Forecasting Stock Market Price Using Multiple Machine Learning Technique

Md.Tanvir Rahman
Electrical and Electronic Engineering
East Delta University
Chittagong, Bangladesh
Email:Tanviredu2018@gmail.com

Rafia Akhter
Electrical and Computer Engineering
University of Georgia
Athens,Georgia,USA
Email:akhter.rafia1@gmail.com

***Abstract*—The stock market is an emerging sector in any country in the world. Many people are directly related to this sector. Stock market prediction is the act of trying to determine the future value of company stock or another financial instrument. When publicly traded, companies issue shares of stock to investors, every one of those shares is assigned monetary value or price. Stock prices can go up or down depending on different factors. Stock prices can be affected by several things including volatility in the market, current economic conditions, and the popularity of the company. The successful prediction of a stock's future price could yield a significant profit. Along with the development of the stock market, forecasting has become an important topic. Since the finance market has become more and more competitive, stock price prediction has been a hot research topic in the past few decades. Predicting stock price is regarded as a challenging task because the stock market is essentially non-linear, on-parametric, noisy, and a chaotic system. The trend of a market depends on many things like liquid money human behavior, news related to the stock market, etc. All this together controls the behavior of trends in a stock market with the advancement of the computing technology we use machine learning techniques, like Support Vector Regression, K-nearest-neighbor, liner Regression, Random Forest Regression, for analyzing time-series data to predict stock price. In this paper, we try to develop a forecasting model by stacking multiple methods to find the best forecast of the stock price.**

Keywords— Stock market, Regression, machine-learning, Stacked generalization

I. Introduction

In the business and economic environment, it is very important to predict various kinds of financial variables to develop proper strategies and avoid the risk of potentially large losses. The forecast for a variety of economic indices has a profound impact on the development of the economy. Especially in the case of stock markets, the task becomes more important because of the dynamic change of market behavior and immeasurable economic benefits. According to the prediction of stock market indices, risk managers and practitioners can realize whether their portfolio will decline in the future and they may want to sell it before it becomes depreciated. Therefore, the research of predicting the future trends of financial indices is significant and necessary for people who are interested in the stock markets.

However, the behavior of stock markets depends on many factors such as political, economic, natural factors, and many others. The stock markets are dynamic and exhibit wide variation, and the prediction of the stock market is a highly challenging task due to the highly nonlinear nature and complex dimensionality. Time series forecasting is the basic study to analyze data processes over a while. This is a series of statistical observations recorded over time series. It can be used to realize past behavior of the series and based on past behavior it can forecast future behavior of the series. The target of sales forecasting is to help the organization to determine the demands of products and improve their strategy for the future. Our purpose is to create a stock price prediction model for various international companies. The resulting model is intended to be used as a decision support tool or as autonomous tools that predict the future value of the stock prices by analyzing the previous stock price data.

This study seeks the goal is to take time-series data, find the equation that best fits the data, and be able to forecast out a specific value. Time series data is a continuous data statistical observations recorded over a specific period of time. This model will try to understand the pattern of continuous data by combining different methods and produce the best fit line that fits the data. The target is to determine the future stock price and improve their strategy for future. Regression models are the most known models used in the machine learning community and recently many researchers have examined their sufficiency in bagging [1]. Although many methods of ensemble design have been proposed, there is as yet no obvious picture of which method is best. One notable successful adoption of ensemble learning is the distributed scenario. In this work, we propose an efficient distributed method that uses the same training set with the parallel usage of an averaging methodology that combines Linear regression model , KNN regression model , Support Vector Regression [2], Random Forest Regression. We performed a Stacked generalization method for stacking the output of individual estimator and use a base regressor to compute the final prediction and the performance of the proposed method was better in most cases. we expect to obtain better results because both theory and experiments show that stacking helps most if the errors in the individual regression models are not positively correlated.

The paper is organized as follows. Section 2 describes literature reviews on the stock price prediction. .In section 3 contains the Proposed methodology. The Eexperimental design is elaborated in section 4, Section 5 contains Implementation of machine Learning Algorithm, Section 6 contains Evaluation and performance analysis. Section 7 contains a conclusion and suggested future work

II. Literature Review

Many algorithms of data mining have been proposed to predict stock price. Neural Network, Genetic Algorithm, Decision Tree and Fuzzy systems are widely used. Pattern discovery is beneficial for stock market prediction and public sentiment is also related to predicting stock price. There is a certain correlation between them. Previous studies on stock price forecasting show the use of technical indicators with artificial neural networks (ANN) for stock market prediction One of the well-researched and most important algorithm in the field of Data mining is Association Rule Mining (ARM), Decision trees are excellent for making financial decisions

In [3]. Y. Yoon and G. Swales predict the stock market data based on Neural network Approach. They take both the quantitative and qualitative variables for decision making. They use a four-layer deep neural network. Their Neural network shows a higher performance 77.5% than the traditional MDA model but it has some limitation in explaining the importance of the input parameter. The hidden unit is useful to extract the feature but it makes difficult to separate the contribution of the input parameter to the output value

In [4]. Ping-Feng Pai, Chih-Sheng Lin, used the autoregressive integrated moving average (ARIMA) model for time series forecasting. ARIMA model can’t easily capture the nonlinear patterns so they Support vector machine (SVM) and neural net to capture the non-linear pattern of the time series data

In [5]. K. Mohan use Deep neural network approach for multivariate time series analysis. They obtain a very close fit during training and the model outperformed other model during prediction

.

III. methodology

For stock price prediction we use four techniques of the machine learning algorithm. K – Nearest Neighbor, Linear regression, Support Vector Regression, and Random Forest Regression.

The prediction system has a two-tier architecture top tier is dedicated to preparing the data set from multiple information sources. The data we take is time-series data.

Time series data are taken by a variable over time (such as daily sales revenue, weekly orders, monthly overheads, yearly income, daily stock prices and tabulated or plotted as chronologically ordered numbers or data points. there are two fundamental ways, how time-series data are recorded. The first way, values are measured just for the specific time stamps, what may occur periodically, or occasionally according to concrete conditions, but anyway, the result will be a discrete set of values, formally called discrete time series. This is a very common case and frequently observed in practice.

In the economic sector, most of the indicators are measured periodically with specific periods, therefore economic indicators represent an appropriate example of discrete time series. The second option is, that data is measured and recorded continuously along with the time intervals. Electrical signals from sensors, various indicators from medicine, like ECG, or any other scientific sensors, they all represent a continuous measurement of the corresponding physical quantity. This kind of process produces a continuous time series. To yield valid statistical inferences, these values must be repeatedly measured, often over a four to five-year period.

Timeseries consist of four components:

- Seasonal variations that repeat over a specific period such as a day, week, month, season, etc.
- Trend variations that move up or down in a reasonably predictable pattern,
- Cyclical variations that correspond with business or economic 'boom-bust cycles or follow their peculiar cycles
- Random variations that do not fall under any of the above three classifications.

There are two main requirements of time series analysis:

- Identification of the important parameters and characteristics, which adequately describe the time series behavior.
- Identification of the best time series model.

The next tier is composed of two major parts. The first part is the data prepossessing. In this process, we process the data by adding more features and removing unnecessary features and removing the bad data and also the absence of the data. The second part is the data alignment.

The second tier is dedicated to market volatility analysis and prediction through model integration and training, which uses multiple kernel learning methodology to train the model It consists of three tasks: First, we build one regression model per source. Second, we train the model with the same data sets, then we make the Take the prediction of the individual model then We take the data as the input of the final Regression model that will give us the final prediction based on the previous four model's prediction.

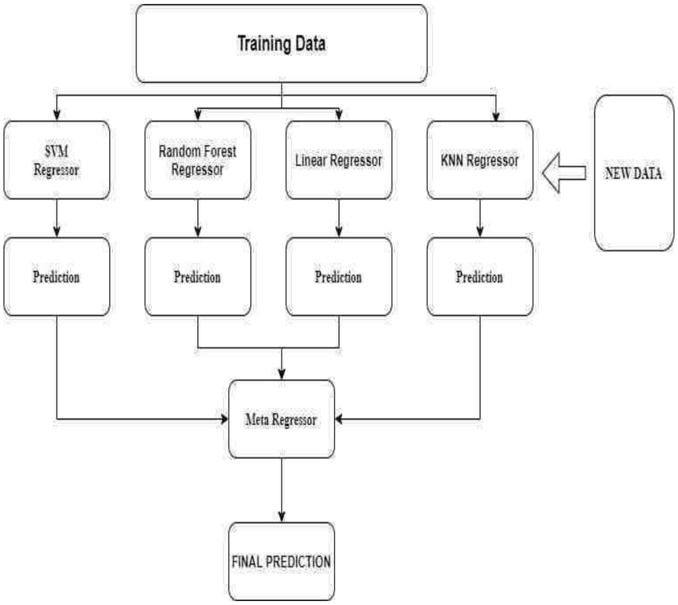


Fig. 1. Proposed Stacked Regression Model

IV. Experimental Design

A. Experimental Setup

Dataset is taken for Quandl . It is a platform for financial, economic, and alternative data that serves investment professionals. Quandl sources data from over 500 publishers. All Quandl's data are accessible via an API.API access is possible through packages for multiple programming languages including R, Python, Matlab, Maple and Stata. Quandl's sources include open data from providers such as the UN, World Bank and central banks; core financial data from providers such as CLS Group, Zacks, and ICE; and alternative data from Dun & Bradstreet, along with numerous confidential sources. There are datasets on the website which are publicly available. For example, the database of the NIKE Corporation, Intel, Microsoft, etc.

B. Data Pre Processing

In the real world, many data sets are very messy. Most stock price/volume data is pretty clean, rarely with missing data, but many data sets will have a lot of missing data. filter out other unimportant features from the feature because not all the features will be included in the final feature list. The reason behind it is the unnecessary feature and those value which has no relation with the stock market prediction will reduce the accuracy of the prediction. in our study, we used the following attributes Adjusted Close price, Volatility, Percentage change, Adjusted open price, Adjusted Volume

V. Machine Learning Implementation

In this work we used four type of machine learning algorithm for the regression task. SVM algorithm, Random Forest algorithm, Linear Regression, KNN algorithm. A detail description of four machine learning algorithm can be found in. Jupyter Notebook is used to implement all the four different algorithms also training and testing this algorithm. We also use the python programming language and we use the **Scikit-learn** is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries

1) Support Vectoe Machine:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a single category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier's detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. we are working with SVM as a regression model, we consider the points that are within the decision boundary line. Our best fit line is the hyperplane that has a maximum number of points

2) Random Forest Algorithm

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set. Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, we don't have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. we can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds

3) K-Nearest Neighbors (K-NN) Algorithm

K Nearest-Neighbor (KNN). KNN algorithms have been identified as one of the top ten most influential data mining algorithms for their ability of producing simple but powerful classifiers. It has been studied at length over the past few decades and is widely applied in many fields. The KNN rule classifies each unlabeled example by the majority label of its k-nearest neighbors in the training dataset. KNN is a non-parametric lazy learning algorithm

We give a single number "k”. This number decides how many neighbors (where neighbors is defined based on the distance metric) influence the Regression Model.

4) Linear Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, a function of the independent variables called the regression function is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data

VI. Evaluation

All the algorith run effciently and our proposed algorithm give better result in all the model based on the 8 different company’s stock market data.

TABLE I. REGRESSION MODEL ACCURACY BASED ON THE CROSS VALIDATI0ON

	Accuracy(%)				
Company	**Proposed Method**	KNN	SVR	RF	Linear Model
Google	**98.7**	*89.1*	74.3	88.6	87.9
Apple	**94.8**	*91.4*	91.0	93.4	92.2
Microsoft	**91.4**	*91.4*	91.4	91.4	91.4
IBM	**80.8**	71.3	77.8	78.4	73.6
NIKE	**96.8**	94.9	95.0	92.9	94.0
MAC DONALD	**94.7**	91.7	91.0	92.7	93.7
Walt Disney	**81.43**	69.6	77.3	75.0	74.3
Intel	**72.77**	61.3	65.7	67.36	52.6

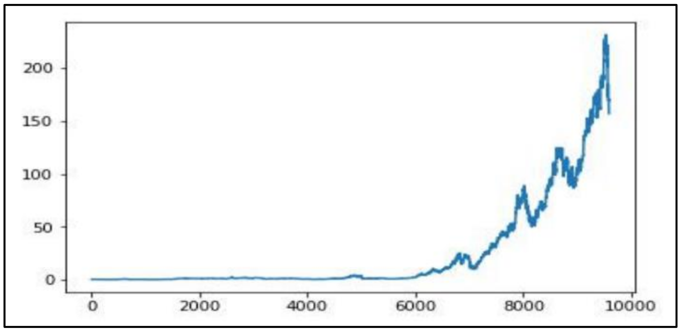


Fig. 2. Previous Stock value of apple.inc

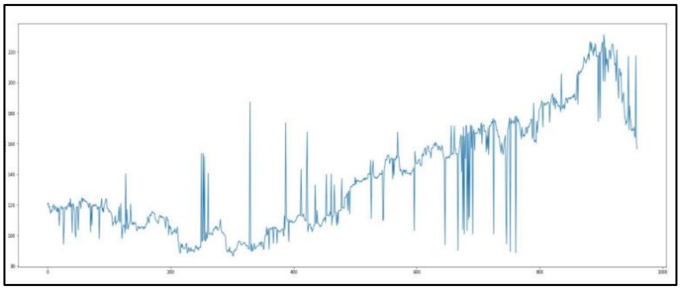


Fig. 3. Predicted stock using kNN regressor

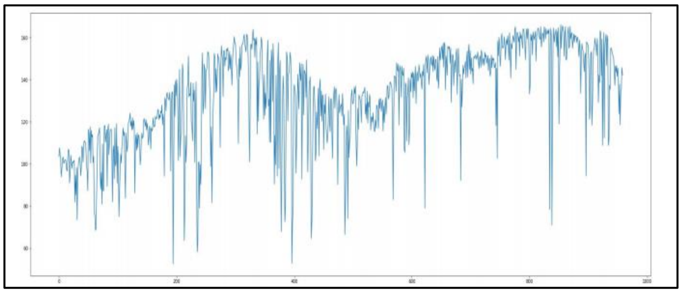
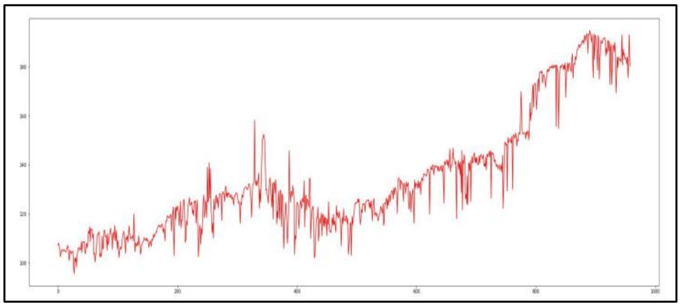


Fig. 4. Predicted stock using SVM regressor



VII. CONCLUSION

Model stacking is an efficient ensemble method in which the predictions, generated by using various machine learning algorithms, are used as inputs in a second-layer learning algorithm. This second-layer algorithm is trained to optimally combine the model predictions to form a new set of predictions. For example, when linear regression is used as second-layer modeling, it estimates these weights by minimizing the least square errors. However, the second-layer modeling is not restricted to only linear models; the relationship between the predictors can be more complex, opening the door to employing other machine learning algorithms. Ensemble methods are commonly used to boost predictive accuracy by combining the predictions of multiple machine learning models. The traditional wisdom has been to combine so-called “weak” learners. However, a more modern approach is to create an ensemble of a well-chosen collection of strong yet diverse models. Building powerful ensemble models has many parallels with building successful human teams in business, science, politics, and sports. Each team member makes a significant contribution and individual weaknesses and biases are offset by the strengths of other members. It is known that if we are only concerned for the best possible correlation coefficient, it might be difficult or impossible to find a single regression model that performs as well as a good ensemble of regression models. In this study, we built an ensemble of regression models using four different learning methods. After determining and comparing with other models. In our proposed model we have attained the highest accuracy among all others.

REFERENCES

[1] L. Breiman, Bagging Predictors. Machine Learning, 24(3),1996

[2] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995

[3] Y. Yoon and G. Swales, “Predicting Stock Price Performance: A Neural Network Approach,” Proc. 24th Hawaii Int’l Conf. System Sciences (HICSS-24), IEEE CS Press, 1991, vol. 154, pp. 156-162.

[4] Pai, P.F. and Lin, C.S. (2005), “A hybrid ARIMA and support vector machines model in stock price forecasting”, Omega, Vol. 33, pp. 497-505.

[5] Chakraborty, K., Mehrotra, K. Mohan, C. and Ranka, S., ‘Forecasting the behavior of multivariate time series using neural networks’, Neural Networks, 5 1992, pp. 961–70.

[6] Specht, D.F. (1991), “A general regression neural network”, IEEE Trans. on Neural Networks, Vol. 2 No. 6, pp. 568-76.

[7] Tsaur, R.C. (2004), “Planning and analyzing for stock investment – a study for stocks of banks”, Hsuan Chuang Management Journal, Vol. 1 No. 2, pp. 1-16

[8] Wu, M.L. (2007), SPSS Statistical Application Learning Practices, Acore Book, Tapei

[9] Elman, J., ‘Finding structure in time’, CRL Technical Center for Research in Language, University of California, San Diego, 1988

[10] Freisleben, B., ‘Stock market prediction with back- propagation networks’, in Belli, F. and Rad- emacher, J. (eds), Lecture Notes on Computer Science, Vol. 604, pp. 451–60, Springer-Verlag, Heidelberg, 1992

[11] O stermark, R., ‘Predictability of Finnish and Swedish stock returns’, OMEGA International Journal of Man agement Science, 17, No. 3 1989, pp. 223–36.

[12] Shiller, R., ‘The volatility of stock market prices’, Science, 235, No. 4784, 1987, pp. 33–7.

[13] Tan, C., ‘Trading a NYSE-stock with a simple arti- ficial neural network-based financial trading system’, in Proceedings of New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, pp. 294–5, Dunedin, 1993

[14] Virtanen, I. and Yli-Olli, P., ‘Forecasting stock marke prices in a thin security market’, OMEGA Inter national Journal of Management Science, 15, No. 2, 1987, pp. 145–55.

[15] White, H., ‘Economic prediction using neural net- works: the case of IBM daily stock return’, in Proceedings of International Conference on Neural Net- works, pp. II-451-II-458, San Diego, CA, 1988.

[16] B. LeBaron, W.B. Arthur, and R. Palmer, “Time Series Properties of an Artificial Stock Market,” J. Economic Dynamics and Control, vol. 23, nos. 9-10, 1999, pp. 1487-1516.