cognizant	0.4756	0.0217
HCL	1.8365	0.0172
SBI	0.6150	0.0198
ICICI	0.9952	0.0255
Infosys	1.1023	0.4737

Infosys has the highest mean return with lowest risk. Though the standard deviation appears to be significant, we can expect it to be lower considering we are holding the stock for a long period of time. (Recall, $\hat{\sigma} = \sigma/\sqrt{n}$ and mean per day would remain more or less the same). Considering a risk-averse situation, we'll go ahead and choose this stock and try to forecast using Time Series Modelling. Note that the ideal selection of stock would vary on various factors including risk tolerance of a person, how stocks in a portfolio reacts to the same market situation, corporate announcements and many such factors. We are going ahead with the Infosys stock based on the fact that it has low risk and high returns in the given period analysed.

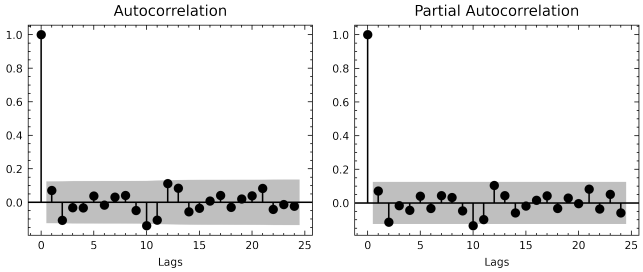
B. Time Series Modelling



On Visualizing the stock, one could see that the stock prices are not stationary and posses some integrating effect. Analysing ACF and PACF plots, we could see that



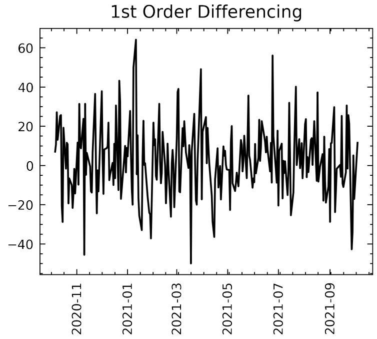
ACF plots shows long memory trend and PACF suggests a strong relationship between $v[k]$ and $v[k-1]$. This means the following could be integrating of order 1. The following was then confirmed with ADF test that it is indeed integrating of order 1. On differencing,



Of the differenced series, ACF and PACF clearly shows zero MA and AR processes. Hence

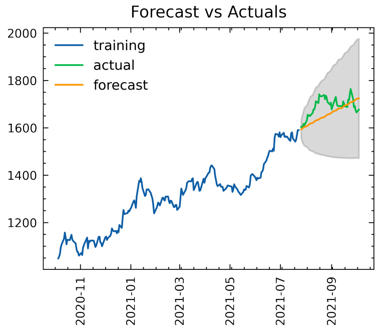
$$v[k] = v[k-1] + e[k] + c$$

is hence the modelling choice of our Infosys stock. The residuals of the differenced series shows white nose/unpredictable like characteristics and thus further strengthens our model assumption.

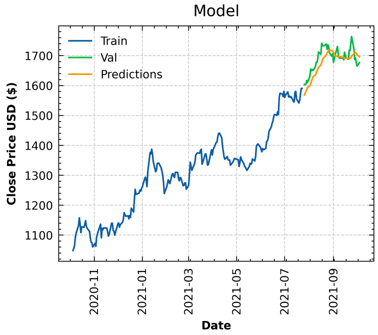


C. Forecast

A Time Series model was analysed and fitted till Oct-2020. The remaining part was forecasted and compared with the true stock predictions.



Through statistical analysis, we are also able to obtain confidence intervals for predictions. Though the model captures the necessary upwards trend and presents a good idea of error estimates, the model seems to relatively simple. A more sophisticated approach would be to use Recurrent Neural Networks to capture patterns with architectures such as LSTM's (Long-Short Term Memory). This will yield much closer results compared to traditional time series analysis and could capture non-linearity in a better fashion.



The above plot shows model predictions by a LSTM model trained and forecasted on the same data. This neural network clearly outperforms our ARIMA model. But due to complex nature and less relevance in the course-content, the mathematics behind such approach were not explored.

IV. Conclusions

Based on our analysis, the key takeaways we had from the following course are:

1) A great amount of information could be obtained by studying and exploring data. It is often essential to know more about the dataset to gain valuable insights rather than just feeding data through algorithms.
2) Real life scenarios such as stock market analysis could be cleverly constructed as Time Series through concepts learned through Linear Regression and Hypothesis testing.

Several additional factors such as following recent trends, having a portfolio reacting to different trends to minimize risks and effective trading strategy could come a long way to improve financial strength of a candidate.

References

[1] Principles of System Identification: Theory and Practice, Arun K. Tangirala.

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning

[3] An Introduction to Statistical Learning, Gareth James et al.

[4] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems.