Stock Market Analysis and Prediction

Saurabh Kumar Mishra[1], Mohd. Aqdas[2], Ajay Kumar[3]

(Computer Science &Engineering Department)
Institute of technology and Management Gida Gorakhpur

Abstract

In this project we tend to plan to implement machine learning approach to predict stock costs. Machine learning is effectively enforced in statement stock costs. the target is to predict the stock costs to create additional knowing and correct investment choices. we tend to propose a stock value prediction system that integrates mathematical functions, machine learning, and alternative external factors with the aim of achieving higher stock prediction accuracy and issuing profitable trades. There are a unit 2 kinds of stocks. you'll understand intraday commerce by the unremarkably used term "day commerce." Intraday traders hold securities positions from a minimum of in the future to consecutive and sometimes for many days to weeks or months. LSTMs area unit is terribly powerful in sequence prediction issues as a result of they're ready to store past info. this is often necessary in our case as a result of the previous value of a stock is crucial in predicting its future value. whereas predicting the value of a stock is associated uphill climb, we {will |we are able to} build a model which will predict whether or not the worth will go up or down.

Key Words: *LSTM, CNN, ML, DL, Trade Open, Trade Close, Trade Low, Trade High, SVM*

1.INTRODUCTION - Data analysis is employed in all business for data-driven deciding. In share market, there are a unit several factors that drive the share value, and therefore the pattern of the amendment of value isn't regular. this can be why it's powerful to require a strong call on future value. Artificial Neural Network (ANN) has the aptitude to find out from the past information and build the choice over future. Deep learning networks like Convolutional Neural Network (CNN), continual Neural Network (RNN) etc. works nice with variable statistic information. we have a tendency to train our model from the past stock information and calculate the longer-term value of that stock. This future value use to calculate the longer-term growth of a corporation. Moreover, we found a future growth curve from completely different firms. therefore, we will analyse and investigate the similarity of 1 company's future curve over another. Stock value of a listed company in {an exceedingly in a very} stock market varies when an order is placed for sell or obtain and a dealing completes. AN exchange collects all sell bids with expected value per stock (normally it's quite {the value| the worth| the value} paid whereas bought by the capitalist) and every one obtain bids with or while not a value limit (normally AN investor expects the longer-term value of the stock are quite this price he's paying now) and a buy-sell dealing is committed once each bid have a match i.e. marketing damage is same with shopping for damage of some buy-bid Fame in 1970 [1] projected economical market hypothesis that says that in an economical market (where all events area unit far-famed {to all | to all or Any | to any or all} stakeholders as an once it happens) the effect all market events area unit already incorporated available costs thence it's out of the question to predict victimization past events or costs. The stock value of a corporation depends on several intrinsic also as outside attributes. Macro-economic conditions to play a very important role in growth or decline of a sector as a full. a number of the intrinsic factors may be company's net, liabilities, demand stability, competion in market, technically advanced line, surplus money for adverse things, stakes in staple provider and finished product distributors etc. Those factors that area unit on the far side the management of the corporate like fossil oil value, dollar rate of exchange, political stability, government policy call etc. return underneath outside attribute. several researchers have tried victimization the historical stock costs because the basis for statistic analysis to forecast future stock costs. many various applied math models were applied since long like moving average (MA), autoregression (AR), weighted moving average, ARIMA, CARIMA etc. Later some non-linear models were additionally tried like GARCH. Recently completely different neural network models, biological process algorithms wre being applied for stock prediction successfully. Deep neural networks like CNN, RNN are used with completely different parameter settings and options. during this paper we have a tendency to shall explore a special sort of RNN referred to as LSTM to predict future company growth supported past stock costs.

.

2. Related Studies

There are variant analysis add stock exchange prediction also as in LSTM. nearly each data processing and prediction techniques were applied for prediction of stock costs. many alternative options and attributes were used for identical purpose. There ar 3 main classes of stock exchange analysis and prediction like (a) Fundamental analysis, (b) Technical analysis and (c) statistic analysis. Most of the stock prognostication techniques with statistic information unremarkably use either a linear like AR, MA, ARIMA, ARMA, CARIMA, etc. [1],[2] or non-linear models (ARCH, GARCH, ANN, RNN, LSTM, etc.). Authors in [3] have analyzed many alternative macro-economic factors by planning an information warehouse that affects share worth movement like rock oil worth, charge per unit, gold price, bank rate, political stability, etc. Researchers in [4] utilized frequent itemset mining technique to search out a lagged correlation between worth movement between totally different plane figure index in Indian share market. Roondiwala et al. in [5] has used RNN-LSTM model on NIFTY-50 stocks with four options (high/close/open/low worth of every day). they need used twenty-one days window to predict ensuing day worth movement. a complete of five years information has been used for prediction and RMSEas error metric to attenuate with backpropagation. Kim et al. in [6] projected a model, 'the feature fusion long short memory-convolutional neural network (LSTM-CNN) model'. they need used CNN to find out the options from stock chart pictures. They found that the candle holder charts ar the simplest candidate for predicting future stock worth movement. Next, they utilized LSTM and fed with historical worth information. they need tested on minute-wise stock worth and used thirty-minute window to forecast thirty fifth minute worth. they need tested on S&P five hundred ETF information with stock worth and trade volume exploitation CNN. They use the CNN and LSTM on an individual basis on totally different illustration of identical information and so used the combined feature fusion model for identical purpose. it's determined that the combined model outperforms individual models with less prediction error. so this work establishes the actual fact that {different totally totally different completely different} illustration of identical information (raw stock worth and trade volume and stock chart image) with combined models wherever every individual model is optimized for separate formatting will learn a lot of intrinsic information dynamics and options that is analogous to reckoning on identical object from different perspective angles that provides new insight. Hiransha et al.in their paper [7], utilized 3 totally different deep learning network architectures like RNN, CNN and LSTM to forecast stock worth exploitation day wise past closing costs. they need thought of 2 company from IT sector (TCS and Infosys) and one from the company sector (Cipla) for experiment. the individuality of the study is that they need trained the models exploitation information from one company and used those models to predict future costs of 5 totally different stocks from NSE and securities market (Network Stock Exchange). They argued that linear models attempt to match the information to the model however in deep networks underlying dynamics of the stock costs ar unearthed. As per their results, CNN outperformed all different models also as classical linear models. The DNN might forecast securities market listed firms despite the fact that the model has learned from NSE dataset. the rationale may well be the similar inner dynamics of each the stock exchanges. Gers & Schmid Huber projected a variation of LSTM by introducing "peephole connections" [18]. during this model the gate layers will look into the cell state. In another case the model coupled forget and input gates. during this case, call to feature new data or to forget it's taken along. It forgets only it must input one thing in its place. This design inputs new values to the cell state once it forgets something older. Cho, etal. [19] projected another widespread LSTM variation called the Gated perennial Unit (GRU). It aggregates each the forget Associate in Nursing input gates into an "update gate." The cell state and hidden state are integrated together with a couple of different minor modifications to form the ultimate model a lot of easy than the initial LSTM. because of the on top of reason this model is turning into widespread day by day. These are by no suggests that Associate in Nursing complete list of modified-LSTMs. There ar several different variants like Depth Gated LSTMs by Yao, et al. [20]. Koutnik, et al. [21] projected 'Clockwork RNNs' to tackle semipermanent dependencies in a very fully totally different manner.

3.LSTM Architecture

3.1 An overview of Recurrent Neural Network (RNN)

In a classical neural network, final outputs rarely act as associate degree output for successive step however if we have a tendency to listen to a real-world development, we have a tendency to observe that in several things our final output depends not solely the external inputs however conjointly on earlier output. as an example, once humans scan a book, understanding of every sentence depends not solely on the present list of words however conjointly on the understanding of the previous sentence or on the context that's created victimization past sentences. Humans don't begin their thinking from scratch each second. As you scan this essay, you perceive every word-based on your understanding of previous words. this idea of 'context' or 'persistence' isn't offered with classical neural networks. Inability to use context-based reasoning becomes a serious limitation of ancient neural network. repeated neural networks (RNN) square measure conceptualized to alleviate this limitation.

RNN square measure networked with feedback loops among to permit persistence of data. The Figure 1Error! Reference supply not found. shows an easy RNN with a feedback circuit and its unrolled equivalent version aspect by aspect.

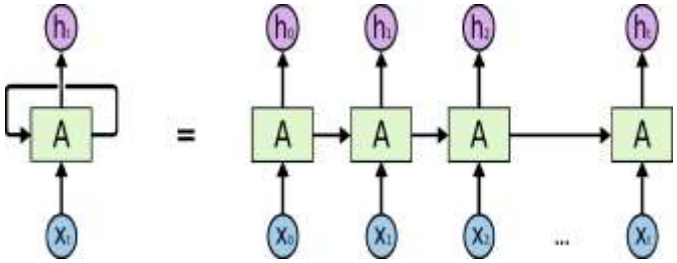


Figure 1: An unrolled recurrent neural network

Initially (at time step t) for a few input noise the RNN generates Associate in Nursing output of ht. within the next time step (t+1) the RNN takes 2 input Xt+1 and ht to come up with the output ht+1. A loop permits data to be passed from one step of the network to successive. RNNs don't seem to be free from limitations although. once the 'context' is from close to past it works nice towards the right output. however once Associate in Nursing RNN must depend upon a foreign 'context' (i.e. one thing learned long past) to supply correct output, it fails miserably. This limitation of the RNNs was mentioned in nice detail by Hochreiter [8] and Bengio, et al. [9]. They conjointly derived back to the elemental aspects to grasp why RNNs might not add semipermanent eventualities. the nice newsies that the LSTMs area unit designed to beat the on top of downside.

3.2 LSTM Network

Hochreiter & Schmidhuber [10] introduced a special variety of RNN that's capable of learning future dependencies. later on many completely different researchers improved upon this pioneering add [11] [12] [13] [14]. LSTMs square measure shaped over the time to mitigate the long dependency issue. The evolution and development of LSTM from RNNs are explained in [15] [16]. Recurrent neural networks square measure at intervals the type of a series of continuation modules of the neural network. In normal RNNs, this continuation module contains an easy structure type of one tanh layer as shown in Figure a combine of.

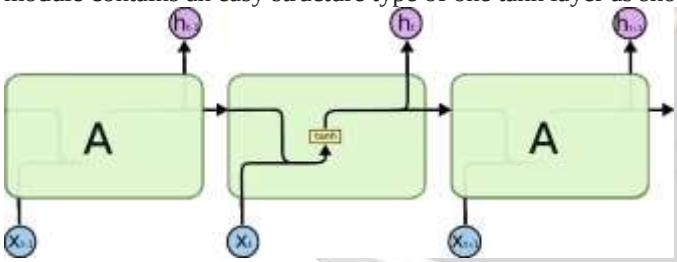
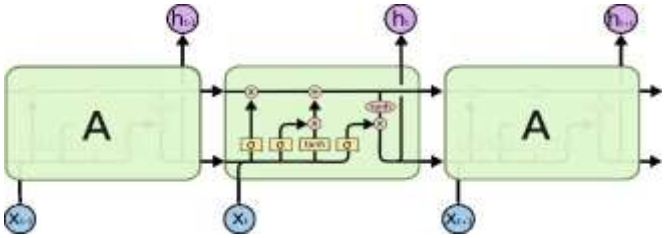


Figure 2: The repeating module in a standard RNN contains a single layer.

LSTMs follow this chain-like structure; however, the repeating module has a different structure. Instead of having a single neural network layer, there are four layers, interacting in a very special wayas shown in Figure 3.

Figure 3: The repeating module in an LSTM contains four interacting layers.



In Figure 3, every line represents an entire feature vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural

network layers. Lines merging denote concatenation, while a line forking denotes its content being copied and the copies going to different locations.

3.3 The Working of LSTM

The key to LSTMs is that the cell state, the horizontal line running through the highest of the diagram. The cell state is sort of a belt. This runs straight down the whole chain, having some minor linear interactions. LSTM has the power to feature or take away data to the cell state, controlled by structures known as gates. Gates are used for optionally let data through. Gates are composed of a sigmoid neural internet layer and a pointwise multiplication operation. The sigmoid layer outputs numbers between zero and one, describing what quantity of every component ought to be let through. a price of zero suggests that "let nothing through," whereas a price of one suggests that "let everything through!" An LSTM has 3 of those gates, to shield and management the cell state. The first step of LSTM is to determine what data are to be thrown out from the cell state. It is made by a sigmoid layer known as the "forget gate layer." it's it's and xtxt, and outputs a number between 00 and eleven for every range within the cell state Ct−1Ct−1. A eleven represents "completely keep this" while a 00 represents "completely take away this." In the next step it's set what new data ar about to be keep within the cell state. it's 2 elements. First, a sigmoid layer known as the "input gate layer" decides that values are to be updated. Thereafter, a tanh layer creates a vector of recent candidate values, C~tC~t, that would be side to the state. within the next step, these two ar combined to form associate update to the state. it's currently time to update the previous cell state, Ct−1Ct−1, into the new cell state CtCt. We multiply the previous state by ftft. Then we tend to add it∗C~tit∗C~t. this is often the new candidate values, scaledby however much we tend to arrange to update every state worth. Finally, we'd like to determine on the output. The output are a filtered version of the cell state. First, we run a sigmoid layer that decides what elements of the cell state we're about to output. Then, we place the cell state through tanh-tanh (to push the values to be between −1−1 and 11) and multiply it by the output of the sigmoid gate, in order that we tend to solely output the elements we tend to set to.

4. Proposed Framework to Forecast Share Price & Company Growth in Different Time Span

In this section, we shall first analyze some existing techniques and their merits to finally arrive at our methodology. Next, we shall discuss the algorithmic and implementation steps in detail. It is implemented in Python.

4.1 Methodology

The purpose of our framework is to investigate that is that the best time span to predict the longer-term share value of a company from a specific sector. Our objective is to predict the longer-term value and calculate the longer-term growth of the corporate within the totally different time span. Then we tend to analyze the prediction error for every company of various sector. supported that we tend to conclude which era span is best for future prediction of that specific sector. We initial predict the longer-term damage of five totally different firms from some pre-decided sectors with the assistance of LSTM. This prediction are going to be done on historical information & the longer term prediction are going to be finished 3-month, 6- month, one year & three years. In these four totally different time spans (3 & half dozen months, one & three years), we tend to calculate the growth of these firms. Then by analyzing the deviations of damage for every time span, we have a tendency to took the resultant time span that has most growth, i.e. less error for the actual sector, e.g. firms A, B, C, D & E from a sector S1 has a lot of growth in 3-months' time span of prediction then we have a tendency to draw AN conclusion that for sector S1, our framework offers the simplest prediction for next 3-months for that specific sector. In our analysis, let's contemplate we tend to square measure victimization the info for Months. Then the burden of a corporation is outlined as:

weight = 1/ (P ∗(P+1)/2)

In our case, month-wise weight (Yi) will be calculated using the following algorithm:

N: = M weight: = 1/(M*(M+1)/2)

FOR i = 1 to M Begin

Yi: = weight * N; /* Yi is the weight of previous i th month*/Q = Q - 1.

i:= i + 1End End FOR

Suppose the expansion rate between totally different time periods is Gri wherever i=1 to M, considering current year as 0 th year. Therefore, Gri is that the rate of growth of (i-1) th fundamental measure w.r.t its immediate earlier year i.e. i th year. to maximize the impact of current growth over the expansion of older year, we would develop a mathematical formula explicit below. Suppose the expansion rates of an organization area unit Gr1; Gr2…Grm respectively from gift to M years earlier. Then the corporate web rate of growth (CNGR) by the subsequent formula.

CNGRj= Y1∗ Gr1 + Y2∗ Gr2 + · · · + Yi∗Gri+ · · · + Yp∗ Grm

Where CNGRj is that the Company web rate of growth of the j th company (where j=1to m)

4.2 Implementation Steps

Step 1: Raw Stock value Dataset: Day-wise past stock costs of selected corporation's area unit collected from the BSE (Bombay Stock Exchange) official web site.

Step 2: Pre-processing: This step incorporates the following:

a) information discretization: a part of information reduction however with explicit importance, particularly for numerical information .

b) information transformation: standardization.

c) information cleaning: Fill in missing values.

d) information integration: Integration of information files.

once the dataset is reworked into a clean dataset, the dataset is split into coaching and testing sets therefore on appraise. making a knowledge structure with 60 timesteps and one output.

Step 3: Feature Selection: during this step, information attributes area unit chosen that area unit reaching to be fed to the neural network. In this study Date & shut value area unit chosen as selected options.

Step 4: Train the NN model: The NN model is trained by feeding the coaching dataset.

The model is initiated victimization random weights and biases. projected LSTM model consists of a ordered input layer followed by three LSTM layers then a dense layer with activation. The output layer once more consists of dense layer with a linear activation perform.

Step 5: Output Generation: The RNN generated output is compared with the target values and error difference is calculated.

The Backpropagation formula is employed to reduce the error distinction by adjusting the biases and weights of the neural network.

Step 6: look at Dataset Update: Step a pair of is recurrent for the look at information set.

Step 7: Error and companies' internet growth calculation: By shrewd deviation we tend to check the percentage of error of our prediction with relevance actual value.

Step 8: Visualization: victimization Keras[21] and their perform arthropod genus the prediction is unreal.

Step 9: Investigate totally different |completely different} time interval: we tend to recurrent this method to predict the value at different time intervals. For our case, we tend to took 2-month dataset as coaching to predict 3-month, 6-month, one year & three years of close value of the share. during this completely different time span, we tend to calculate the share of error within the future prediction. this could show a discrepancy for various sectors. So, this may facilitate to search out a frame for the actual sector to predict future companies' internet growth.

5. RESULT

The projected LSTM based mostly model is enforced victimization Python. In Table one the Error worth for various companies belong to Banking Sector supported the historical information of one month, 3-month, 6 month, 1 Year, 3 Year span is shown.

Table 1: Error Value for Different Banks

In the same method calculation is completed for alternative sectors additionally supported the highest-level firms belong to it sector. The error values for the world is shown in Table a pair of. It has been ascertained from the result that for pretty much all the sectors the error level comes down drastically with the check knowledge for extended periods. therefore, we advise to use this LSTM primarily based model to predict the share worth on very long-time historical knowledge.

Fig -4: State bank of India Stock Prediction

Charts

Bank Names	1 month	3 month	6 month	1 year	3 year
SBI	93.30438	9.371283	19.5584	5.148866	0.830179
HDFC	532.8527	523.4962	162.8642	24.40721	0.987856
ICICI	71.80286	9.881709	10.76914	4.575525	0.863681
Avg Error	232.6533	180.9164	64.39726	11.3772	0.893905

6. CONCLUSIONS

In this paper, we tend to analyze the expansion of the businesses from completely different sector and take a look at to search out out that is that the best time span for predicting the long run worth of the share. So, this attracts a vital conclusion that firms from an exact sector have constant dependencies additionally because the same rate of growth. The prediction is often additional correct if the model can train with a bigger variety of knowledge set. Moreover, within the case of prediction of assorted shares, there could also be some scope of specific business analysis. we will study {the completely different |the various} pattern of the share worth of various sectors and might analyze a graph with additional different time span to fine tune the accuracy. This framework generally helps in market research and prediction of growth of various firms in numerous time spans. Incorporating different parameters (e.g. capitalist sentiment, election outcome, government stability) that aren't directly related to with the price could improve the prediction accuracy.

REFERENCES

[1] F. a. o. Eugene, "Efficient capital markets: a review of theory and empirical work," Journal of finance, vol. 25, no. 2, pp. 383-417, 1970.

[2] Z. A. Farhath, B. Arputhamary and L. Arockiam, "A Survey on ARIMA Forecasting Using Time Series Model," Int. J. Comput. Sci. Mobile Comput, vol. 5, pp. 104-109, 2016.

[3] S. Wichaidit and S. Kittitornkun, "Predicting SET50 stock prices using CARIMA (cross correlation ARIMA)," in 2015 International Computer Science and Engineering Conference (ICSEC), IEEE, 2015, pp. 1-4.

[4] D. Mondal, G. Maji, T. Goto, N. C. Debnath and S. Sen, "A Data Warehouse Based Modelling Technique for Stock Market Analysis," International Journal of Engineering & Technology, vol. 3, no. 13, pp. 165-170, 2018.

[5] G. Maji, S. Sen and A. Sarkar, "Share Market Sectoral Indices Movement Forecast with Lagged Correlation and Association Rule Mining," in International Conference on Computer Information Systems and Industrial Management, Bialystok, Poland, Sprigner, 2017, pp. 327-340.

[6] M. Roondiwala, H. Patel and S. Varma, "Predicting stock prices using LSTM," International Journal of Science and Research (IJSR), vol. 6, no. 4, pp. 1754-1756, 2017.

[7] T. Kim and H. Y. Kim, "Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data," PloS one, vol. 14, no. 2, p. e0212320, April 2019.

[8] S. Selvin, R. Vinayakumar, E. A. Gopalkrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in International Conference on Advances in Computing, Communications and Informatics, 2017.

[9] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen Netzen," Diploma, Technische Universität München, vol. 91, no. 1, 1991.

[10] Y. Bengio, P. Simard, P. Frasconi and others, "Learning long-term dependencies with gradient descent is difficult," IEEE transactions on neural networks, vol. 5, no. 2, pp. 157-166, 1994.

[11] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long-time lag problems," in Advance in neural information processing systems, NIPS, 1997, pp. 473--479.

[12] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 2, pp. 107-116, 1998.

[13] J. Schmidhuber, D. Wierstra, M. Gagliolo and F. Gomez, "Training recurrent networks by evolino," Neural computation, vol. 19, no. 3, pp. 757-779, 2007.

[14] L. Pasa and A. Sperduti, "Pre-training of recurrent neural networks via linear autoencoders," in Advances in Neural Information Processing Systems, NIPS, 2014, pp. 3572-3580.

[15] J. Chen and N. S. Chaudhari, "Segmented-memory recurrent neural networks," IEEE transactions on neural networks, vol. 20, no. 8, pp. 1267-1280, 2009.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[17] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction, MIT Press, 2018.

[18] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in Proceedings of the IEEEINNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, IEEE, 2000, pp. 189- 194.

[19] K. Cho, B. Van Merri{\"e}nboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[20] K. Yao, T. Cohn, K. Vylomova, K. Duh and C. Dyer, "Depth-gated LSTM," arXiv preprint arXiv:1508.03790, 2015.

[21] J. Koutnik, K. Greff, F. Gomez and J. Schmidhuber, "A clockwork rnn," arXiv preprint arXiv:1402.3511, 2014.

[22] R. Kotikalapudi, "Keras Visualization Toolkit," [Online]. Available: https://raghakot.github.io/keras-vis. [Accessed 31 May 2019].

BIOGRAPHIES

My Name is Saurabh Kumar Mishra Currently I am Student of B. Tech (Computer science & Engineering) This Research Paper Belongs To Analysis and Prediction of Stock Market.