INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING

IJTE

ISSN: 2057-5688

A study on Deep Learning based Financial Data Prediction Algorithm

[1]GUMMA RAMAN, [2]M. ANJENEYELU

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Assistant professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

***Abstract**:* Given the ease with which gradient explosion and disappearance can occur during deep network reverse conduction when using cyclic neural network RNN prediction methods, we investigate a short-term memory (LSTM) cyclic neural network stock price prediction method employing a variable-length Batch strategy. First, using the publicly traded company's historical time series data as the study object, time series spanning many days are built for the network's input, and Early stopping technology is included during training to avoid the learning from over-fitting the data. Finally, the test set's closing price is predicted using the variable-length Batch and state parameter transfer. We validate the superior generalisation ability and reduced prediction error of the LSTM prediction model and the parameter optimisation technique based on variable length Batch by comparing their results to those of the conventional machine learning regression model.

***Keywords**: l*ong-and short-term memory model, cyclic neural network, deep learning, time series, over-fitting.

I. INTRODUCTION

The stock market is a kind of time series data that is both dynamic and noisy.

Stock prices, given their role as an economic barometer, often follow the general pattern of economic growth throughout the country. Researchers in the disciplines of banking, statistics, computer science, and so on have all focused on the challenge that is stock market forecasting[1].

There are three main methods for predicting the stock market: statistical analysis, machine learning, and deep learning. Statistical analysis is the primary tool of early prediction.

Stock prices are an example of time series data, and as such, they tend to be highly volatile and non-stationary, making them challenging to analyse using conventional statistical approaches due to the presence of complicated nonlinear interactions.

Experts and academics from a wide range of disciplines have successfully used a number of conventional machine learning algorithms to the problem of stock prediction in recent years, thanks to the explosion of research into the area of machine learning.

The models' generalisation ability is inadequate, despite their capability of modelling the nonlinear link between the stock index's past data and the future stock price. Since many variables might have an impact on a stock's price, a strong nonlinear approximation ability and the capacity to learn enormous amounts of data are two advantages of a deep neural network. Even in the field of stock forecasting, progress has been made. Circular network of neurons While RNNs have shown useful for predicting time series data, their usage is limited by the fact that networks with more than 10 layers are vulnerable to gradient disappearance or explosion during reverse conduction. Since RNN isn't well-suited for processing deep network time series data, researchers have developed the LSTM model, which builds off of RNN with an amnesia gate mechanism to address gradient disappearance and gradient explosion[2]. Predicting time series data is a strong suit for the LSTM neural network. As a result, this paper integrates state-of-the-art LSTM optimisation methods, employs Early stopping methods to avoid overfitting, builds a variable batch mechanism to actualize an LSTM three-layer network, and conducts experiments on the TSLA stock index to achieve optimal prediction error.

II. REVIEW OF LITERATURE

According to research by Yang et al. (2020), the newly identified infection, dubbed Covid 19, may be caused by SARS CoV 2. According to the preliminary research, around 18,000 individuals in China perished as a result of Covid 19. Most misfortune may be attributed to co-occurring disorders. In addition, they said that this is the third major Coronavirus epidemic in the recent two decades. After having an effect in 2002 and 2003, SARS has resurfaced, this time in its most severe form as Covid 19. Cold to severe sickness and death were among the symptoms.

It is widely thought that this strain was first exposed to people in the wild life market in Wuhan, Hubei Province. Transmission is mostly blamed on bats, pangolins, and reptiles.

In order to break the transmission cycle, the group also proposed strict isolation practises, including home isolation (Lee 2019). This is reminiscent of the Spanish flu, the most savage onslaught on humanity ever recorded in history. There is a need for strong measures to restrict mobility, such as closing borders and limiting air travel. Vaccine development research is also recommended. Preparations for antiviral medications will be prepared to stop the spread. The worldwide economy and healthcare systems were also evaluated in this research. They further extrapolated that his rapid transformation would have an emotional effect on people all across the world and may lead to acts of racism against Asians.

Yao et al. (2020) suggested using hydroxychloroquine to treat Covid 19 in the same year. The same context applies to this investigation, with Covid 19 again being traced back to its birthplace in Wuhan.

In the past, chloroquine was used to treat the SARS virus family. Chloroquine was initially the drug of choice for doctors treating this virus, according to historical accounts (Lei et al., 2018, 2020). Malaria and other illnesses spread by mosquito bites were treated with this medication. This treatment for autoimmune disorders was also shown to be without side effects. According to their findings, Hydroxychloroquine has the potential to be more effective than Chloroquine. The immuno-modulatory effect of the latter medicine is cited as the justification (Li et al., 2008; Li, 2011). This novel idea is also useful in combating the cytokine impact, which is seen in patients with severe disease. According to their findings, more robust clinical data is needed to back up their research. Hydroxychloroquine was proven to be more successful than Chloroquine in treating the novel SARS CoV 2 strain, according to extensive clinical investigations.

According to Yin-Wonget al. (1998), SARS and the coronavirus family are the most common causes of pneumonia. (Yin and Wunderink, 2018; Letko et al., 2020). They noted that prior research has deemed human coronavirus to be less dangerous. However, the spread of respiratory disorders like SARS and MERS has highlighted the role of human coronavirus as a major pathogen. This review focuses on the histology, clinical findings, and epidemiological investigations of human coronavirus.

In 2020, Zhao et al. conducted profile research of ACE.

By the third week of February, the World Health Organisation had documented approximately 76,000 instances of coronavirus infection, with about 2,500 deaths as a consequence. This illness is also rapidly spreading around the world. Human lung tissue demonstrates ACE2 RNA profile (Li and Xia, 2020). This research provides important insight into the treatment approach. This study provided the biological groundwork for future research.

This paper was published in Lancet by Zhou et al. in 2020. Cohort research using a retrospective design. This is unique research since it focuses on young adults and adults. Patients from prestigious medical centres and pulmonary medicine practises are considered for inclusion in this research. Their demographic, clinical, and psychological observations are all included in the sample data. EMRs are compared between those who did and did not make it (Lim et al., 2020).

Analyses of mortality risk variables by regression analysis. The deductions made are intriguing. Once again, old age is a potential risk factor, but the innovative aspect of this research is its emphasis on early illness prediction. This research suggests that early diagnosis in patients may allow for the use of isolation measures and alternative medicine therapies.

Since at least the 1980s, influenza has been the leading cause of mortality in the United States (Soliman et al., 2019). We uncovered an intriguing study that lends credence to our findings; in it, the authors anticipate the seasonal spread of influenza in Dallas, Texas, USA using deep learning algorithms (Lin et al., 2020).

The group used a combination of statistical models and feed forward neural networks. Selected operators were employed in conjunction with the regression, average, and absolute shrinkage approaches. The prediction of this influenza virus also made use of the Bayesian model. Using a cross-validation approach, they were able to determine that 75 hidden nodes and 2 hidden compute layers were necessary for the deep learning process. The results show a mean square error of 0.005 (Min Chen, 2017). Mortality rates were shown to be significantly affected by characteristics such as age, location, prior diagnosis of diabetes, blood pressure at 24 hours, and previous diagnosis of cardiovascular disease (Masuda and Holme 2013, Lina et al., 2020). Section 2.2 presents a number of such strategies (Masuda and Holme, 2013, Lupia et al., 2020).

III. MODEL

Linear regression Linear regression is the most basic machine learning algorithm, which builds a model based on data to reflect the relationship between input characteristics (independent variables) and targets (dependent variables), and then uses unused data to predict.

$$Y = \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n \quad (1)$$

LSTM, or Long-Term, Short-Term Memory

To address the issue of gradient disappearance and gradient expansion common to RNNs during reverse network development, Hochester and other researchers have turned to memory gate technology. Training, prediction effect, and execution time for speech, signal, financial time series data, and so on are all areas where the LSTM model excels[3].

The key structure of the LSTM model consists of three gates: the forgetting gate ft, the input gate it, and the output gate to. The forgetting gate ft determines which information is filtered out by the cell, while the input gate it and the output gate to determine which values passing through the input gate are used to update the memory state. The gated realisation formula for a typical LSTM architecture is shown in Formulas 2–5. To determine how far back data should be stored, ft first takes as input the current time, at, and as input the state of the memory unit at the previous time, ht-1, and then uses the sigmoid function to generate a number between 0 and 1. Weight is denoted by Wi, Wf, and Wo, whereas the partial FRQILGHQFH&WGHQRWHVPHPRUX QLWDQGGHQRWHVDFWLYDWLRQ function is shown by bi, bf, and bo[4].

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \sigma(W_{\tilde{C}}[h_{t-1}, X_t] + b_{\tilde{C}}) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

IV EXPERIMENT AND ANALYSIS

A. Data set

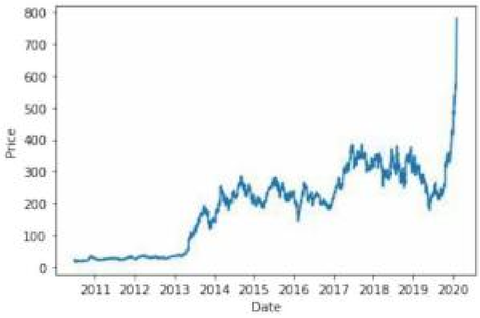


Figure 1 Historical closing price from 2010 to 2020

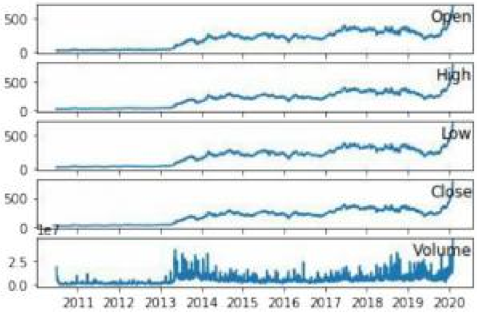


Figure 2 Data Distribution of each feature from 2010 to 2020

In the experiments, we employ 10 years' worth of trade data (from June 29, 2010 to February 3, 2020, for a total of 2416 daily line data) on Tesla stock. There are two stages to the experiment. The first section of the experiment is grounded on the classic machine learning regression model. Single-feature and multi-feature experiments are chosen using the regression coefficient (Figure 1). In the second, we use a deep learning LSTM model that doesn't take feature engineering into account (see Figure 2).

B. Measure of effectiveness

Root mean square error MSE (Mean Squared Error) is used in this study to represent the predictive power of the model. MSE, or mean squared error, is the evaluation index used to measure how well a prediction model performed in comparison to the true value.

The better the impact of prediction, the lower the value. In equation (6), M is the total number of test samples, yi is the actual value, and y quoi is the prediction.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 \qquad (6)$$

C. Analysis and experimentation using logical regression

1) Results from experiments

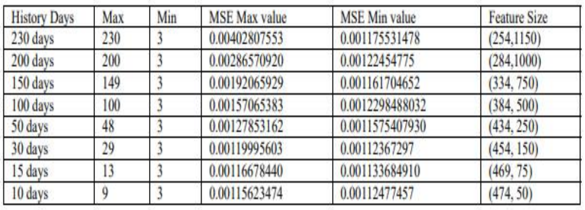
To forecast the subsequent closing price, the first set of experiments takes as inputs the previous day's starting price, highest price, lowest price, previous day's closing price, and trading volume.

We split the dataset in half, using the first 80%, or a total of 1931 daily data, as the training set and the remaining 20%, or a total of 485 data, as the test set, to determine the predictive power of the regression model.

2) Dissecting the Outcome

The MSE for logical regression predictions is shown in Table I for various timespan reviews. From Table I, we can see that the best MSE is achieved when using the first three days of data as input features for learning and prediction, regardless of the number of days in the historical review.

TABLE I THE LOGICAL REGRESSION OF FORECASTING THE FUTURE CLOSING PRICE OF STOCKS



History Days	Max	Min	MSE Max value	MSE Min value	Feature Size
230 days	230	3	0.00402807553	0.001175531478	(254,1150)
200 days	200	3	0.00286570920	0.00122454775	(284,1000)
150 days	149	3	0.00192065929	0.001161704652	(334, 750)
100 days	100	3	0.00157065383	0.0012298488032	(384, 500)
50 days	48	3	0.00127853162	0.0011575407930	(434, 250)
30 days	29	3	0.00119995603	0.00112367297	(454, 150)
15 days	13	3	0.00116678440	0.001133684910	(469, 75)
10 days	9	3	0.00115623474	0.00112477457	(474, 50)

The best number of days to look back can be selected by comparing the MSE curves of different days of review. As shown in Table I, the MSE curves of the first 10 days, 15 days, 30 days, 50 days, 100days and 150days are studied and predicted according to the review data of the first 10 days, 15 days, 30 days, 50 days, 100days, 150days and so on. The MSE curves all get the local minimum at 3 days, that is, for this data set, the data of the first three days should be used for logical regression learning. The minimum MSE can be obtained by forecasting the regression model.

V. CONCLUSION

The research investigates the LSTM neural network model with a batch size that may be adjusted. Root means square error (MSE) is reduced in this model compared to that of the standard logical regression model. Following this study, researchers may investigate automated parameter optimisation or gather market sentiment elements like news as input characteristics to enhance the accuracy of their predictions.

REFERENCES

[1] Yu Haishu, Cai Jihua, Xia Hong. Application of Arima Model in Stock Price Forecast [J]. Economist, 2015 (11): 156-157.

[2] Ariyo A A,Adewumi AO,Ayo CK.Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. 26-28 March2014, Cambridge, UK.IEEE, 2014:106-112.

[3] Shi Jia, Liu Wei, Feng Zhichao, et al. Analysis and Forecast of Stock Market Price Law based on ARIMA Model [J]. Statistics and applications, 2020, 9 (1): 101-114.

[4] Patel J,Shah S,Thakkar P, et al. Predicting stock market index using fusion of machine learning techniques[J].Expert Systems with Applications,2015,42(4):2162-2172.

[5] Abdelhamid B., (2009), 'Radial Basis Function Nets for Time Series Prediction', International Journal of Computational Intelligence Systems, Vol.2, 2, pp. 147-157.

[6] Abdou, H.A. and Pointon, J. (2011), 'Credit scoring, statistical techniques and evaluation criteria: a review of the literature', Intelligent Systems in Accounting, Finance, and Management, Vol. 18, 2–3, pp.59–88. [7] Ak, O. Fink, E. Zio, (2016), 'Two Machine Learning Approaches for Short-Term Wind Speed Time-Series Prediction', IEEE Transactions on Neural Networks and Learning Systems, Vol.27,8, pp.1734-1747.

[8] Prasadu Peddi (2023), Using a Wide Range of Residuals Densely, a Deep Learning Approach to the Detection of Abnormal Driving Behaviour in Videos, ADVANCED INFORMATION TECHNOLOGY JOURNAL, ISSN 1879-8136, volume XV, issue II, pp 11-18.

[10] B.Majhi, M. Rout, R. Majhi, G. Panda, P. J. Fleming, (2012), 'New robust forecasting models for exchange rates prediction', Expert Systems with Applications, Vol.39, pp.12658-12670.

[11] Bask, A., Merisalo-Ratanen, H., Tinnila, M. and Lauraeus, T. (2011), 'Towards banking: the evolution of business models in financial services', International Journal of Electronic Finance, Vol. 5, 4, pp.333–356.

[12] Naga Lakshmi Somu, Prasadu Peddi (2021), An Analysis Of Edge-Cloud Computing Networks For Computation Offloading, Webology (ISSN: 1735-188X), Volume 18, Number 6, pp 7983-7994.

[13] Prasadu Peddi (2021), "Deeper Image Segmentation using Lloyd's Algorithm", ZKGINTERNATIONAL, vol 5, issue 2, pp: 1-7.