**Methods of Prediction and Forecasting for the**

**NIFTY-50 Stocks**

*Data Analytics Phase-II Project Report*

*Team Name : 60. Paper Plain Dose*

|  |  |  |  |
| --- | --- | --- | --- |
| **Vishnu R Dixit**  *PES1201801448*  *rvdixit23@gmail.com* | **Nikhil Ram**  *PES1201801972*  *nikhilsram.off@gmail.com* | **Yash Gawankar**  *PES1201801482*  *yashgawankar@gmail.com* | **K Vikas Gowda**  *PES1201801957*  *kanthvikas2000@gmail.com* |

# **INTRODUCTION**

Stock forecasting plays a very crucial role in the finance and economics sector. Amount of research in this field has spiked over the past few years. The hallmark of a successful prediction tool is to maximize gains from given historical dataset. Time-series forecasting is a widely used method for prediction of stock values on non-stationary stock value data. Due to the complex behaviour of the market it is often impossible to predict the future trend without taking into account several hundreds or even thousands of factors simultaneously. This would require complex modelling of the periodicity of the data.

# **PROBLEM STATEMENT**

The goal of this project is to visualise various share values over time (short and long term) and observe common patterns over time for particular companies as per the dataset under consideration and hence create a data pipeline for numerical analysis of time series stock data. The predictions to be done include the rise and fall of the value and the magnitude and time period in which the same occurs. The specific problem we seek to solve is to predict the time series data for a given period using supervised learning methods in order to get a highly accurate estimation of the value at a further point in time. Further Goals would include utilizing other factors for prediction of values, and collection of non-numerical data and factoring them into the prediction.

# **APPROACH**

There are two widely used strategies to analyse stock. *Fundamental* and *Technical* analysis

**Fundamental Stock Analysis**

This form of analysis encompasses a wide variety of attributes to model and predict stock such as financial records, economic reports, company assets, and market share. Parameters such as company’s profitability, liquidity, and growth trajectory are taken into consideration. The advantages of this approach is that it is more likely to predict the sudden fluctuations as the data accounts for the company’s internal activities. But acquiring uniform fine grained data for multiple companies is not easy and the models which are used are relatively complex.

**Technical Stock Analysis**

Technical analysis is a methodology which is adopted to analyse data by observing the previous and current stock parameters. Modelling it is relatively much simpler and the predictions made using this strategy is more widely used in the stock market from the investors perspective.

# **PREVIOUS SOLUTIONS TO THE PROBLEM**

**Moving Averages**

Method of Moving Averages is a simple forecasting calculation which computes an average for *k* previous values which are considered to be a part of a window of a fixed size *k*. It is a simple extrapolation of a few previous values with respect to the time of the prediction.

**Drawbacks of Moving Averages**

This model does not model any component of a standard stationary time series data. It only estimates the values based on equal weights to each of the values in the time window. Hence it is neither very accurate nor precise.

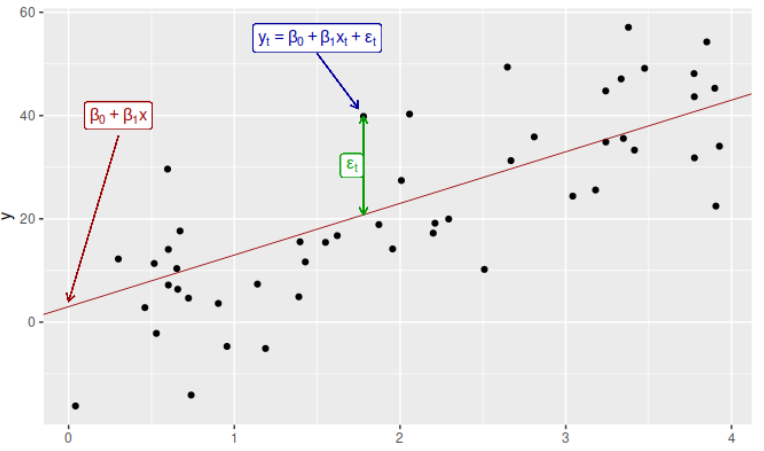
**Regression**

Regression involves modelling the data as an equation of a sum of various components, which are empirically determined and the importance of each component by the estimation of their coefficients, so as to minimize the error of the model with respect to training data. This models the general trend of the data pretty accurately for a short time interval after the last data point in the training dataset and with experimentation the regression function can match the actual model of the data. Periodicity can also be modelling using periodic mathematical functions, thereby increasing the prediction precision of the model.

For example, consider the given equation for a regression function. The regression model is the estimation of the coefficients which fit best the given model, which is determined by xt.

yt=β0+β1xt+εt

**Drawbacks of Regression:-**



Linear Regression

FIG 1.1

The drawbacks of regression for analysis of time series data are as follows. It isn’t precise as it is often not possible to model all the components due to the complexity of the patterns in the time series data. It also gives equal weightage to all the points in the training set for predicting all the points that come thereafter. This, in reality, is not the case as the movement of the data in recent times usually tends to be more important in determining the rise or fall of the data.

**Exponential Smoothing:-**

This method involves prediction of a future value using a set of non uniform, exponentially decaying sets of weight as the readings go back in time. This technique models the future values to be biased towards the most recent values of the time series and hence is a more accurate prediction than what regression would be. Multiple components of the series can be represented in the model by using multiple decaying constants.

**Drawbacks of Exponential Smoothing:-**

Exponential smoothing makes use of the current data and does not add in anything to account for the trend of the series. Hence it is always lagging behind the actual trend. This is what is sacrificed for the smoothing in the model. It also does not work very well for series which include seasonal and cyclic behaviour.

**PROPOSED SOLUTION**

**Exploratory Data Analysis**

*Data Description*

The analysis is being performed on a dataset which consists of data pertaining to **NIFTY 50 Companies.** The following companies are the top 50 companies that are listed in the *National Stock Exchange.* The data we have acquired has the daily logs of the stock for the last two decades with the following fields: Open Price, Close Price, High, Low, Trades, Deliverable Volume and Percentage, Turnover and Volumes. The data acquired is rich with no deficiencies. ‘Trades’ is the only field for which the last 10 years data is accounted into.

*Data Cleaning*

The Data acquired was largely clean with almost all attributes consisting of all the pertaining values. ‘Trades’ is the only attribute for which only the last 10 years' data is available over the entire 2 decades. There was one crucial factor which needed to be taken into account. As the value of each *unit share* increases over a long period of time, the shares tend to split up into smaller units (generally 1 share split into 2 / 2 share into 3).

There are two main reasons for companies to split their shares. The first is to establish a value which is the upper bound of a value a share can take and the second is to increase the liquidity which allows a better flow of assets through the market as an original share can be owned by one but a split share can be owned by many and this allows for a more widespread change of ownership . This hence tends to be represented as sudden false dips in the price. Hence a new column called the ‘split factor’ is introduced which indicates the factor by which the share has been split into smaller fragments with respect to the initial value hence keeping intact the true trend.



False Depression (Share Split)

FIG 2.1

**ARIMA Modelling**

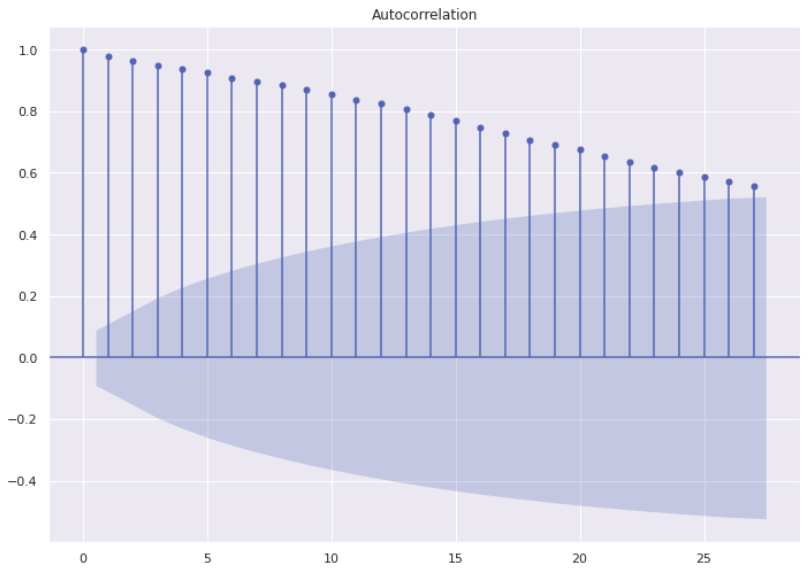
Time Series Data is composed of 4 components which inturn can be represented individually mathematically. *Trend* is the long term growth or dip analysed over the entire time period under consideration. *Seasonal* component is that pertaining to the periodic repetitions of cycles visualised repeat with a fixed time period. These include the general yearly patterns eg: financial year quarters when it comes to the finance sector. *Cyclicity* is the component which considers the systematic patterns which do not have a periodic nature. The periodicity can either grow over time or dip over time but the pattern is evident. *Randomness* is the last component which is affected by underlying natural factors which hence results in variation which cannot. be accurately predicted.

ARIMA stands for Auto Regressive Integrated Moving Average. The Trend as well as Cyclicity of the data is taken into account to generate the appropriate parameters and the ARIMA model is fit according to these values. The Integrated component ‘**d**’ is necessary inorder to make the time series data stationary where in the ARMA model is then fit onto the data. This is set by analysing the slope of the trend. Then the ARMA model which is fit to the stationary data requires 2 parameters p, q which are estimated by analysing the PACF and ACF plots. Significant seasonality was not observed in stock data hence SARIMA model is not necessary.

**ACF and PACF**

Auto Correlation Function is a method used to measure the values of auto-correlation of a stationary time-series with its previous values .Correlogram or an ACF plot is a plot illustrating the serial correlation in data over a time period .We can identify the number of significant terms needed by looking at the ACF plot.This gives us ‘**q**’ which is the number of errors in the lagged forecast.

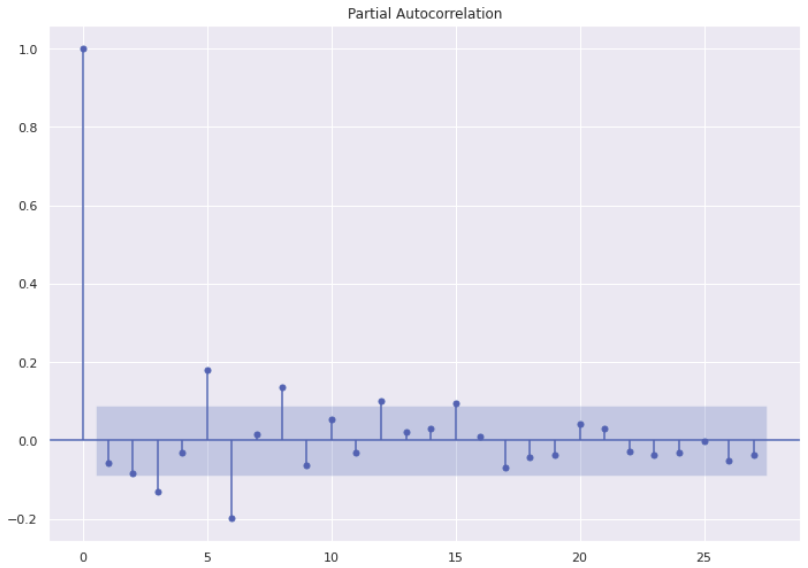
On the other hand Partial Auto Correlation Function (PACF) is used to measure the correlation of present value with the residuals. ‘**p**’which is the number of auto-regressive terms is estimated using the PACF plot.



Autocorrelation (d=0,ɑ=0.05 )

*HDFC stock - NIFTY50*

FIG 2.2



Partial Autocorrelation (d=0,ɑ=0.05 )

*HDFC stock - NIFTY50*

FIG 2.3

**Benefits of ARIMA**

ARIMA is well suited for generating computationally simpler models with less amount of data. Hence it is appropriate for *Cold Start*.

**Drawbacks of generalised ARIMA Solution**

ARIMA in general has one major drawback where a generalised model cannot be trained and used to model all stocks. The autocorrelation and partial autocorrelation parameters could vary from stock to stock and time to time. Another drawback of ARIMA is that it cannot significantly predict the long term trends well.

**Rectified ARIMA Modelling**

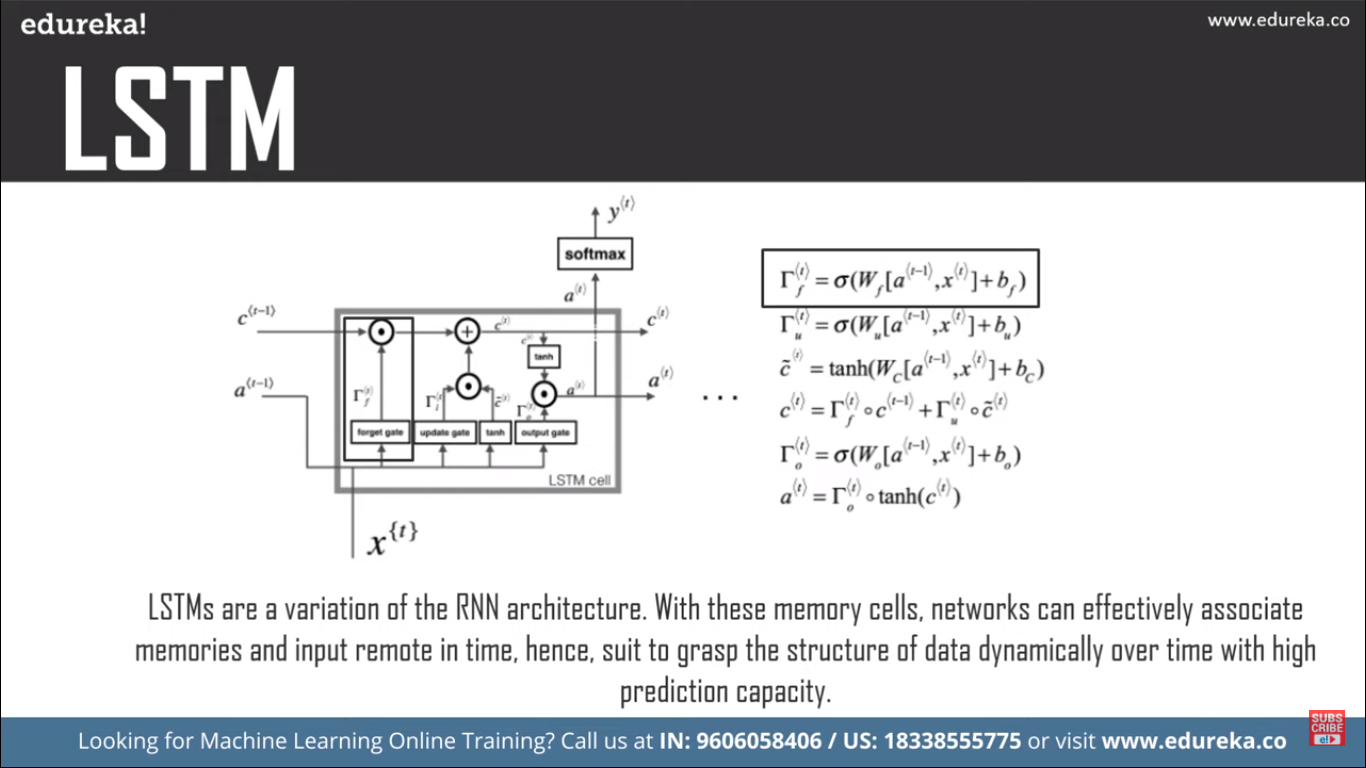
The major drawback of ARIMA is rectified by defining a custom function which analyzes the ACF and PACF plots with a confidence of *95%* and the most significant p, q attributes are used to train the model. Hence this function is extrapolated for each stock independently and separate models are trained.

**LSTM**

LSTM stands for Long-Short-Term Memory. It is a deep learning methodology whose infrastructure is mainly a recurrent neural network accompanied with data structures called gates used to store history in a raw form or in states. The three types of gates are input gate, output gate and forget gate and these allow information storage within the network.

These gates allow for the network to have *memory*  of the past data and also ensure proper flowing and retention of data and dependencies between several data items. A brief description of the gates are as follows.

Input gates ensure that inflow of the data into a cell is given the appropriate weight or importance. The deterioration or retention of a value in the network are controlled by the forget gate. At every iteration of the computation in the network, one phase of computation (not the output) is sent back as the input to the network and using this data, the output of the network which is the desired result is computed. The output gates computed how important this intermediate result is to the output of the network.



LSTM Functioning

FIG 2.3

Since LSTMs model memory and the unequal importance of the data inflowing at each point in time, this allows for highly accurate and precise computations of the future values of the time series data.

**Algorithmic Trading**

Algorithmic trading is a concept which involves the automation of the process of trading. The system input includes the program which determines the decisions to make using a combination of several predictive methods which model variables like price, time and volume and indicator variables along with a minimal amount of human oversight post the automated process. This method is sought after due to the capacity to book large financial gains off of small deltas within the price in a time series stock data. The employment of algorithms to book these profits for small differences in stock values in a short period of time is called the scalping trading strategy.

**DMAC**

The Dual Moving Average Crossover is an algorithmic trading system that makes use of two moving averages, one short and one long. The trading of the stocks occurs when the two moving averages cross each other.In the statistics of time series, and in particular the stock market technical analysis, a **dual moving-average crossover** occurs when, on plotting two moving averages each of which are based on different degrees of smoothing, the traces of these moving averages cross. This algorithm makes use of two (or more) moving averages, a slower moving average, ie. a long term average and a faster moving average, ie. a short term average. A short term moving average is *faster* as it only considers prices over a short period of time and is thus more reactive to daily price changes. On the other hand, a long term moving average is deemed *slower* as it monitors the stock prices over a longer period of time, but tends to smooth out price noises which are often reflected in short term moving averages.

When a faster moving average crosses a slower moving average, a crossover takes place. In the process of stock trading, this point of intersection is used either to enter (buy or sell) or exit (sell or buy) the stock market. Such a crossover can be used to signal a change in trend and can be used to trigger a trade in a black box trading system.

**EXPERIMENTAL RESULTS AND OBSERVATIONS**

**Short-Term Trend Prediction (ARIMA)**

ARIMA prediction tends to perform extremely well in predicting the next days’ immediate stock trends given the time series subset until the current day.

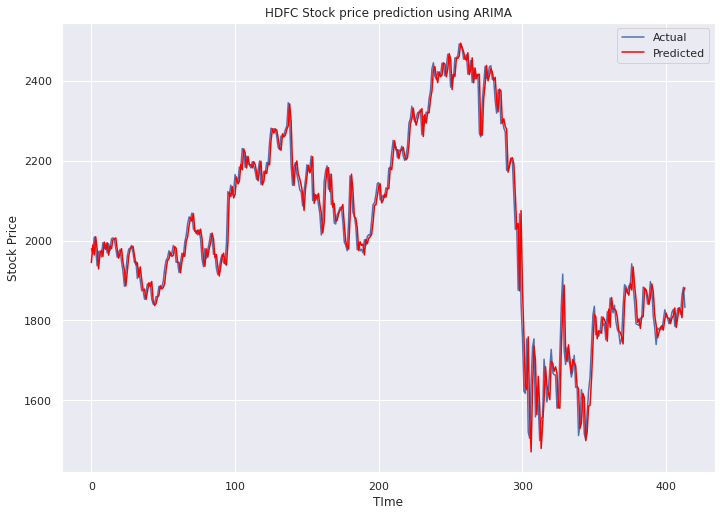
The model was trained and tested on the *Nifty50* stock *HDFC*. The subset was taken from 2015 to mid 2020 and a 66:33 train-test split was made.

*The observations made were as follows* for estimating the next immediate days’ trend*.*

**RMSE: 59.99**

**MAE: 2.1759**

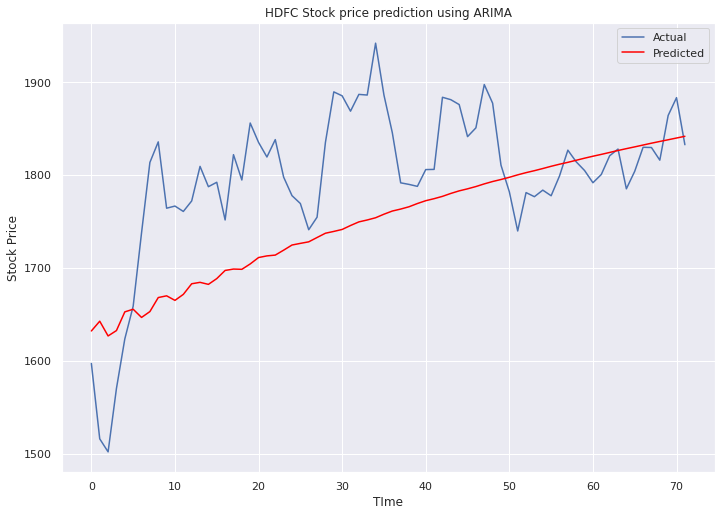
**MAPE 0.0221**



ARIMA Prediction

*HDFC stock - NIFTY50*

FIG 3.1



ARIMA Failure in predicting extended future

FIG 3.2

ARIMA is not good at predicting the trends far into the future. It tends to linearly extrapolate the existing trend by not taking into account the fluctuations.

**Prediction using LSTM**

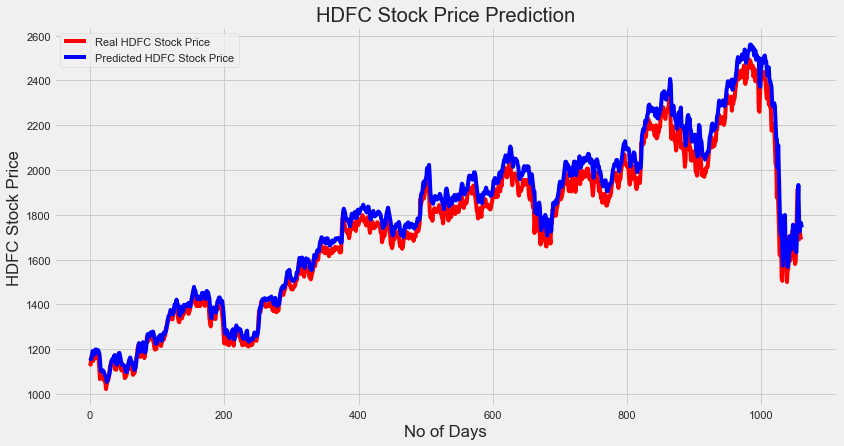
LSTMs are designed using Recurrent Neural Networks and are used for sequence prediction problems along with infrastructure to maintain observed data in either raw or in the form of states within the network. Thus it seems to perform very well in Stock Prediction when it comes to predicting the long term trends as well.

The model was trained and tested on the *Nifty50* stock *HDFC*. The subset was taken from 2000 to mid 2020 and a 4000:1000 days train-test split was made.

**RMSE: 40.74410744**

**MAE: -22.20397173**

**MAPE 0.0190386**



LSTM Prediction

*HDFC stock - NIFTY50*

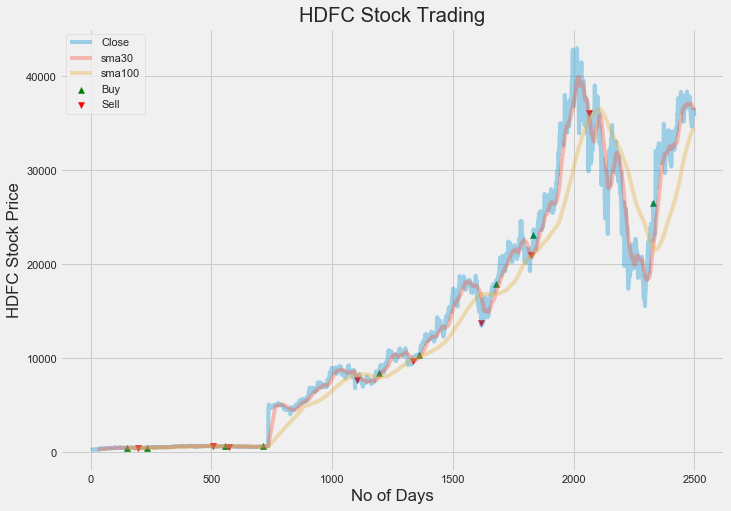
FIG 3.2

**Trading Using the Dual Moving Average Crossover (DMAC)**

In the dual moving average crossover trading strategy, these crossovers are points of decision to buy or sell the currencies. The Technical Approach suggests that when the Short Term Moving Average (STMA) moves above the LTMA, that represents a Buy (or Long) signal.Conversely, when the STMA moves below the LTMA, the Technical Approach indicates a Sell (or Short) signal.

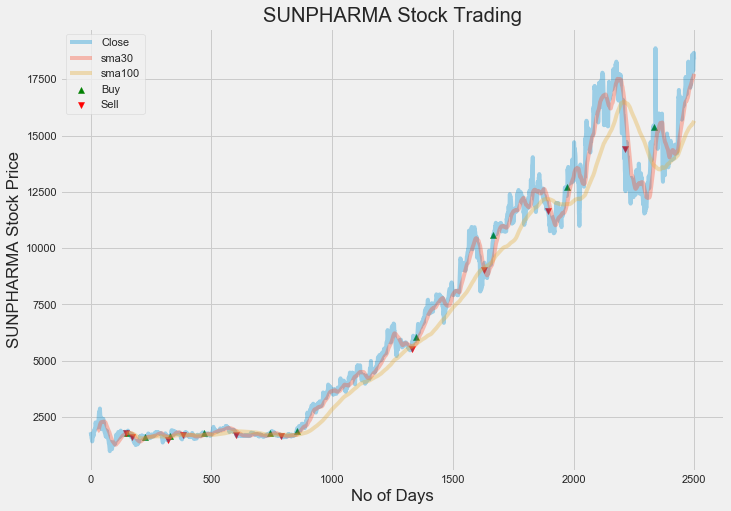
The following graphs represent the Buying and Selling points for two NIFTY-50 stocks, namely HDFC and SUNPHARMA.

The Red Downward arrows indicate a sale of stock and the Green Upward arrows indicate a purchase in stock.



HDFC Stock Trading

FIG 3.4



SunPharma Stock Trading

FIG 3.5

# **FUTURE WORK**

The goal of this report is to extend the current application to the real world. The addition of a few more *Indicators* and an *Interfac*e to access a *Demat API* can make stock trading much more interactive and simpler to automate. The various models which are represented above have different characteristics and a suitable ensemble can be established based on whether the client wants to invest for the short-term or the long-term.

# **CONCLUSIONS**

The modelling of patterns of stocks is often complicated and multi-dimensional. Therefore conventional methods where a fixed model is imposed onto a data set will not always work out given the nature of the problem. Hence it is better to resort to methods where the model of the data itself is dynamically computed geared towards each dataset individually. This ensures that it is not oversimplified or overgeneralised and gives better predictions due to its singular focus for a given dataset. ARIMA provides a very simplistic approach of modelling the trends linearly with the data and the error from previous timestamps. *Long and Short Term Memory (LSTM)* is implemented using a recurrent neural network with data structures to represent *memory* over the entire duration of the time series with the respective weights at each point in time within the prediction model. Since this is a very versatile method with several parameters influencing how the forecaster models the data, there is a lot of scope for variations and tweaking which offers the user flexibility during implementation..

The following analysis has provided deep insights on how data which is affected by real world scenarios such as the economy, industrial sector and the global performance of the companies can be relatively modelled well by just analysing the trends and behavior of the same parameters in the past. In many situations when the source of real world factors of the data are obscure, the inherent pattern from the past can be exploited to make predictions.