International Journal of Recent Advances in Multidisciplinary Topics
Volume 2, Issue 4, April 2021
https://www.ijramt.com | ISSN (Online): 2582-7839

Stock Market Trend Prediction Using Hybrid Machine Learning Algorithms

A. Menaka[1], V. Raghu[2*], B. J. Dhanush[3], M. Devaraju[4], M. Arun Kumar[5]
[1,2,3,4,5]*Department of Information Technology, Adhiyamaan College of Engineering, Hosur, India*

***Abstract*: Machine learning and feature extractions are playing very vital role in health sector and internet sectors also. Stock market prediction is a major exertion in the field of finance and establishing businesses. Stock market is totally uncertain as the prices of stocks keep fluctuating on a daily basis because of numerous factors that influence it. One of the traditional ways of predicting stock prices was by using only historical data. But with time it was observed that other factors such as peoples' sentiments and other news events occurring in and around the country affect the stock market, for e.g. national elections, natural calamity etc. Investors in the stock market seek to maximize their profits for which they require tools to analyze the prices and trend of various stocks. Machine learning algorithms have been used to devise new techniques to build prediction models that can forecast the prices of stock and tell about the market trend with good accuracy. Many prediction models have been proposed to incorporate all the major factors affecting the price of stocks. This paper focuses on portraying distinct machine learning algorithms such as support vector machine random forest boosted decision trees assemble methods and few hybrid method, which have been used to build prediction model and predict the stock prices for different stock exchanges. This paper also covers the various challenges that are encountered while building prediction models. In this case study by using simple machine learning algorithm linear regression we predict the stock market closing amount price. In this case we used 7 different types of real-time datasets also.**

***Keywords*: Linear Regression, Support Vector Machine, Logistic regression.**

1. Introduction

Over the years, stock market has played a vital role in the prosperity of many businesses and also in the GDP of a country. As stock market is too uncertain, there is no surety that the investments made in the market would bear some profits rather it may incur some losses as well. As a part of the economic liberalization, the stock markets have been given the most important place in the financial schemes of the global corporate sector. Many factors have been found out that affect the stock prices out of which the historical data has been the most prominent one. However, it was observed that solely historical data does not give the predictions accurately. Hence more factors were identified which came out to be affecting the stock prices significantly, these were people's sentiments and news events. Thus, in addition to historical data financial news and people's reviews become major sources of such information that help in designing good prediction models that can predict the stock market prices with improved accuracy than their predecessors. However, devising such models is not an easy task. Prediction using solely historical data was also not an easy task as it involved selecting the most essentials features from the large datasets and then using pre-processing techniques on them to filter out the required data as per the specifications of the devised model. Now with the inclusion of other sentiment data, the task becomes even more difficult but the results have improved significantly. With the evolution of computer science, various new disciplines came into existence which provided better prediction models. One such discipline of computer science is Machine Learning. Over the years, machine learning has played a vital role in predictions. Predictions like workload management in cloud, heart disease prediction, house rent price prediction stock market price prediction etc. were now possible with various techniques of machine learning. It helped in building new and improvised prediction models, which gave better results with lesser complexity. In context with stock market prediction, many researchers have been able to devise models for stock market prediction which uses various techniques of machine learning such as SVM (Support Vector Machine), Linear Regressions. This paper also discusses the challenges that are faced or can be faced by researchers while devising prediction models.

2. Literature Survey

1) Title-Forecasting with artificial neural networks: the state of the art

Interest in using artificial neural networks (ANNs) for forecasting has led to a tremendous surge in research activities in the past decade. While ANNs provide a great deal of promise, they also embody much uncertainty. Researchers to date are still not certain about the effect of key factors on forecasting performance of ANNs. This paper presents a state-of-the-art survey of ANN applications in forecasting. Our purpose is to provide (1) a synthesis of published research in this area, (2) insights on ANN modelling issues, and (3) the future research directions.

2) Title: Predicting Stock Market Trends using random forest: a sample of the zagreb stock exchange

Stock market prediction is considered to be a challenging task for both investors and researchers, due to its profitability and intricate complexity. Highly accurate stock market

*Corresponding author: raghureddyrvm@gmail.com

predictive models are very often the basis for the construction of algorithms used in automated trading. In this paper, 5-days-ahead and 10-days-ahead predictive models are built using the random forests algorithm. The models are built on the historical data of the CROBEX index and on a few companies listed at the Zagreb Stock Exchange from various sectors. Several technical indicators, popular in quantitative analysis of stock markets, are selected as model inputs. The proposed method is empirically evaluated using stratified 10-fold cross-validation, achieving an average classification accuracy of 76.5% for 5-days-ahead models and 80.8% for 10-days-ahead models.

3) Title-Machine Learning in stock price trend forecasting

Predicting the stock price trend by interpreting the seemly chaotic market data has always been an attractive topic to both investors and researchers. Among those popular methods that have been employed, Machine Learning techniques are very popular due to the capacity of identifying stock trend from massive amounts of data that capture the underlying stock price dynamics. In this project, we applied supervised learning methods to stock price trend forecasting. According to market efficiency theory, US stock market is semi-strong efficient market, which means all public information is calculated into a stock's current share price, meaning that neither fundamental nor technical analysis can be used to achieve superior gains in a short-term (a day or a week). Indeed, our initial next-day predication has very low accuracy around 50%. However, as we tried to predict long-term stock price trend, our models achieved a high accuracy (79%). Based on our prediction result, we built a trading strategy on the stock, which significantly outran the stock performance itself.

4) Title-Detecting Stock Market manipulation using supervised learning algorithms

Market manipulation remains the biggest concern of investors in today's securities market, despite fast and strict responses from regulators and exchanges to market participants that pursue such practices. The existing methods in the industry for detecting fraudulent activities in securities market rely heavily on a set of rules based on expert knowledge. The securities market has deviated from its traditional form due to new technologies and changing investment strategies in the past few years. The current securities market demands scalable machine learning algorithms supporting identification of market manipulation activities. In this paper we use supervised learning algorithms to identify suspicious transactions in relation to market manipulation in stock market. We use a case study of manipulated stocks during 2003. We adopt CART, conditional inference trees, C5.0, Random Forest, Naïve Bayes, Neural Networks, SVM and kNN for classification of manipulated samples. Empirical results show that Naïve Bayes outperform other learning methods achieving F2 measure of 53% (sensitivity and specificity are 89% and 83% respectively).

5) Title - Forecasting Stock Market Trend using prototype generation classifiers

Currently, stock price forecasting is carried out using either time series prediction methods or trend classifiers. The trend classifiers are designed to predict the behavior of stock price's movement. Recently, soft computing methods, like support vector machines, have shown promising results in the realization of this particular problem. In this paper, we apply several prototype generation classifiers to predict the trend of the NASDAQ Composite index. We demonstrate that prototype generation classifiers outperform support vector machines and neural networks considering the hit ratio of correctly predicted trend directions.

3. Proposed Methodology

In this section, the proposed build linear regression model. Linear regression is a technique used to predict the relationship between the dependent and independent variable. Relationship between the two variables is said to be deterministic if one variable can be accurately expressed by the other. Roy et al [31] proposed a modification of the least square method, the LASSO (Least Absolute Shrinkage and Selection Operator) which was based on a linear regression model. This method was able to produce sparse solutions and performed well when the numbers of features were less as compared to the number of observations. It was used to predict the future price of the chosen stock. The dataset used was Goldman Sachs Group. The model's performance was compared to ridge regression and it was found out the MAPE of LASSO (1.4726) was less, compared to ridge regression. Due to the huge amount of Tick data and to ease the manipulation of data, we have imported our data to My SQL database where sorting data is done when querying. The initial step was to replace missing ticks. Tick data have different time intervals in the data collected between ticks. This is due to data not being recorded over some time. For example, a second might have four prices recorded and other seconds might not have even one price recorded. To fill missing ticks, we look for the nearest tick data to fill our missing seconds. After importing data to our database and fill missing ticks, we group our data into one-minute time interval where we get the last tick received for each minute recorded in our data. Then, we store clean one-minute data in a new table (no weekends, no ticks outside market open time. In this case we predict the closing price by using linear regression algorithm.

4. Result and Discussion

We analyzed data that was unlabeled we did not know to what class a sample belonged (known as unsupervised learning). In contrast to this, a supervised problem deals with labelled data where are aware of the discrete classes to which each sample belongs. When we wish to predict which class a sample belongs to, we call this a classification problem. SciKit -Learn has a number of algorithms for classification, in this section we will look at the Support Vector Machine. We will work on the Wisconsin breast cancer dataset, split it into a training set and a test set, train a Support Vector Machine with a linear kernel, and test the trained model on an unseen dataset. The Support Vector Machine model should be able to predict if a new sample is malignant or benign based on the features of a new, unseen sample:

1) Regression Method

It is a form of predictive modelling technique which

investigates the relationship between a dependent (target) and independent variable (s) (predictor). To establish the possible relationship among different variables, various modes of statistical approaches are implemented, known as regression analysis. Basically, regression analysis sets up an equation to explain the significant relationship between one or more predictors and response variables and also to estimate current observations. Regression model comes under the supervised learning, where we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. Predicting prices of a house given the features of the house like size, price etc. is one of the common examples of Regression.

2) Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. It is represented by an equation:

$$\hat{y} = h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta} \cdot \mathbf{x}$$

where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s). This can be written much more concisely using a vectorized form, as shown:

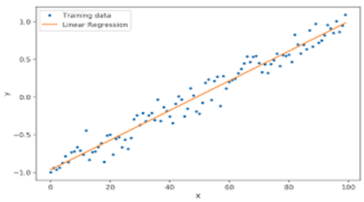

Fig. 1. Linear regression with data points

5. Screen Snippets

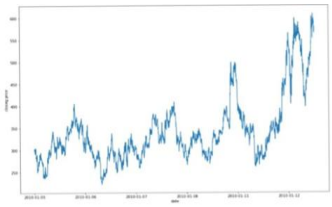

Fig. 2. Closing price vs. Date

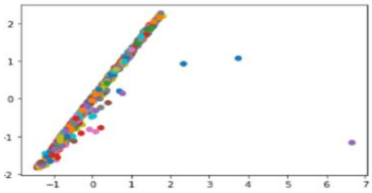

Fig. 3. Clustering Data

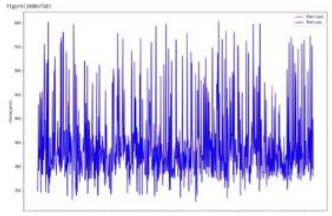

Fig. 4. Closing price vs. Date

6. Result and Discussion

A comprehensive big data analytics procedure using hybrid machine learning algorithms has been developed to forecast the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY). Ideally, researchers look to apply the simplest set of algorithms to the least amount of data, with both the most accurate forecasting results and the highest risk-adjusted profits being desired. We have also considered this standard for this research.

7. Conclusion

A comprehensive big data analytics procedure using hybrid machine learning algorithms has been developed to forecast the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY). Ideally, researchers look to apply the simplest set of algorithms to the least amount of data, with both the most accurate forecasting results and the highest risk-adjusted profits being desired. We have also considered this standard for this research.

References

[1] Chen, Z. Qiao, M. Wang, C. Wang, R. Du and H. E. Stanley, "Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market?", IEEE Access, vol. 6, pp. 48625-48633, 2018.

[2] N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction", 2017 2nd *International Conference for Convergence in Technology*, pp. 1199-1202, 2017.

[3] Q. LI, L. JIANG, P. LI and H. CHEN, "Tensor-Based Learning for Predicting Stock Movements", AAAI Conference on Artificial Intelligence, Feb. 2015.

[4] Xi Zhang, Jiawei Shi, Di Wang and Binxing Fang, "Exploiting investors social network for stock prediction in China's market", *Journal of Computational Science*, vol. 28, 2018.

[5] Thien Hai Nguyen, KiyoakiShirai and Julien Velcin, "Sentiment analysis on social media for stock movement prediction", Expert Systems with Applications, vol. 42, no. 24, pp. 9603-9611, 2015,

[6] D. Yan, G. Zhou, X. Zhao, Y. Tian and F. Yang, "Predicting stock using microblog moods", China Communications, vol. 13, no. 8, pp. 244-257, Aug. 2016.

[7] Aparna Nayak, M. M. ManoharaPai and Radhika M. Pai, "Prediction Models for Indian Stock Market", Procedia Computer Science, vol. 89, 2016.

[8] Jigar Patel, Sahil Shah, Priyank Thakkar and K Kotecha, "Predicting stock market index using fusion of machine learning techniques", Expert Systems with Applications, vol. 42, no. 4, 2015.

[9] Michel Ballings, Dirk Van den Poel, Nathalie Hespeels and Ruben Gryp, "Evaluating multiple classifiers for stock price direction prediction", Expert Systems with Applications, vol. 42, no. 20, 2015.

[10] Jitendra Kumar and Ashutosh Kumar Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution", Future Generation Computer Systems, vol. 81, pp. 41-52, 2018.