Ensemble Learning Approach for Enhanced Stock Prediction

Shikha Mehta, Priyanka Rana, Shivam Singh, Ankita Sharma, Parul Agarwal
Department of Computer Science and Engineering
Jaypee Institute of Information Technology, Noida, India
mehtshikha@gmail.com, Prr.1698@gmail.com,shivamsinghlps6@gmail.com,ankita.sharma.vb@gmail.com, parul.agarwal@jiit.ac.in

***Abstract*—Stock market prediction is the technique for deciding the future estimation of an organization stock or other money-related instrument exchanged on a monetary trade. Fruitful forecasts of stock market lead to high investment gains. Analysis of stock market data has always been a hot area of research due to the large number of factors affecting the stock market. In the last few years, researchers have utilized machine learning techniques for learning the trends of stock market in order to improve the accuracy of predictions. However, authors have applied these techniques individually and compared their results. Since the aggregated opinion of a group of models is relatively less noisy as compared to the single opinion of one of the models, this paper presents an ensemble machine learning approach for predicting the stock market. The weighted ensemble model is built using weighted support vector regression (SVR), Long-short term memory (LSTM) and Multiple Regression. From the results it is observed that ensemble learning approach is able to attain maximum accuracy with reduced variance and hence better predictions.**

Keywords—Long Short Term Memory; Multiple Regression; Ensemble approach; Support Vector Regression.

I. INTRODUCTION

A nation's capital market generally relies upon the width and profundity of the stock base. The financial growth of a nation to a great extent relies upon the extension and advancement of long term capital. The extraordinary renaissance seen everywhere around the world is because of offers and stocks. Enthusiasm for the buying and selling of offers isn't just appeared by huge organizations and financial specialists but also by little speculators, people, salaried individuals, settled salary gathering, etc. Offers are bought when the costs are low in the market and sold when they are high[1]. The edge is the benefit to the financial specialist. Regardless of whether there is a minor enhancement in the execution of securities exchange expectation, still incredible benefit can be accomplished.

Forecasting the trends of the stock market has always been a relevant area of research in every decade. The technological advancements, social, economic and political factors, international policies, etc. pose new challenges for financial market prediction in every era. The mushrooming usage of WWW has further escalated the issues by shortening the life cycle of products and services. Exploring time series financial data has always attracted researchers due to it's non-stationary and pseudo-chaotic in nature[1], [2]. Time series modeling is a challenging task as the means and the standard deviation of the series changes over a period of time and thus making the correlation between past and present unpredictable. In literature numerous time series techniques forecasting techniques such as Autoregressive[3], Autoregressive Moving Average, Moving Average (MA)[4], Exponential Moving Average (EMA), etc. have been applied substantially for financial time series predictions attaining good results [3]. However, their performance varied drastically with the change in the time period and sources of time series such as stocks, indexes, and currencies. With the rising complexity of the problem, conventionally time series techniques are proving inefficient to deal with the challenges of the digital era. This motivated the researchers to explore machine learning algorithms for predicting the bulls and bears and analyze the behavior of stock markets for making trade decisions.

In literature, a number of machine learning algorithms have been applied for predicting market stocks. Yu, Chen, and Zhang [5] applied support vector machine for non-linear classification of market stocks. Authors created a stock selection model and applied PCA (principal component analysis) on a financial dataset in order to extract small dimensional relevant features for successful classification. However, accuracy was achieved nearly 62% only. Moghaddam, Moghaddam, and Esfandyari[6] employed artificial neural network (ANN) for prediction of a stock market index. ANN is considered as one of the efficient classification technique, thus applied on NASDAQ stock exchange rate forecasting. Authors used a back propagation network for daily prediction of stock prices. Dash and Dash [7] developed a framework for the analysis of market stocks using machine learning algorithms. The decision for stock trading was forecasted with their ANN and decision support model and compared with SVM, Naïve Bayesian, K-Nearest neighbor, and decision tree model. Chang et. al. [8] utilized Ensemble neural network model for training and testing of stock market data. In order to determine stock turning points in stock trading, intelligent piecewise linear representation methods are used and then training of turning points was done through ensemble neural networks. Forecasting of turning points was made through this approach and was profitable as compared to other approaches. Another work on financial trading through machine learning was performed by Gerlein et. al [9]. This work is an empirical analysis describing the merits and demerits of financial trading through existing tools and techniques. However, the work

978-1-7281-3591-5/19/$31.00 ©2019 IEEE

did not propose any new method for good stock predictions. Qian and Rasheed[10] applied various machine learning classifiers such ask-nearest neighbor, decision tree and artificial neural network for forecasting stocks. Accuracy of only 65% was achieved through collaborated models. Nunno[11] performed stock market prediction through regression techniques. The work describes types of regression methods such as linear and polynomial regression. Support vector regression was found to be the most effective model for prediction of stocks in the market. However, it needs some advancement to get better results. Another work of support vector regression (SVR) for stock market prediction was made by Meesad and Rasel[2]. Data is preprocessed through different types of windowing operators and then fed to the SVR model. The model is applied to real-time data for a popular company named Dhaka stock exchange. Narayanan and Govindarajan [1]applied SVM and Naïve Bayes model on time series data i.e. stock market prediction. Authors proposed two new models i.e. AdaSVM and AdaNaive for analyzing stock market data. The performance of the proposed algorithm is compared is SVM and Naïve Bayes and it was observed that the proposed algorithm gives better efficacy as compared to existing algorithms. Tsai et. al. [12] predicted prices of various stocks through different classifiers. Financial time series forecasting is made through machine learning techniques by Sung et. al. [13]. The direction of stock prices through different classifiers is also predicted by Ballingset. al. [14]. Although, a number of classifiers have been applied for predicting prices of the stock market, yet an efficient technique assembling the benefits of the best machine learning model is missing.

Most of the machine learning techniques has attained considerable results; each individual technique has its own merits and demerits. To tackle the concerns of individual algorithms or to take benefits or advantages of all algorithms, ensemble strategies are gaining more importance[9]. Techniques such as Deep Neural Networks, SVMs, Recurrent Neural Networks, and Ensembles have become popular not only for their prediction capacity but also for the fact that they can now be actually trained in a stable and timely manner. This is due to the increase in computation power and the advent of new training methods, such as Long-Short team memory for Recurrent Neural Networks. Ensemble strategies overcome limitations of Neural Networks and classification algorithms. In this paper, the Ensemble learning algorithm has been deployed with machine learning classifier namely Support Vector Regression, Multiple Regression and Long short-term memory network (LSTM) for the purpose of stock prediction. Ensemble model presented here improves accuracy and robustness over single model methods. It also overcomes the limitations of a single hypothesis. The target function may not be implementable with individual classifiers but may be approximated by model averaging. Thus, this method is more beneficial for predicting the stocks in the market.

II. Ensemble Model for Stock Prediction

Machine learning is the specialized form of data mining in which models are learned in a supervised or unsupervised fashion. In supervised learning, a model is learned by training it over a given set of examples. Thereafter trained model is used to predict the class of the new instance. In unsupervised learning, the model finds the patterns existing in a given class. Ensemble model combines the predictions/results of various machine learning models in order to improve the overall performance of a system [15]. The basic idea behind the ensemble method is to unite the diverse perspectives obtained from different models to make better prediction quality[16]. Prediction of various machine learning models can be combined in three ways. These methods are max voting, averaging and weighted averaging[16]. In max voting methods, various machine learning classifiers make a prediction for every data point. Prediction made for each data point by each classifier is considered as one vote. The final prediction of class for a data point is the one with the majority of the votes i.e. class given by the majority of the models. This is mainly used for classification algorithms. In the averaging method, the prediction is made by taking the mean of results obtained from multiple models. It is mainly used for regression models. The weighted average method is an extension of the averaging method in which each and every model is assigned a weight which signifies the importance of a model for making the final predictions.

Ensemble approach is a machine learning method that consolidates a few base models with the end goal to create one ideal model. The ensemble has a different number of learners which are called base learners. The ability of generalization of an ensemble is generally more grounded than that of base learners. Combining learners are engaging in light of the fact that it can support powerless learners who are somewhat superior to strong learners and can make precise predictions. "Base learners" are likewise called as "powerless learners". Stock prediction is a supervised classification problem where forecasting of stock prices are made for the future. In this paper, ensemble learning model combines the decisions of three base learners namely, Support Vector Regression(SVR), Multiple Regression and Long short-term memory network (LSTM) for stock market prediction through the weighted averaging method.

Ensemble model improves accuracy and robustness over single model methods. Ensemble learners are used in this research paper because it has overcome the limitations of a single hypothesis. The target function may not be implementable with individual classifiers but may be approximated by model weighted averaging. Thus this method has been deployed for training stock prediction data. In the next subsection, a method for constructing ensemble learning and algorithm used in ensemble learning are discussed.

A. Ensemble Model Construction

Ensemble learning is a technique of combining multiple machine learning algorithms in order to yield better predictions. An ensemble model is built in two stages. In the initial step, all the base learners are formulated. Each of these learners is produced in a parallel style where the generation of a learner has an impact with respect to the other learner. In the following stage, decisions of these base learners are consolidated in two ways- majority voting and weighted averaging. The popular combination method used is majority voting for classification and weighted averaging for regression.

B. Algorithms Used

The ensemble learning algorithm is formulated using three machine learning algorithms such as Support Vector Regression (SVR), Multiple Regression and Long short-term memory network (LSTM).

1) Multiple Regression

Regression [11] is a technique for calculating target values based on the independent predictors. This technique is generally utilized for determining and finding the relation between two variables. Regression methods generally vary in two different ways, one the number of independent factors and second, kind of relationship between the independent and dependent factors. Regression consists of a number of independent variables and there exists a linear relationship between the independent(x) and dependent(y) variable known as Linear Regression.

Multiple regression is an advanced version of simple linear regression. It is used for discovering the relation between at least two continuous factors. One is predictor or independent variable and others are responses or dependent variables. It looks for a statistical relationship but not a deterministic relationship. The relation between factors is said to be deterministic if one variable can be accurately expressed in terms of others[17]. Multiple regression for 'y' dependent variable can be expressed by equation (1):

$$y = (a_1 * \beta_1) + (a_2 * \beta_2) + (a_3 * \beta_3) \ldots + (a_k * \beta_k) + \epsilon \qquad (1)$$

Variable 'y' is linearly dependent on 'k' independent variable a_1, a_2, a_3 ... a_k. Regression coefficients are expressed as β_1,β_2,β_3 ...β_k. ϵ is the tolerance error i.e. difference in observed and fitted value. The motive of the multiple regression algorithms is to find the best values for a_1, a_2, a_3 ... a_k such that the model is trained well. The model is trained in a number of epochs to best fit the model with known data.

2) Support Vector Regression(SVR)

Support Vector Machine [18][19] is likewise utilized as a regression strategy, keeping all the other features intact. The same principle of SVM is employed by Support Vector Regression (SVR), with some minor contrasts[2]. Error value yields a real number which makes it exceptionally hard to predict the outcome of current data, which has infinite possibilities. In the case of SVR, epsilon (margin of tolerance) i.e. a parameter of the regression model is adjusted with respect to SVM[20][4]. The basic motive behind SVR remains same as that of SVM i.e. maximize the margin and minimize the error reserving the tolerated error part.

In the present research work, non-linear support vector regression is used to predict the stock data. Non-linear SVM uses kernel function that reconstructs the data into space of high dimensional attributes so that linear separation can be performed. The Gaussian radial basis kernel function is more preferred over polynomial kernel function as it gives better accuracy. The Equation of Gaussian radial basis kernel function[2]is given in Equation 2.

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \qquad (2)$$

Where $\boldsymbol{\gamma}$ is parameter defining the width of bell curve, x_i and x_j are feature vector of input space.

3) Long short-term memory network(LSTM)

LSTMs are an exceptional type of recurrent neural network (RNN) that are skilled for learning dependencies of the long term given by Hochreiter & Schmidhuber[20]. These are the popular technique used in a large number of applications and is used to surpass the dependency problem of long terms[21]. This is a type of RNN which contains a long sequence repeating part of a neural network. Unlike RNN with single layer neural network in each module, LSTM has a different structure for the neural network in each module. The outcome obtained in the output block is fed as input in next iterations. LSTMs are successful in overcoming the problem of vanishing gradients.

The core of LSTM is the cell state which runs horizontally through neural network modules. LSTM consists of gates through which information flow can be regulated. These gates consist of pointwise multiplications with the sigmoid neural network layer. Sigmoid layer implies that output can be from 0 to 1. Zero means no information is allowed and one means all information is allowed to transmit from cell state.

4) Proposed Ensemble Model

The ensemble approach proposed in this paper uses a weighted averaging method as this method gives more promising results as compared to average and majority voting methods. Initially, the prices of the stocks are forecasted through the base learners i.e. Support Vector Regression, LSTM, Multiple Regression. Thereafter in the second step on the basis of accuracy obtained from base learners, weights are assigned to them. In the third step average of weighted accuracies of the base, learners are obtained as final output to the ensemble approach.

Let A1, w1 is the weight and accuracy of Multiple regression, A2, w2 is the weight and accuracy of Support vector regression and A3, w3 is the weight and accuracy of LSTM, then accuracy of ensemble model is calculated using Equation 3.

$$Accuracy = \frac{A18w1 + A2*w2 + A3*w3}{w1 + w2 + w3} \qquad (3)$$

The next section proves the performance efficacy of the proposed ensemble model against individual base learners through experimental results and analysis.

III. Experimental Results and Analysis

The performance evaluation of different machine learning algorithms i.e. base learners such as Support Vector Regression, Multiple Regression and Long short-term memory network (LSTM) is made on the basis of accuracy against the proposed ensemble model. Experiments are conducted in python spyder 3.1 version. Inbuilt libraries of python used are – Numpy, Scikit-learn, Pandas, and Keras. All techniques are implemented on a computer system with 4GB RAM and intel core i5 processor. The parameter values used in base learners are shown in Table I.

TABLE I. Parameter values of Base Learners

Algorithms	Parameters	Values
Multiple Regression	Test_size	0.1
	Random_state	4
Support Vector Regression	Kernel	Rbf
	C	Le3
	Gamma	0.1
Long Short Term Memory	Init	Uniform
	Activation	ReLu
	Loss	MSE
	Optimizer	ADAM
	Batch_size	512
	Nb_epoch	500
	Validation_Split	0.1
	Verbose	0

Yahoo Stock (1996-2016) dataset is used for evaluation of base learners as well as ensemble approach. Yahoo finance dataset has been collected from Yahoo stock data [22]. The data is collected from April 1996 to April 2016. The stock market took a toll during 2007-2008 financial crises. During this time companies went in loss and stock data of companies absolutely volatile (unpredictable). Training machine learning model using this data would cause the system to be less accurate. This is because of a lack of trend during the crisis period. So, such data is avoided that can result in uncertain behavior. Yahoo stock prediction dataset consists of 8 features and 5039 instances. Features of yahoo dataset are as follows:

1. Date
2. Date Value
3. Open
4. High
5. Low
6. Close
7. Volume
8. Adj Close

In the stock market prediction dataset, prices of stocks are defined for different dates. Fig.1, Fig. 2, Fig. 3 and Fig. 4 illustrate the line graphs that depict the prices of stock predicted from multiple regression, support vector regression, LSTM and proposed ensemble model respectively. The x-axis of the line graph denotes time in a number of days where 1 unit is 20 days. Y-axis denotes prices where 1 unit is 2 rupees. The predicted value of each learner is compared with the target/desired output value which is already defined in dataset. The desired output enables predictors (learners) to calculate its performance efficacy through accuracy. Red line in the graph denotes the prices of stocks predicted by the learner and blue line denotes the actual stock price.

The weight values in base learners are assigned on the basis of prediction accuracy obtained. Multiple regression is assigned highest weight i.e. 3 it gives better accuracy in contrast to the other two. Support vector regression is assigned 2 and long short term memory network 1 i.e. least weight. On the basis of this, the weighted average method is applied in an ensemble model proposed here.

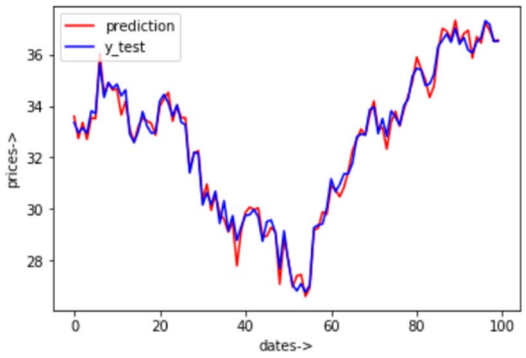

Fig.1. Predicted prices of stocks using Multiple Regression

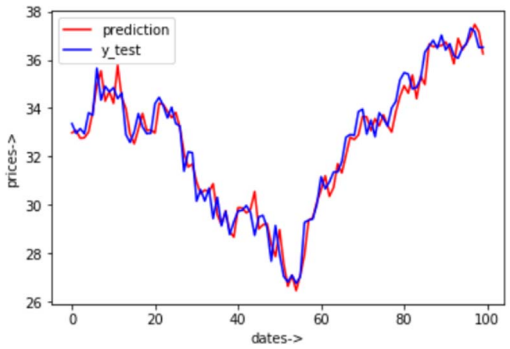

Fig.2. Predicted prices of stocks using support vector regression

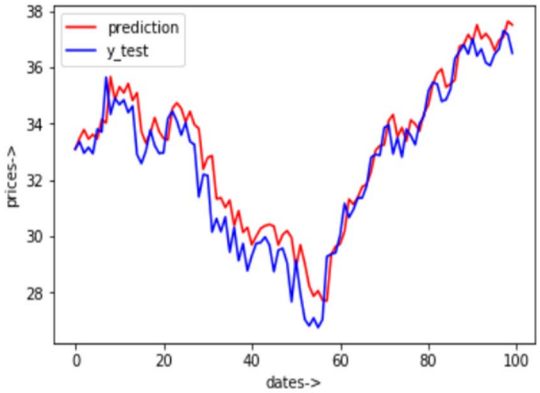

Fig.3. Predicted prices of stocks using LSTM

It can be observed from the above graphs that predicted and the actual value largely deviates in LSTM learner. In case of support vector regression, predicted results better than LSTM. However, prices of stocks predicted by multiple regression give good overlap of predicted value with actual value. Best results are obtained through our proposed ensemble model. Prices predicted by the proposed model closely mirror the actual prices of stock. It can also be noticed from plots (given in fig 1to fig 3) that peak of predicted graphs gets dilute from LSTM, SVR, multiple regression to ensemble model. This is because the proposed model takes the best of all learners. It assigns the highest weight to the learner with maximum accuracy and lowest weight to the minimum accuracy learner.

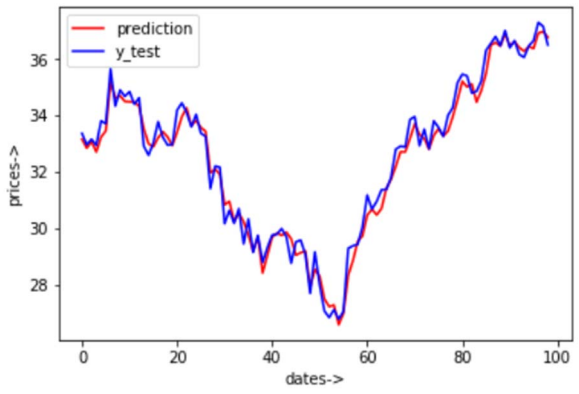

Fig.4. Predicted prices of stocks using the proposed ensemble model

Further results are established through accuracy rate obtained through predicted and actual value shown in Table II. The accuracy rate is computed by multiplying the outcomes of SVR, LSTM, Multiple Regression with the respective weights that were assigned to them on the basis of their accuracies (shown in Equation 4). The output is then divided by the sum of all weights assigned to the base learner.

$$accuracy = 100 - \left(\left(\frac{original\ array}{len(predicted\ array)} \right) * 100 \right) \quad (4)$$

TABLE II. Accuracy rate of base learners and ensemble model on stock market prediction

Algorithm	Accuracy Rate
Multiple Regression	99.02
Support Vector Regression	98.56
Long Short Term Memory Network (LSTM)	97.63
Proposed Ensemble Model	99.12

Hence it has been observed from the above results that the proposed ensemble model gives better results as compared existing base learners.

IV. Conclusion

The paper presented an ensemble model based on various machine learning models i.e. multiple regression, support vector regression and long short term memory networks. The model applies the weighted average method based on the accuracies obtained from prediction of Yahoo stock data. Experimental results depict that though accuracy rates of the algorithms are not enhanced much (because accuracy rate of Multiple regression is 99.02 and ensemble learning is 99.12), but the deviation between the actual and predicted price has significantly reduced in ensemble model as compared to any other of the three algorithms for any particular day (shown in graphs). Therefore, ensemble learning is a more appropriate model than other single models and can be further used for other classification problems.

References

[1] B. Narayanan and M. Govindarajan, "Prediction of Stock Market using Ensemble Model," Int. J. Comput. Appl., vol. 128, no. 1, pp. 18–21, 2015.

[2] P. Meesad and R. I. Rasel, "Predicting stock market price using support vector regression," 2013 Int. Conf. Informatics, Electron. Vision, ICIEV 2013, pp. 1–6, 2013.

[3] M. Asad, "Optimized Stock market prediction using ensemble learning," 9th Int. Conf. Appl. Inf. Commun. Technol. AICT 2015 - Proc., pp. 263–268, 2015.

[4] H. Qu and Y. Zhang, "A New Kernel of Support Vector Regression for Forecasting High-Frequency Stock Returns," Math. Probl. Eng., vol. 2016, pp. 1–9, 2016.

[5] H. Yu, R. Chen, and G. Zhang, "A SVM stock selection model within PCA," Procedia Comput. Sci., vol. 31, pp. 406–412, 2014.

[6] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," J. Econ. Financ. Adm. Sci., vol. 21, no. 41, pp. 89–93, 2016.

[7] R. Dash and P. K. Dash, "A hybrid stock trading framework integrating technical analysis with machine learning techniques," J. Financ. Data Sci., vol. 2, no. 1, pp. 42–57, 2016.

[8] P. C. Chang, C. H. Liu, C. Y. Fan, J. L. Lin, and C. M. Lai, "An ensemble of neural networks for stock trading decision making," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5755 LNAI, pp. 1–10, 2009.

[9] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," Expert Syst. Appl., vol. 54, pp. 193–207, 2016.

[10] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," Appl. Intell., vol. 26, no. 1, pp. 25–33, 2007.

[11] L. Nunno, "Stock Market Price Prediction Using Linear and Polynomial Regression Models," pp. 1–6, 2014.

[12] C. F. Tsai, Y. C. Lin, D. C. Yen, and Y. M. Chen, "Predicting stock returns by classifier ensembles," Appl. Soft Comput., vol. 11, no. 2, pp. 2452–2459, 2011.

[13] M.-C. Sung, T. Ma, M.-W. Hsu, J. E. V. Johnson, and S. Lessmann, "Bridging the divide in financial market forecasting: machine learners vs. financial economists," Expert Syst. Appl., vol. 61, pp. 215–234, 2016.

[14] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," Expert Syst. Appl., vol. 42, no. 20, pp. 7046–7056, 2015.

[15] C. Zhang, Y. Ma, and Editors, Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012., 2012.

[16] T. G. Dietterich, "Ensemble Methods in Machine Learning," pp. 1–15, 2007.

[17] D. Enke, M. Grauer, and N. Mehdiyev, "Stock market prediction with Multiple Regression, Fuzzy type-2 clustering and neural networks," Procedia Comput. Sci., vol. 6, pp. 201–206, 2011.

[18] "support vector Regression." [Online]. Available: https://www.saedsayad.com/support_vector_machine_reg.htm. [Accessed: 28-Apr-2019].

[19] F. E. H. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," Int. J. Manag. Sci., vol. 29, pp. 309–317, 2001.

[20] C. Lin, "Large-scale Linear Support Vector Regression," Jmlr, vol. 13, pp. 3323–3348, 2012.

[21] K. Greff, R. K. Srivastava, J. Koutn'\ik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: Search Space Odyssey," CoRR, vol. abs/1503.0, no. 10, pp. 2222–2232, 2015.

[22] "Yahoo Stock Data." [Online]. Available: https://github.com/ranapriyanka1604/Ensemble-Approach-to-Stock-Prediction/blob/master/yahoostock.csv. [Accessed: 28-Apr-2019].