Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 10, October-2019

Big Data Analysis in Stock Market Prediction

Prit Modi
Dept. of Computer Engineering,
Institute of Technology,
Nirma University,
Ahmedabad, India

Shaival Shah
Dept. of Computer Engineering,
Institute of Technology,
Nirma University,
Ahmedabad, India

Himani Shah
Dept. of Computer Engineering, Institute of Technology,
Nirma University,
Ahmedabad, India

***Abstract*— Big data analytics can be used in many domains for accurate prediction and analysis of the large amount of data. They facilitate the discovery of significant information from large data, which is hidden otherwise. In this paper, we describe an approach for analysis of the stock market to understand its volatile nature and predict its behavior to make profits by investing in it. We first provide literature survey of past works on this domain. Then we provide a methodology of our approach which contains data collection and machine learning algorithms.**

Keywords—Stock market prediction, Apache Hadoop, Apache Spark, NSE, machine learning, Support Vector machine

I. INTRODUCTION

A financial exchange has two essential functionalities: First is to encourage the procedure for the organizations by methods for which they can exchange. The second is to sort out and deal with the setting, where exchange can appropriately happen. Contributing and benefitting from the market has never been basic, and that is because of clear vulnerability and highly unpredictable nature of the market i.e. shares/values can possibly improve and fall in value quickly. Instability is a factual proportion of the scattering of profits for a given security or market file. Usually, the higher the unpredictability, the more hazardous the security. Recorded instability likewise 'known unpredictability' is the instability of genuine costs of basic stocks. They have demonstrated to be most testing yet fulfilling and advantageous for venture.

There are different examinations on the conduct of the market. Specifically, subordinates, for example, fates and alternatives have taken a lot of considerations, recently. Anticipating these subordinates isn't significant for the hazard the board purposes. Other than that precise expectation of the market's course can assist financial specialists with gaining huge benefits with a limited quantity of capital. Financial exchange forecast can be seen as a difficult time-dependent expectation [10]. There are numerous variables that are persuasive on the monetary markets, including political occasions, cataclysmic events, financial conditions, etc. In spite of the multifaceted nature of the developments in market costs, market conduct isn't totally arbitrary. Rather, it is represented by a very nonlinear dynamical framework.

Predicting the future costs is completed dependent on the specialized investigation, which concentrates the market's activity utilizing past costs and the other market data. The market examination is in logical inconsistency with the Efficient Market Hypothesis (EMH). EMH was created in 1970 by financial expert Eugene Fama whose theory stated that it is impossible for an investor to outperform the market as all the available information is already there in the stock prices. If the EMH was true, it would be impossible to use machine learning methods for market prediction. Nevertheless, there are many successful technical analyses in the financial world and the number of studies appearing in academic literature that are using machine learning techniques for market prediction [9].

II. LITERATURE REVIEW AND RELATED WORKS

In one of the approaches, the first fetch the data and then pre-process it, after that it is transformed from high-frequency data to a ratio matrix and then the outlier algorithm finds the anomalies in it. Then the predictions are made based on the position of the anomalies and the result is evaluated. The evaluations on real exchange data show that this approach is more effective in predicting than the other traditional data mining algorithms. [1]

To make a profit by investing in the stock market there is a need for intensive planning due to various uncertainties and its volatile nature. One way to analyze trends of the stock market and to make a rewarding investment is to consider the historical volatility of stocks. One of the proposed ways is to collect historical data of 8 years (2009-2016) from the NSE website and also present data of stocks and indexes. The closing price of the stock is considered in this approach. After collecting data, it is arranged and processed using Apache Hive. After that standard deviation was calculated for the quarter, 4 – year, 8 – year period and was compared with Nifty 50 index and the stock with greater standard deviation was picked. Based on this analysis for quarterly and year wise returns on stocks, consistent stocks and promising companies can be figured out which was verified using current data on stocks. [3]

Another way to help investors is to determine when the low and high prices of stocks occur so as to figure out when to buy and sell stocks. For this one of the proposed solutions is to use a feed-forward neural network. ReLu was used as an activation function and an adaptive moment estimation optimizer was used. The data was collected for 9 years, was pre-processed and divided into training and test dataset. Then it was fed into various models like neural networks, ARIMA model, Support vector machines, Multi-layer perceptron model, etc. after that results of various methods were compared with actual data by calculating the mean error, root-mean-square error, etc. Is was found out that the feed-forward neural networks gave the best accuracy for opening the price of a stock. [4]

Big Data investigation is utilized essentially in different divisions for exact prediction and examination of the huge data sets. They permit the revelation of critical data from huge informational indexes. In this paper, a methodology on Cloudera-Hadoop based information pipeline is proposed to perform investigations for any scale and kind of information, in which US stocks are examined to foresee every day increases dependent on continuous information from Yahoo Finance. The Apache Hadoop Big-Data Framework is used to deal with enormous informational collections through disseminated stockpiling and preparing, stocks from the US financial exchange are picked and their everyday gain information is isolated into preparing and test collection to anticipate the stocks with high day by day pickups utilizing Machine Learning module of Spark. [5]

Another proposed way was to analyze financial news and social media data to build a prediction model that uses big data processing techniques, machine learning and social media analytics for predicting stock market trends. it shows that sentiment analysis facilitates various analysis methods. Using social media contents with numeric data helps the quality of input and gives better predictions [6]

One of the proposed ways for short term prediction is to use a method based on hierarchical clustering, stepwise regression and ANN model for determining similar historical patterns for stocks and to predict daily stock price by optimal significant variables using feature selection. also, the processing is done using a big data framework based on R and Hadoop. and the accuracy is determined using RMSE values of stock items.[8]

III. METHODOLOGY

One of the methodologies for stock market prediction is described in this section. For Stock market prediction using event-based supervised learning one of the novel approaches as suggested in the literature is: - Choosing the significant event criteria followed by choosing appropriate news based on the chosen significant event criteria. Then, assign an appropriate label for each news based on its associated event (for example, positive for a rise or negative for fall) and train a classifier on the labeled tweets. Predict sentiment of new future news and aggregate tweet sentiments. And at last, take a long/short position based on the net aggregated sentiment [11].

A. *Data Collection*

One of the most significant steps for predicting the behavior of the stock market is data collection. For this purpose, two sets of data are used: 1) the daily stock market information, and 2) the earnings calendar data.

Daily stock market information can be collected from various websites like www.google.com/finance and finance.yahoo.com. Also, there are various APIs which can be used to fetch the historical data of market information.

Earning Calendar data can be obtained from websites www.nasdaq.com and finance.yahoo.com which provides historical data for various companies, but there are no publicly available API to collect the earnings announcements data.

The data are collected for the period of time between Jan. 1st, 2009 to Nov. 1st, 2014 and are stored in structured databases. Our objective is to find the sign of the jump of stock price after the earnings announcement. Various companies announce their earnings at different times of the day (either before the market opens or after the market closes). In order to calculate the price rise right after the earnings announcement, we also need to have the announcement. For example, Apple Inc. announced its last earnings on October 20th 2014 after the market was closed. Therefore, the true jump for our consideration is the difference between the opening price on October 21st and the closing price on October 20th.

B. *Feature Selection*

Having collected the big data set for stock prices and earnings amounts, many numerical features can be defined out of it. For every company and for each earnings amount, we consider all the data from a year before the announcement date. This brings about one learning example data. Each company creates several learning examples (based on the number of earnings announcements that are available). For every example we have 54 numerical features. To ensure that most useful features are selected from all these 54 features, scores were assigned to each feature on the basis of the absolute value of the correlation between features and the objective and filter were used.

Among those features some of the most important features are Surprise factor, earning per stock and the difference between previous ESP, Market Cap, and Earning Jump and some operations on EPS and Market Cap. It also includes the Standard Deviation of the last 90 days of data.

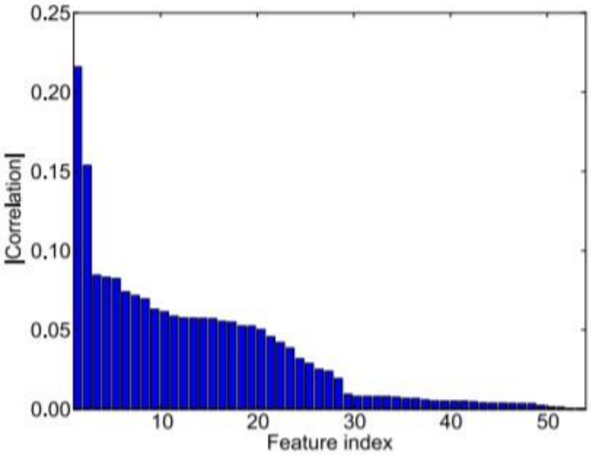

Fig. 1. Scores of different features

C. *Machine Learning Algorithms*

The data is prepared into two sets. The first set which is used as the training set includes all the data prior to April 1st, 2014, and the second one that is used as the test set consists of the data from April 1st, 2014 to November 1st 2014. Basically, this is to do cross-validation with about 85%

training set and 15% test set. We are going to investigate different machine learning algorithms for our prediction goal.

1) Logistic regression with regularization

Firstly, implementation was done using Logistic regression by using regularization. Here, goal is to find a vector θ that maximizes the log-likelihood function.

For this, stochastic gradient descent algorithm is used[12]:

$$y = [\theta \, |x > 0] \qquad (1)$$

$$\theta := (1 - \gamma)\theta + \alpha(y\ (i) - h(\theta\ Tx\ (i)\ ))x\ (i) \qquad (2)$$

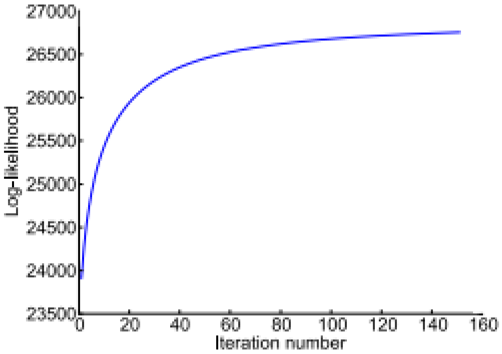

Fig. 2. Convergence of the stochastic gradient descent method for maximization of the log-likelihood function for the logistic regression method.

The convergence plot for the stochastic gradient descent (SGD) algorithm is shown in Fig. 2. The obtained training error is 37.8% and test error is 36.4%. Predicting the market movement is an extremely tough and challenging job, and the observed accuracy is in fact quite reasonable. Previously, many other works have obtained similar or lesser accuracy. To verify the results of the SGD further, the Newton’s method to obtain the θ of logistic regression is used. The runtime of the Newton’s method was nearly the same as SGD for the current size of feature vector (10 dimensions + 1 for interception). The training error was 37.8% and the test error was 36.1% [12].

2) SVM with different kernels

There are many similar works that have used Support Vector Machines (SVMs) to obtain the right balance between the VC-confidence interval and empirical error. It was observed that the results are better on generalization performance compared to the logistic regression method. So, SVM with l2-regularization and different kernels is implemented. The regularization factor in SVM provides a trade-off between variance and bias. It is observed that SVM can completely overfit the data whereas the test error will increase for the large values of the coefficient C.

IV. CONCLUSION

In general, the task of stock market prediction is quite challenging, and achieving very high accuracy is not possible. Nonetheless, machine learning techniques can provide reasonable market movement predictions, that can be utilized by investors. The calculated results show that using support vector machines with Gaussian kernel and regularization have a better performance than the logistic regression and SVM with other kernels.

Also, there are different methods to find the most accurate model for prediction of prices of the stock. The number of neural network nodes can be increased and using Neural Networks we can achieve better results. Reviewing various methods, we have found out that the Feed Forward Neural network gives the highest accuracy for the opening price of the stock. We have also observed that various methods can be efficient depending on the types of stocks and their prices.

REFERENCES

[1] L. Zhao and L. Wang, “Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm,” in 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 93–98.

[2] M.D. Jaweed and J. Jebathangam, “Analysis of stock market by using Big Data Processing Environment” in International Journal of Pure and Applied Mathematics, Volume 119

[3] S. Tiwari, A. Bharadwaj, and S. Gupta, “Stock price prediction using data analytics,” in 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, 2017, pp. 1–5

[4] P. Singh and A. Thakral, “Stock market: Statistical analysis of its indexes and its constituents,” in 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bangalore, 2017, pp. 962–966.

[5] Z. Peng, “Stocks Analysis and Prediction Using Big Data Analytics,” in 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 2019, pp. 309–312.

[6] G. V. Attigeri, Manohara Pai M M, R. M. Pai, and A. Nayak, “Stock market prediction: A big data approach,” in TENCON 2015 - 2015 IEEE Region 10 Conference, Macao, 2015, pp. 1–5.

[7] W.-Y. Huang, A.-P. Chen, Y.-H. Hsu, H.-Y. Chang, and M.-W. Tsai, “Applying Market Profile Theory to Analyze Financial Big Data and Discover Financial Market Trading Behavior - A Case Study of Taiwan Futures Market,” in 2016 7th International Conference on Cloud Computing and Big Data (CCBD), Macau, China, 2016, pp. 166–169.

[8] S. Jeon, B. Hong, J. Kim, and H. Lee, “Stock Price Prediction based on Stock Big Data and Pattern Graph Analysis:,” in Proceedings of the International Conference on Internet of Things and Big Data, Rome, Italy, 2016, pp. 223–231.

[9] R. Choudhry and K. Garg, “A Hybrid Machine Learning System for Stock Market Forecasting,” vol. 2, no. 3, p. 4, 2008.

[10] K. Kim, “Financial time series forecasting using support vector machines,” Neurocomputing, vol. 55, no. 1–2, pp. 307–319, Sep. 2003.

[11] M. Makrehchi, S. Shah, and W. Liao, “Stock Prediction Using Event-Based Sentiment Analysis,” in 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 2013, pp. 337–342.

[12] H. Pouransari and H. Chalabi, “Event-based stock market prediction,” p. 5.