Stock Market Prediction using Machine Learning

Amit Gupta
Dept. of IT
BVCOE, New Delhi, India

Raunaq Kataria
Dept. of IT
BVCOE, New Delhi, India

Neha Gupta
Dept. of IT
BVCOE, New Delhi, India

Abstract – Stock Market has been subjected to a lot of intense research work over last couple of decades, all aimed at predicting its movement. It has been based on the ability to predict the performance of the stocks using various fundamental and technical indicators. The conventional stock market analysis is based on the application of technical indicators by intraday stock traders and fundamental indicators by chartered financial analysts to verify the state of the company and the strength of its fundamentals. These approaches although quite effective in predicting the market, are static in nature. Static algorithms fail to understand the market and stop giving appropriate predictions. In this paper, we have tried to solve this problem by using various machine learning classification techniques and algorithms. We aim to build a robust and dynamically intelligent algorithm that constantly learns from the market and gives results on the basis of changing market conditions.

Index Terms- Machine Learning, Stock Market, Logistic Regression, SVM, SMA, EMA, MACD`

I. INTRODUCTION

Stock traders and modern day market analysts complement and base their understanding of Stocks' price and performance in the near future upon various prediction tools. Prediction of stock market prices, its rise and fall of values has constantly proved to be a perilous task
mainly due to the volatile nature of the market[1-3]. Although such tool based analyses and prediction of stock trends has been there since quite a while, but the static nature of the conventional techniques used for analyses have proven to be ineffective in the longer run. These static techniques are based on the technical and fundamental market indicators that stop producing results eventually. Hence, such static algorithmic techniques needed to be replaced and revamped with a dynamic approach.

Modern day analyses of stock market are based on the state of the art Machine learning. To be specific, several studies have been conducted to predict the stock market's movement based on various machine learning techniques such as those based on Naive Bayes Theorem, Support Vector Machine (SVM), Logistic Regression and Reinforcement learning among other algorithms. However, in most of such researches, the input features in these algorithms are extracted from the data within the same market under consideration. Such consideration neglects the important data carried by other external features and hence make the prediction analysis more localized. Hence, several efforts have been done to incorporate external information through recent finance specific news or personal views such as those based on social media, which guide the basic public sentiment or opinion towards a specific stock or a company. Such approaches, covered under the umbrella name of Sentiment Analysis are based upon the attitudes and opinions of various domain experts such as renowned analysts and market experts, and are aimed to interpolate the minds of general investors. However, sentiment analysis results in a failed prediction when opinions are biased and also when the opinions are based on the results rather than results based on opinions.

Previous Researches like Shashaank,, Sruthi, Vijayalashimi and Jacob [4] focused on using the stock prices to predict the turnover for the company. In one of the similar researches to find the best technique for predicting the market Iqbal et. al.[5] rates several algorithms such as ANN, RNN and ARIMA on basis of time taken and efficiency but fails to show whether the stock beats the market or not. Some of the other researches like [6], [7] aim to predict the volume of the stock and discusses the various approaches to predict the stock price, but cannot correlate the change in stock price relative to the market index. In this paper, we examine and compare the efficiencies of machine learning algorithms like 'logistic regression' and 'Support Vector Machines' by using them to predict whether a particular stock will outperform the benchmark of the market.

The first step is to extract the OHLC (Open High Low Close) data for any stock under study along with the data of the benchmark stock, in this case, IBM and S&P 500. The value of that stock relative to that of the benchmark is calculated called the Market Relative Price (MRP). The difference between the present day's MRP and the 'n' days into the future MRP is calculated which is called Market Relative Price Change or MRPC. This is the value that is predicted, which describes how a particular stock will perform with respect to the market. If it's positive, it means stock's price will go up and if it's negative, it will go down. Next step is to build a dataset, whose columns consist of features responsible for predicting the MRPC. The different rows have the corresponding values of the features based on the day of trading. The dataset is then divided into training and test set. The machine learning algorithm is trained on the training set and a model is generated. The test set when passed through the model, gives the predicted values. In the next section, II, we discuss the various algorithms that are implemented to fit the data and predict the sentiment. Next, in III, we discuss the steps to implement a machine learning system incorporating various ML algorithms In IV, we analyse the performance of all the algorithms and decide which one performed the best. In V. we have concluded the paper.

II. ALGORITHMS USED IN RELATED WORK

Predicting the Rise or fall in the price of a stock requires separating stocks into two categories namely the stocks that gained value within certain amount of time in future or lost it. This is a classic case of a Classification Problem in Machine Learning. For testing our hypothesis and correlation we tested the data for different algorithms famous for classification. In this section we'll discuss the various algorithms used for fitting the data. First the most prominent algorithm in binomial classification is the Logistic Regression, which uses the Sigmoid function to build the model. Next we used the Support Vector Machines that use 'kernels' to perform classification task. We'll try to explain more about the mentioned algorithms and their uses.

A. Logistic Regression

One of the most widely used regression algorithms; the Logistic Regression is mainly used to train classifiers to give solutions in a range from 0 to 1 in an iterative procedure. It involves calculating a hypothesis h_θ which can be written mathematically as (1). Hypothesis is calculated using sigmoid function 'g' by running the algorithm on the Training dataset $x^{(i)}$.

$$h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) \quad (1)$$

It then predicts the particular label of the test set, using the same model. Along with predicting the labels, it also calculates the gradient θ and cost of the predicted values $J(\theta)$, represented as (2). It tries to reduce the cost of error, with each iteration. y_i is the set of given values.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \quad (2)$$

In the later section, we'll prove with the statistical data that it was one the best performing Algorithms.

B. Support Vector Machines

Kernel-based techniques (e.g. support vector machines, Bayes point machines, kernel principal component analysis (PCA), and Gaussian processes) represent a major development in machine learning algorithms. Support vector machines (SVM) is a group of supervised learning methods that can be applied to classification or regression. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik-Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis [8].

III. IMPLEMENTATION OF ALGORITHMS

In this section, we'll discuss how to predict the probability of a stock's price to go up or down. First we'll explain how the data was extracted. What is the pre-processing done in order to clean the extracted data? What are the specific features used in preparing the dataset? How are these contributing to the prediction? We'll begin with the extraction of the OHLC data.

A. Extraction of Stock Price Data

Every stock generates an Open High Low Close data every trading day. 'Open' is the price of the stock when the market opened. 'High' is the highest price that the stock achieved during that day. 'Low' is the lowest price and 'Close' is the last price the stock achieved at the end of the day, before market got closed. To extract the data, we used the 'quantmod' package to pull all the OHLC data related to a particular stock. E.g. 'IBM' in this case. We pulled data using 'getYahooData' method. We also entered the starting and the ending date, i.e. the time period of which we wish to receive the data of. We have gathered a total of 13298 days of trading data. We'll be performing two kinds of forecast. One will be the 5-day forecast and the other will be the 30-Day forecast. So, the calculation of MRPC is done accordingly.

B. Preprocessing

After getting all the data. We pass the data through a preprocessing pipeline. The pipeline is nothing but a sequence of mathematical functions functions performed on the data to calculate the factors that later will be useful in predicting the price change. These factors are also known as indicators or more specifically as Technical and Fundamental Indicators. So the Indicators we've used are as follows:

i) *Moving average convergence divergence (MACD)*: It is a trend-following momentum indicator that shows the relationship between two moving averages of prices. The MACD is calculated by subtracting the 26-day exponential moving average (EMA) from the 12-day EMA. A nine-day EMA of the MACD, called the "signal line", is then plotted on top of the MACD, functioning as a trigger for buy and sell signals.[9]

ii) Simple Moving Average (SMA): It is a simple, or arithmetic, moving average that is calculated by adding the closing price of the security for a number of time periods and then dividing this total by the number of time periods. Short-term averages respond quickly to changes in the price of the underlying, while long-term averages are slow to react. [10]

iii) Exponential Moving Average (EMA): It is a type of moving average that is similar to a simple moving average, except that more weight is given to the latest data. The exponential moving average is also known as "exponentially weighted moving average"[11]

iv) *Double Exponential Moving Average (DEMA):* It is a fast-acting moving average that is more responsive to market changes than a traditional moving average. It was developed in an attempt to create a calculation that eliminated some of the lag associated with traditional moving averages [12].

C. Model Building and Prediction

After the data has been divided, we run various Machine Learning Algorithms to fit the data. Running the algorithm over the data generates a model called the Hypothesis. The hypothesis is tested against the test set to predict the sentiment of the tweet. We used 2 different algorithms and used the following commands for model building: 'glm' for Logistic Regression and 'svm' for Support Vector Machines. After building the models, we predict the value of sentiments using the 'predict' function, which takes two parameters the model and the test data.

IV. COMPARATIVE ANALYSIS OF ALGORITHMS

In this section, a detailed description of the performance metrics of the different models is given. We'll first elaborate on the importance and meaning of parameters used in Table I. The accuracy tells us about the percentage number of times a particular algorithm predicted the correct result. More the accuracy better the algorithm. P-value is the probability of observing a test **statistic** that is as extreme as or more extreme than currently observed assuming that the null hypothesis is true [13]. Lower the P-value, better the prediction. The k-value of kappa value is the measure of correlation of variables. A high correlation means better relationship between the variables. Sensitivity is measure of True Negatives in percentage. Specificity is the measure of True Positives. The best performances are as follows. The most accurate is the Logistic Regression, getting 63% times the right answer. The lowest P-Value was for SVM. SVM had the best k-value of 0.044.

TABLE I. Comparison of 5-Day Forcast Performance of Algorithms

Algorithm/ Parameters	**GLM (size: 13000 approx)**	**GLM (size: 1000)**	**SVM (size: 13000 approx)**	**SVM (size: 1000)**
Accuracy	0.5309	0.6333	0.5457	0.6267
P-Value	1	1	0.003344	1
Kappa	0.0606	0.1668	0.0911	0.1673
Sensitivity	0.5297	0.5714	0.5473	0.5467
Specificity	0.5326	0.6498	0.544	0.5467
Negative Predicted Value	0.6156	0.3025	0.5694	0.3445
Positive Predicted Value	0.4449	0.8508	0.5217	0.8122

V. CONCLUSION

The final conclusion is that according to the statistics Logistic Regression proved to be the best algorithm for the purpose of predicting stock market. By running the algorithms and getting around 63% overall accuracy in GLM and around 85% in predicted positive value in GLM, we were able to prove that numerical data which doesn't seem to take sides, is actually biased, and can predict stock market.

REFERENCES

[1] Abhishek Gupta, Dr. Samidha , D. Sharma - "Clustering-Classification Based Prediction of Stock Market Future Prediction" - (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5(3) , 2014, 2806-2809.

[2] Dharamveer , Beerendra , Jitendra Kumar –" Efficient Prediction of Close Value using Genetic algorithm based horizontal partition decision tree in Stock Market " Volume 2, Issue 1, January 2014 International Journal of Advance Research in Computer Science and Management Studies Research Paper Available online at: www.ijarcsms.com.

[3] Kannan, K. Senthamarai, et al. "Financial stock market forecast using data mining techniques." Proceedings of the International Multiconference of Engineers and computer scientists. Vol. 1. 2010.
[4] Shashaank D.S1, Sruthi.V2, Vijayalashimi M.L.S3 and Shomona Garcia Jacob "Turnover Prediction Of Shares Using Data Mining Techniques: A Case Study "
[5] Zahid Iqbal, et al." Efficient Machine Learning Techniques for Stock Market Prediction." Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.855-867
[6] Gharehchopogh et al. "A Linear Regression Approach To Prediction Of Stock Market Trading Volume: A Case Study" International Journal of Managing Value and Supply Chains (IJMVSC) Vol.4, No. 3, September 2013
[7] Agrawal et al" State-of-the-Art in Stock Prediction Techniques", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 4, April 2013
[8] "SVM - Support Vector Machines", Available: http://www.support-vector-machines.org/
[9] "Moving Average Convergence Divergence - MACD", Available:http://www.investopedia.com/terms/m/macd.asp
[10] "Simple Moving Average (SMA)", Available: http://www.investopedia.com/terms/s/sma.asp
[11] "Exponential Moving Average - EMA)", Available: http://www.investopedia.com/terms/e/ema.asp
[12] "Double Exponential Moving Average - DEMA)", Available: http://www.investopedia.com/terms/e/dema.asp
[13] Bud Gerstman, "P-Value", Available: http://www.sjsu.edu/faculty/gerstman/EpiInfo/pvalue.htm