Analyzing and Predicting Stock Market Using Data Mining Techniques – A Review

Suraj Baviskar[1], Nitin Namdev[2]
[1]ME Scholar, Department of CSE, JIT Borwan, Khargoan, (M.P), India.
[2]Assit.Prof. Department of CSE, JIT Borwan, Khargoan, (M.P), India.

***Abstract*— Stock market is generating enormous amount of valuable trading data. This paper presents a review on various data mining techniques applied to stock trading data to analyze and predict stock trend, outline of proposed model for stock prediction system based on technical, fundamental and external environmental factors. This paper focus on understanding stock market related information for investors and analyze National Stock Exchange data to predict stock future movements. Review on Technical indicators, price based and volume based, which are widely used to analyze stock data. Stock analyst, investors and stock brokers are trying to analyze stock trading data to predict future movement of stocks. Data mining is used to extract meaningful previously unknown rules and hidden patterns in data.**

***Index Terms*— Data Mining, Stock data analysis, SMA, EMA, MACD, RSI, MFI, PVT.**

I. INTRODUCTION

Public companies raise funds by issuing shares to public and institutions. By purchasing shares, an investor acquires partial ownership of company. A stock market is a place where public listed company stocks are traded. Stocks are exchanged among buyer and seller which generate transaction data. Prices are changing as per demand and supply of stocks. All trading data is captured by stock exchange where stock companies are listed. Stock trading data is nonlinear, fluctuating hence highly time variant. A lot of information is hiding in this huge data captured by stock exchange is difficult and time consuming for human being to extract without powerful tools. Thus predicting future price of stock is highly challenging.

Stock Market

Stock prices are changing due to demand and supply of stocks in stock market. Based on past history and current information, investors are buying or selling stocks of listed companies. Investors, stock analyst and stock brokers are predicting demand and supply of stocks after studying fundamentals and technical information of stocks. In order to study and analyze stocks data, investor need to go through vast amount of data to discover hidden patterns which is very cumbersome and tiring task. Usually stock analysts analyze stocks based on three vital aspects - external environmental factors, fundamental analysis and technical analysis.

External Environmental factors - These are factors which impacts running business in specific environment such as interest rates, currency exchange rates, industry specific information such as government policies, growth rate of industrial production and consumer price, oil and commodity rates etc. Usually this information is published in media.

Fundamental Analysis - Fundamental analysis refers to company specific financial information such as income statement, dividend yields, acquisition, winning new deals, revenue growth, order book growth, earning per share (EPS), face value and book value (BV) of share etc. In short, fundamental analysis tells us value of the company stock with regards to it potential growth in earnings. This information is relatively static for financial quarter and can be easily found on companies' official web-sites or with stock exchange.

Technical Analysis - Technical analysis refers to the price and volume movement of stock. Technical indicators are used to identify patterns and trends based on stocks' historical price and volume data. Short-term, mid-term and long-term data is considered for mapping and identifying corresponding stock trends as per investment strategies. Technical information is rapidly changing with time. Apart from intraday trading strategies, investors are considering closing price of stock as market price for that trading day.

On a typical trading day, analysts are looking for previous day's stock information like open, high, low, close price of stock along with volume traded. This basic trading information gives lot of hidden information like stock volatility, increase or decrease in volumes, support and resistance levels, overbought and oversold condition etc. In order to quick visualize this information, analysts are using either OHLC (open, high, low, close) charts or Candlestick charts. Analyzing hundreds of stocks using these methods are time consuming, error prone and tiring. Hence, trading firms, stock brokers and investors are rely on powerful software tools using data mining technologies to study and predict stock trends to take appropriate position in the market.

Tools or technical indicators used to analyze stocks are classified into two broad categories:

A. Price-Based Indicators-

Moving averages- There are two important moving averages like Simple moving averages (SMA) and Exponential moving averages (EMA).

Moving averages indicates average value of stock over a period of time (n). SMA applies equal weight to all prices over a period of time whereas EMA applies more weight to recent prices compare to previous prices.

$$SMA\ (n) = (P1 + P2 + P3 + ... + Pn\ ) / n$$

EMAc – Current Exponential Moving Average
EMAp – Previous Exponential Moving Average

$$EMAc = ClosePrice * (2/(TimePrd + 1)) + EMAp\ (1 - (2/TimePrd+1)))$$

Moving Average Convergence and Divergence (MACD)- The MACD is difference between 12-day exponential moving average (EMA12) of a stocks' price and 26-day exponential moving average (EMA26) of its price.

$$MACD = EMA12 - EMA26$$

The result is an indicator that moves above and below zero. MACD above zero imply 12-day moving average is higher than 26-day moving average which is upward shift in demand supply. If MACD falls below zero, it suggest downward shift in demand supply. 9-day EMA of the MACD is known as Signal line. The first value of signal line is simply a 9-day trailing average and all other values are given by below equation, where the time period (TimePrd) is 9.

$$SIGNALn = MACDn * (2 / (TimePrd + 1)) + SIGNALn -1 (1 (2 / TimePrd + 1)))$$

Relative Strength Index (RSI)- Relative Strength Index is calculated based on the comparison of the gain of stock to the loss of stock. It is the ratio of the Upward (U) exponential moving average of stock to the Downward (D) movements for the stock. It is expressed in values from 0 to 100.

If RSI of stock is above 70 then stock is considered as overbought and if RSI is below 30 then stock is considered as oversold. Usually RSI of stock is calculated for 14 days timeframe.

Pc = current Closing price,
Pp = previous Closing price,
U=Upward stock moment,
D=Downward stock moment,
On a day when the stock has closed up,
U = Pc – Pp and D = 0,
On a day when the stock has closed down,
D = Pp – Pc and U = 0,

Then EMA is calculated for U and D for a period n, represented respectively by EMAup and EMAdn.
Now RSI is calculated as follows:

$$RSI = 100 * (EMAup / (EMAup + EMAdn))$$

Pivot point- Pivot point is considered as point of rotation. It is a crucial point indicating which way market is heading during the course of the day. Pivot point shows critical support and resistance levels at which stock price can change. Pivot is calculated by open, high, low and close price of stock from previous trading day.

$$Pivot = (High + Low + Close) / 3$$
$$Support1 = (2 * P) - High$$

*Resitance1 = (2*P) - Low*
Support2 = P - (R1+ S1)
Resistance2= P - (S1 + R1)

Bollinger Bands- Bollinger Bands are volatility indicator and drawn using standard deviation above and below a simple moving average. Usually stock price stays within upper and lower Bollinger band. Bollinger band widen during high volatility in price and it narrows when price is less volatile. If stock price is near to upper band means stock is overbought and if it's near lower band stock is oversold.

n=n-day Time period
D= Number of standard deviation

MiddleBand = Sum (CLOSEn) / n

*UpperBand = MiddleBand + {D * sqrt (sum (CLOSE - MiddleBand)^2 / n)}*

*LowerBand = MiddleBand - {D * sqrt (sum (CLOSE - MiddleBand)^2 / n)}*

52week high and low- Highest and lowest stock price over the last year.

B. Volume-Based Indicators

On Balance Volume (OBV)- OBV is momentum volume based indicator which considers daily stocks volume in its construction. If stock is closing positive then volume is added otherwise volume is subtracted. Thus OBV keeps a running total of volume. OBV shows if volume is flowing into or out of stock. If high volume flows into the stock with same or high price indicate more demand for stock.

Price Volume trend (PVT)- Price Volume Trend relates stock price with stock volume traded. PVT is calculated by multiplying the day's volume by percent that the stock's price changed from previous day close and adding this value to a cumulative total. PVT is more accurate compare to OBV as it indicates accurate flow of money into stock.

*PVT = {((close - PreviousClose) / PreviousClose) * Volume + PreviousPVT}*

Money Flow Volume (MFI)- MFI measures strength of money flowing into the stock and money flowing out of stock. Thus, MFI ensures the reliability of current stock trend. MFI uses volume and price to measure demand and supply or buying and selling pressure. MFI is positive when stock prices are up and it is negative if stock prices are down.

*MFI = {[(close - low) - (High - low)] / (High - low)} * Volume*

Data Mining Techniques

Data mining refers to extracting or mining knowledge from large data sets. The ultimate goal of data mining is forecasting. Some of its functionalities are associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis.

Association Rules- The association's rules used to discover the relationships between items or variables that occur synchronously in the data set. Association rule discovery systems return all rules that satisfy user specified constraints thus permitting the user to identify which specific rules have the greatest value. It determines which things are found together. It focuses on categorical data rather than on numerical data.

Classification- Classification refers to examining the features of a newly data set and assigning it to one of predefined set of classes. Classification task has to build a model that can be applied to unclassified data to classify it. Data classification can be done in many different methods; most commonly used method is the classification by using Decision Tree.

Decision Tree- It represents set of rules which can easily expressed by humans in a tree structure. They are used to divide large heterogeneous data into smaller, more homogeneous data group with target variables. Decision trees are easy to build with known examples and can be visualize easily. It works smoothly on large amount of data and on diverse data types. Once decision tree is built it can be applied to new data to classify it.

Clustering- Clustering finds groups of records or items that are similar to one another. High similarity within a group indicates good cluster whereas low similarity between patterns shows they belong to two different groups. Clustering is an example of undirected data mining. Clustering solves classification problems by distributing cases into groups or segments so that degree of association can be stronger between members of same cluster and the degree of association is weak between members of different clusters. Clustering is used to find customer segments out of large amount of data. Partitioning methods, hierarchical agglomerative clustering and density based clustering are few classical clustering approaches normally used.

Neural network- Neural network mainly address the classification and regression tasks of data mining. They can find nonlinear relationship among input and predictable attributes. Decision trees and artificial neural networks can be trained by using an appropriate learning algorithm.

Multilayer perception is most popular type of directed modelling neural network. Normally, it consists of input layer, many hidden layers where calculations are performed and finally passed to the output layer. Major advantage of neural network is its ability to approximate any continuous function without making any assumptions about the form of function. Training neural network is done by adjusting weights for the network. Back propagation calculates overall errors of the network by comparing it with values produced on training set. Then it adjusts weight accordingly to reduce overall error. Normally it takes longer to learn or train neural network compare to decision trees.

Time Series- Time Series is a sequence of data points measured at successive uniform time intervals. Goal of data mining is to identify what is happening over time. Time series data has wealth of information for that time frame of data. Regression analysis is most popular tool for modelling time series, finding trends and outliers in such data set. Similarity search used to find data sequences that differ slightly from given query sequence.

II. RELATED WORK

Data mining technologies are widely used in various applications such as financial domain, telecom domain, healthcare domain, agriculture domain, e-commerce, marketing etc. In order to predict stock market future movement, researchers have implemented data mining methodologies like decision tree, association rule, clustering, artificial neural network, support vector machine, fuzzy system, genetic algorithm, time series mining and mixed methods.

Muh-Cherng Wu et al.[6] presented a stock trading method by combining the filter rule and the decision tree technique to generate candidate trading points. Authors also considered both the past and the future information in clustering the trading points.

Jianfei Wu et al. [7] has compared clustering algorithm with core patterns and observed that core patterns are more stable as stock price evolves. Their algorithm which accepts only one parameter is more effective when compared with the DBSCAN clustering algorithm.

The authors of [8] made an empirical study on building a stock buying/selling alert system using back propagation neural networks. The system was trained and tested with past price data from Hong Kong and Shanghai Banking Corporation Holdings over the period of one year in 2004.

Lee et al.[9] has used technical indicators as input variables to feed forward neural network for prediction of NASDAQ and Taiwan Stock Exchange.

Aditya Nawani et al. [10] designed a market prediction system using neural network data mining techniques and used MATLAB GUI to make accurate predictions for trading firms.

El-Baky et al., [11], proposed a new approach for fast forecasting of stock market prices using new high speed time delay neural networks (HSTDNNs). They used the MATLAB tool to simulate results to confirm the theoretical computations of the approach.

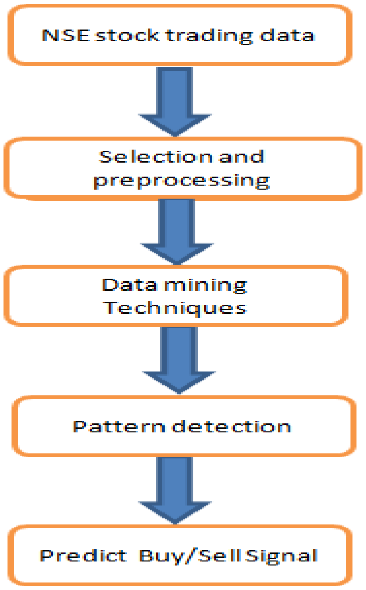
III. PROPOSED WORK

To predict National Stock Exchange market precisely is very complex task till date. We propose to build simple user friendly stock analysis and prediction system based on technical data, fundamental data and external environment factor data derived from raw trading data generated by National Stock Exchange. Analysis based on this model will improve accuracy in identifying good stocks to invest as a part of individuals' financial diversified portfolio. Proposed analysis and prediction system will be easy to operate, understand and provide clear categories of uptrend, down trend stocks, fundamentally strong stocks for mid to long term investment, buy, sell or wait signals for intraday and very short term investment. Depending on investors' financial requirements and risk taking capacity, investor can choose investment strategies suitable to his style with the help of this system.

Here is high level end to end process of proposed system:

1. NSE stock trading process understanding.
2. NSE stock trading data understanding.
3. Data preparation process.
4. Modelling.
5. Evaluation of models.
6. Deployment or implementation.
7. Preserve data.

Flow diagram for proposed process:



Steps for Implementation-

1. **Data collection** - Retrieve daily trading data in comma separated format from NSE website after trading hours which reflects day closing prices of NSE listed stocks.

2. **Select and cleanse data** – Select, cleanse and map gathered data to predefined format.

3. **Data Processing** - Process gathered data to predefined model.

4. **Apply data mining techniques** - Identify patterns from mapped and processed data using various data mining techniques.

5. **Prediction** - Make a prediction based on analyzed data to flash buy or sell or wait signal for investors.

NSE stock trading data description-

It is important to understand stock data before analysis. National Stock Exchange is recording daily trading data and publishes it at the NSE website after end of trading hours. This data is available in comma separated value (CSV) format and database format. Historical data is also available for all past trading sessions which can be used for analysis purpose.

Here are few sample records from typical stock CSV format data.

```
SYMBOL,SERIES,OPEN,HIGH,LOW,CLOSE,LAST,PREVCLOSE,TOTTRDQTY,TOTTRDVAL,TIMESTAMP,TOTALTRADES,ISIN,
20MICRONS,EQ,30.85,31.45,30.6,30.75,30.75,30.75,50904,1582699.2,15-JUL-2015,219,INE144J01027,
3IINFOTECH,BE,4,4.35,3.95,4.35,4.35,4.15,1927403,8183028.75,15-JUL-2015,994,INE748C01020,
3MINDIA,EQ,8199.9,8329,8161.3,8200.85,8310,8194.15,1179,9676426.6,15-JUL-2015,214,INE470A01017,
8KMILES,EQ,900,929,890.3,897.35,892,894.85,47010,42748144.55,15-JUL-2015,2829,INE650K01013,
A2ZINFRA,EQ,24.95,25.9,24.7,25.3,25.55,24.9,755748,19118912.65,15-JUL-2015,2309,INE619I01012,
AARTIDRUGS,EQ,669.4,694.95,666.1,681.75,686,662.95,113338,77643890.75,15-JUL-2015,3728,INE767A01016,
```

Proposed system model-

System will collect raw history data from various locations primarily from National Stock Exchange website and store it in database. We prefer to use open source database to scripting language which is compatible for simple web based client server architecture. Various data mining methodologies are used on filter, classify and categorize data by extracting useful patterns or rules to represent it in simple form for investor. Investors, as per his investment goals, can choose stocks analyzed by this system. Investor requires any web browser to access stock analyzed web-page. This system will predict and recommend stock name, action either buy or sell, entry price level, target price and stop loss.

IV. CONCLUSION

Data mining has been extensively used to extract vital information from history stock data to analyze and predict its future trends. This paper presents stock market related information, process, technical indicators and tools to analyze stock exchange data in addition to that it covers review on various techniques of data mining to predict stock market using various strategies for investor and brokers. Based on this information, proposed system model can be build to formulate various stock trading strategies with suitable data mining techniques to cater investors' requirements.

REFERENCES

[1] Book "Data Mining Techniques", by Michael J.A. Berry and Gordon S. Linoff, 2nd edition

[2] Dunham, M. H. & Sridhar S.(2006), "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition.

[3] Han, J., Kamber, M., Jian P. (2011). "Data Mining Concepts and Techniques". San Francisco, CA: Morgan Kaufmann Publishers.

[4] Sang C. Suh, Ed. Practical Applications of Data Mining.

[5] Pring, M.J. (1991). Technical analysis explained. New York: McGraw-Hill.

[6] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) "An effective application of decision tree to stock trading", Expert Systems with Applications, 31, pp. 270-274.

[7] Jianfei Wu, Anne Denton, Omar Elariss, Dianxiang Xu, "Mining for Core Patterns in Stock Market Data", ICDMW, 2009, 2013 IEEE 13th International Conference on Data Mining Workshops, 2013 IEEE 13th International Conference on Data Mining Workshops 2009, pp. 558-563

[8] Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T.L. (2007) "Design and implementation of NN5 for Hong Kong stock price forecasting", Engineering Applications of Artificial Intelligence, 20, pp. 453-461.

[9] Lee, C-T., and Chen,Y-P. 2007. "The efficacy of neural networks and simple technical indicators in predicting stock markets." In Proceedings of the International Conference on Convergence Information Technology, pp.2292-2297.

[10] Aditya Nawani, Himanshu Gupta, Narina Thakur, "Prediction of Market Capital for Trading Firms through Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 70– No.18, May 2013

[11] Hazem M. El-Bakry, and Wael A. Awad, "Fast Forecasting of Stock Market Prices by using New High Speed Time Delay Neural Networks", International Journal of Computer and Information Engineering 4:2 2010. Pp 138-144.

[12] D. Venugopal Setty, T.M.Rangaswamy, K.N.Subramanya."A Review on Data Mining Applications to the Performance of Stock Marketing". International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 3.

[13] Yusuf Perwej and Asif Perwej, "Prediction of the Bombay Stock Exchange (BSE) Market Returns Using Artificial Neural Network and Genetic Algorithm", Journal of Intelligent Learning Systems and Applications, 4, 108-119, 2012

[14] Hajizadeh E., Ardakani H., and Shahrabi J., "Application of data mining techniques in stock markets: A survey", Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010.

[15] R. Wang. "Stock Selection Based on Data Clustering Method". In 2011, 7th International Conference on Computational Intelligence and Security, 2011.

[16] Web Source - "Yahoo Finance". http://finance.yahoo.com/

[17] Web Source - "NSE India"- http://www.nse-india.com/products/content/all_daily_reports.htm

[18] Web Source - "Investopedia" - http://www.investopedia.com/active-trading/technical-indicators/