Literature Survey and Exploratory Data Analysis for the NIFTY-50 Stocks

Phase-1 report

Team Name : 60. Paper Plain Dose

Vishnu R Dixit	Nikhil Ram	Yash Gawankar	K Vikas Gowda
PES1201801448	*PES1201801972*	*PES1201801482*	*PES1201801957*
rvdixit23@gmail.com	*nikhilsram.off@gmail.com*	*yashgawankar@gmail.com*	*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. In order to make it real time and dynamic, feeding in a real time data API would allow for buying and selling through an buying and selling API provided by APIs provided by financial service companies (Zerodha). Further Goals would include utilizing other factors for prediction of values, and collection of non-numerical data and factoring them into the prediction.

EXPLORATORY DATA ANALYSIS

The first step was the observation on raw data .Currently we have selected a dataset which consists of chronological data of every day of 50 prominent stock labels for a period of around 20 years. This includes the fundamental fields like open, close, high, low, trades etc for every day.

Our end-goal is to analyse this data and develop a model which can be applied onto a much larger dataset which has these fundamental parameters and can be used to predict the result based on either long term/short term necessities of the client.

Plot of share values over time

Fig. 1.1

Some of the methods used to visualize the data given are as follows.

Data Description

Currently, 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).

Companies choose to split their shares so they can lower the trading price of their stock to a range deemed comfortable by most investors and increase the liquidity of the shares. 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 1.2

Candle-Stick Plot

This is a unique plot which uses the box plot pattern to summarize the stock events of every unit time interval. The 2 tips of the box plot are used to indicate the high and the lowest price of the day.

The upper and lower ends of the lobe of the box represents either the open/close price or vice-versa. This representation depends on the color coding of the box. Red and green colours are used to indicate whether the close price is higher than open price or vice versa respectively.

Candle Stick Plot

Fig 1.3

Moving Average Plot

These plots are a smoothened version of a time series plot which average a fixed number of previous values in the time series data. This number is often called the window size. Figure 1.4 is a plot of the time series data with a 12 value window moving average of the closing price of any given stock.

Moving Average

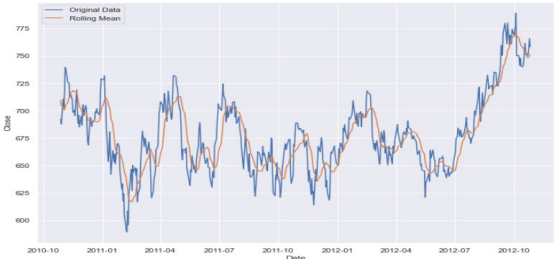

Fig 1.4

Moving Standard Deviation

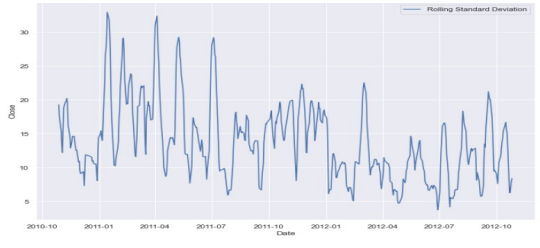

Fig 1.5

Moving standard deviation plot

A moving standard deviation plot calculates the standard deviation for all the values in the window. This would indicate the magnitude of the fluctuations of the value of the stock. When moving standard deviation is calculated with the difference between the high and low price, it gives an estimate

of the extent of fluctuation of fluctuations of the share value. *Fig 1.5* shows the rolling standard deviation 12-window plot for the close value of a given stock.

LITERATURE SURVEY

Reference [1]:

Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models

There are fundamentally two methods for analysis of share value time series data. Namely, Statistical prediction (Technical Analysis) and Non Statistical prediction(Fundamental Analysis).

Statistical prediction of stock prices is possible through several different means. Namely, Multivariate linear regression, Multivariate adaptive regression spline (MARS), Regression tree, Bootstrap aggregation (Bagging), Extreme gradient boosting (XGBoost), Random Forest, Artificial Neural Networks, CNNs, RNNs (LSTM) and Support Vector Machine

Non statistical means include text mining and natural language processing of data from several sources on the internet to assess the intrinsic value of the stock and whether it is under or overvalued by the market.

Inability to identify features within such a serial time series data would mean that the ideal approach for forecasting would involve deep learning methods or arbitrary choices of features.

A non deep learning method used could be running a regression with one of the value variables like open value or close value with all the other variables recorded for a given day.

An extension of this could be done where the dependent variable is the daily delta. Although this does not capture the complete mathematical picture, it is known to have substantial accuracy. In order to efficiently capture all the data given efficiently such that each independent variable holds more meaning, a few normalized variables are used for building predictive models.

For example : (high - H_{min})/(H_{max} – H_{min}) would capture the comparative magnitude of the high value in comparison to several other values within a time frame. The same can be done for all the other variables low, close, range and volume

RNNs have a problem of vanishing and exploding gradients as they either never converge or get stuck on a local minima. This is done by implementing gates in a network, namely *forget gates*, *input gates* and *output gates*. This network architecture with the backpropagation through time algorithm provides a high degree of power in the forecasting of time series data. LSTM networks consist of memory cells which “forget” the past with time hence preventing them from getting stuck on a local minima

Reference [2]:

A Prediction Approach for Stock Market Volatility Based on Time Series Data:-

This paper gives insight on the structured methodology which must be incorporated in order to analyse time series data and build a model to predict the solution. 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 yearply 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. Hence, The goal is to develop a model which can accurately substantialize the former three parameters let alone Randomness without overfitting the past data.

One of the most prominent approaches to dealing with Time-Series data and hence modelling employing a systematic approach. Initially the trend of the associated data is visually analysed by line plots and moving average plots. This is performed to determine whether the data represents a stationary or non-stationary trend. If not stationary, whether there exists a gradual dip or rise as well as analyse the overall linearity.

Finally, ARIMA modelling is performed on the data and testing is performed. The benefit of the ARIMA model is that it incorporates even certain characteristics of randomness hence is sometimes

successful at predicting the randomness of the trend in the future.

Identify the time series type

Dickey - fuller Test is used for this purpose. A statistical tool used to test if a given time series is stationary .(stationary time series is the one which does not have trend or seasonality). This is achieved by testing the null hypothesis $\alpha = 1$ (alpha is the first lag on Y) in the following model equation .

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e$$

ACF and PACF

Autocorrelation and Partial autocorrelation. They are helpful in identifying the the seasonal and cyclic variation along the rolling mean

ARIMA

AR - Auto Regression : The regression is done based on the list of previous values. Hence y_t is predicted based on y(t-1),y(t-2).... and so on

MA - Moving Average : The value y(t) is predicted based on the errors of the prediction of y(t-1),y(t-2) and so on…. Hence e(t-1),e(t-2) and so on

I - Integrating : Integration is implied as the inverse of differencing. It is the degree of differencing that needs to be done on data. In order to transform a ‘‘non-stationary time series into a stationary one’’, the series needs to be differenced

ARIMA Model

It encompasses all the following factors which over time has proven to model the future reasonably well

Reference [3]:

Trend Trading: The 4 Most Common Indicators

The method of trend trading tries to capture gains through the analysis of an asset's momentum in a particular direction. The prominent trend indicators to quantify this momentum of the stock value as mentioned in the article are as follows

Moving averages : This is an average of a certain value over a well defined time interval. It does not predict the future value. It is one empirical method to identify buy and sell value.

Moving Average Convergence Divergence : MACD is an oscillating indicator about a central band. Indicates trends and momentum. Consists of two lines (Fast line and the slow line). The Crossover of these indicate that the said time will precede a rise or a fall.

Relative Strength Index : This index is an oscillating indicator between 0 and 100. It represents the over or undervalue of a share at any given time. A good rule of thumb to decide when to buy and sell it if the value of the RSI is less than thirty and over 70 respectively.

On-balance Volume Indicator : It is calculated using the volume of shares sold/bought .Volume is known to confirm trends of a stock as a rise in price would be accompanied by a rise in OBV. OBV rising without the price indicates a future rise .OBV flat lining and price rising would mean a pause in growth or the start of a fall of the value of the stock.

REFERENCES

[1] Sheikh Mohammad Idrees, M. Afshar Alam, and Parul Agarwal, *“A Prediction Approach for Stock Market Volatility Based on Time Series Data”*(2019)

[2] Sidra Mehtab, Jaydip Sen and Abhishek Dutta, *“Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models”*

[3] Kamalakannan J, Indrani Sengupta and Snehaa Chaudhary, *“Stock Market Prediction Using Time Series Analysis”* (2018 IADS International Conference on Computing, Communications and Data Engineering)

[4] Xinyi Li, Yinchuan Li, Hongyang Yang, Liuquing Yang, Xiao-Yang Liu, *“DP-LSTM: Differential Privacy - inspired LSTM for Stock Prediction Using Financial News”* (Dec 2019)