Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230512

Stock price prediction using decision tree classifier and LSTM network

Hongyi Xu

School of Computer and Information Technology, Beijing Jiaotong University (BJTU), Beijing, 100044, China

21281057@bjtu.edu.cn

Abstract. Nowadays, stock price prediction has become a popular research topic, many researchers try to predict stock prices in various ways. However, there are many different tools, but not all of them have good performance, so it is necessary for researchers to evaluate and compare different tools. In this paper, to achieve the goal of predicting stock price precisely, the main approach chosen is building deep learning models and use them to make predictions. Two methods, decision tree and long short-term memory (LSTM) neural network, are used in this study. In the model using the decision tree classifier, the daily state of the stock is divided into two types: the rise and fall of the stock price. The task of the model is to make predictions about daily stock prices and classify them. The other model uses the LSTM network, which is used to make accurate closing price predictions. In the end, the performance of the two models is assessed for further work.

Keywords: Stock Price, Machine Learning, Decision Tree, LSTM Network.

1. Introduction

In this era of rapid economic development, more and more people are trying to speculate in the stock market. In order to make profits, stockholders select the stocks that have an upward price trend in their mind and buy them. In addition, the stocks whose prices are expected to go down will be sold [1]. It is precisely because of these behaviors that the stock market can operate normally. However, in many cases, stockholders' forecasts are not accurate, because there are many factors that affect stock prices [2], such as national policies, investor irrationality (herd mentality, etc.), macroeconomic conditions, company management capabilities, and so on. Therefore, the idea that people want to predict stocks came into being. Deep learning is a very good tool for various predictions. It can learn a large amount of input data through a neural network, and finally dig out the potential patterns of the data. As a result of that, deep learning has become a very popular stock forecasting tool. In this paper, this research's goal is to find a deep learning method that can predict stock prices reliably. The dataset is split into test and training sets [3] and used for training in two models. One model uses a decision tree classifier and the other one uses LSTM networks.

2. Related Work

To achieve higher accuracy, many hybrid and innovative ideas have been proposed in succession. Wasiat Khan expanded the source of information, combining deep learning models with information from social

DOI: 10.54254/2755-2721/37/20230512

media [1]. Erhan Beyaz evaluated the accuracy of the deep learning model using Fundamental and Technical analysis and tried to mix these two methods together [4]. Besides, Pratik Patil presented two kinds of hybrid models based on a novel approach using graph theory, which used convolutional neural networks and a traditional machine learning approach respectively [5]. In 2022, Yu-Xuan Luo and Yi Ji presented a hybrid model that combined IPSO (improved particle swarm optimization) and LSTM neural network for prediction [6]. In 2018, Althelaya and his group did a comparison between different LSTM architectures used in short-term and long-term stock prediction. In addition, they also evaluated their performance and compared it with simpler neural networks and LSTM whose form is more basic [7]. Other methods like Decision tree linear regression are also used to build models that predict stock. In 2021, Rezaul Karim and his teammates designed two models using Decision Tree and Linear Regression respectively to improve the efficiency [8]. In order to evaluate the results, some researchers also summarized and compared various models. Ernest Kwame Ampomah and his group do a comparison between the performance of different tree-based ensemble ML models [3]. Furthermore, Sidra Mehtab built several regression models and compared their performance in detail [9].

3. Methods

3.1. Decision Tree and Gini Index

The Decision Tree is a kind of tree structure, which can determine the classification of the input data. Each non-leaf node on the decision tree represents a test. Besides, every leaf node holds a class. The classification begins from the root, and it is over when it reaches the bottom, and the leaf node's class is seen as the result.

In this kind of classification, the key problem is how to find out the best classification rules. In general, as the process goes on, the purity of each class's content is expected to increase. Decision tree classifiers have three optional criteria to measure the quality of a split, which are "gini", "entropy" and "log loss". The default method of Gini index calculation is used in this paper. The lower the Gini index, the higher the purity of the divided datasets. At each step, the classifier calculates the Gini index of each feature separately and selects the one with the least Gini index. Its formula for calculating purity is as follows:

$$\mathrm{Gini(p)} = \sum_{k=1}^{K} P_k(1 - P_k) = 1 - \sum_{k=1}^{K} {P_k}^2 \quad (1)$$

Where P_k is the possibility of correct classification, which means that the sample is contained in class K.

3.2. LSTM Network

LSTM network makes improvements in the original RNN network. When dealing with the degradation of the deep learning model, it can perform better. The basic function of the LSTM network is to remember long-period useful information, and the selection of information is accomplished by three gates. It uses a chain structure. The workflow is displayed in Figure 1:

DOI: 10.54254/2755-2721/37/20230512

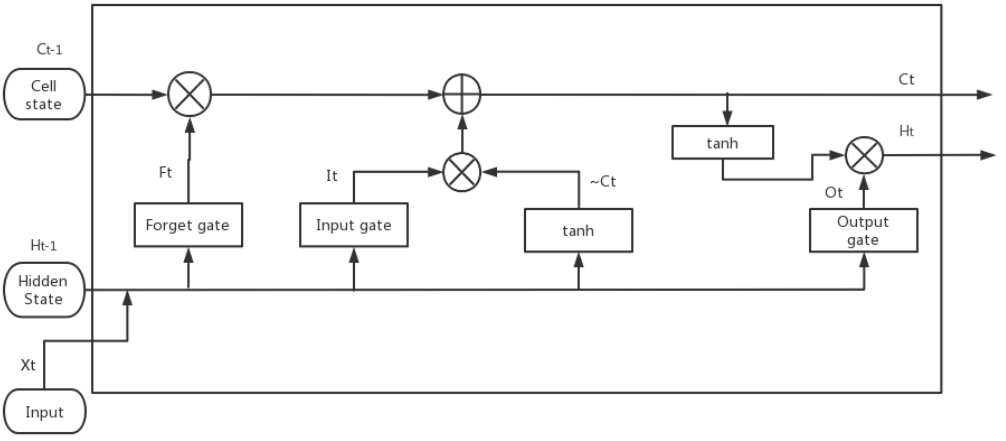

Figure 1. The schematic of LSTM Network's Working Process.

The following describes the workflow of an LSTM network in more detail:

Firstly, the forget gate determines what information needs to be discarded. Its output value F_t is greater than 0 and less than 1, which shows to what extent the previous cell state C_{t-1} needs to be retained. The formula for calculating the output is as follows:

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \tag{2}$$

Where X_t is input to the LSTM unit, H_{t-1} is the hidden state from the last LSTM unit, W represents the weight matric and b represents the bias parameter, σ is the sigmoid function.

The second gate in the LSTM network, the input gate, determines what to keep from the candidate memory in the next step. The formula for calculating the I_t is as follows:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \tag{3}$$

After that, a generator using the tanh function generates a candidate memory. The output of the generator is multiplied by I_t. Next, the cell state C_t is updated using the result. The formula for calculating the $\tilde{C}_t$ and C_t is as follows:

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{4}$$

$$C_t = F_t * C_{t-1} + I_t * \tilde{C}_t \tag{5}$$

Where $\tilde{C}_t$ is candidate memory.

Finally, the hidden state is updated. The last gate will complete a task which is to select and calculate the output information, see formula(6). The C_{t-1} is passed into the hyperbolic tangent (tanh) function, then multiplied by O_t to generate a new result. This result H_t is the new hidden state, see formula (7):

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{6}$$

$$H_t = O_t * \tanh(C_{t-1}) \tag{7}$$

4. Results

4.1. Dataset

The dataset used in this paper is the Infosys 5-year dataset from 01-07-2015 to 29-07-2020. It is taken from NSE INDIA National Stock Exchange of India Ltd. Figure 2 shows INFOSYS's stock price in a short period.

DOI: 10.54254/2755-2721/37/20230512

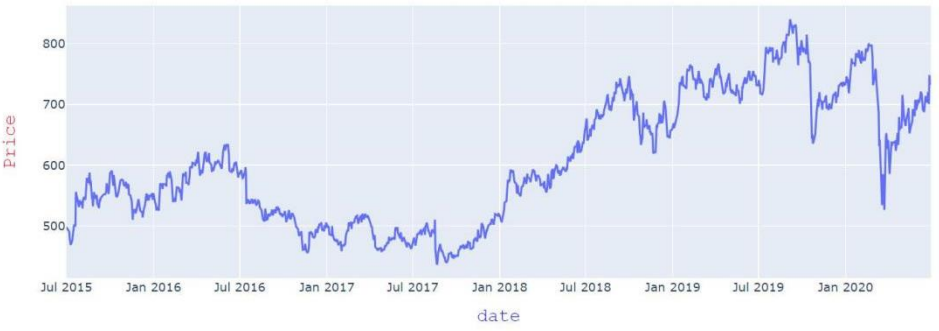

Figure 2. Stock Price of INFOSYS.

The dataset contains several key columns, which represent every day's stock price and its trend in detail [1]. The information contained in the dataset can be seen in Table 1.

Table 1. Information of INFOSYS Stock Price Dataset.

Date	Open	High	Low	Close	Adj Close	Volume
2015/7/1	494.5	502.5	493	498.700012	415.561249	6880852
2015/7/2	499.5	500.700012	492.524994	494	411.644745	4007568
2015/7/3	494	496.5	491	495.149994	412.603058	2695306
2015/7/6	492.5	494	487.5	491.649994	409.686493	4305602
2015/7/7	492.5	495	489.5	490.25	408.519897	3497418

4.2. The Experimental Result of Decision Tree

Table 2 shows the input and output of the model. If the price when the market is closed is significantly higher than the price when the market is opened, the stock price is very likely to rise in the short term. In the opposite situation, it tends to decline in the short term. Hence their difference is chosen as input. To make the result more accurate, the difference between the daily maximum and minimum prices is also introduced as input.

A decision tree classifier can classify the input into different labels, so the selected output is the label. The purpose of this model is to predict the rise and fall of stocks, so two categories are chosen for the decision tree to classify, namely "price increase" and "price decrease".

Adjclose (Adjusted Close) refers to the price after the closing price of the stock has been reset. It is a very important price in the historical data of the stock [6]. When a stock is listed and traded, its price is affected by various factors, such as dividends, stock splits, joint stock, etc., and these factors will have an impact on the stock price. Adjclose is the closing price adjusted by reweighting the historical data of the stock to make the stock price comparable. Therefore, for evaluation, the difference between the adjusted clothing price of two adjacent days is chosen as the classification standard. For each day's data, the adjusted close of the previous day is subtracted from the adjusted close of the day, and when the value is greater than 0, it is classified as a "stock price increase".

Table 2. The Input and Output of Decision Tree Model.

Input	Output
Open-Close High-Low	State (-1 for price increase, 1 for price decrease)

Figure 3 shows the results of model training. In Table 3, accuracy indicates the share of the samples that are predicted accurately, micro avg combines the results of all the categories of predictions and then calculates the average for the whole, macro avg averages the performance metrics for each category

DOI: 10.54254/2755-2721/37/20230512

separately and then averages them, weighted avg indicates the weighted average value. -1 means that the stock price will fall, and 1 means that the stock price will rise. In this classification tree model, the accuracy is 0.74.

	precision	recall	f1-score	support
-1	0.78	0.72	0.75	131
1	0.71	0.77	0.74	115
accuracy			0.74	246
macro avg	0.74	0.75	0.74	246
weighted avg	0.75	0.74	0.74	246

Figure 3. The Results of Model Running.

Figure 4 shows the results of the decision tree model in more detail. An ordinate of 1 means that the stock price goes up, an ordinate of -1 means that the stock price goes down, and the abscissa is the date. When the blue and orange dots overlap, the stock's state (whether the price went up or down that day) was predicted correctly for that day. Due to the large amount of data, this graph only shows the forecast for about two months.

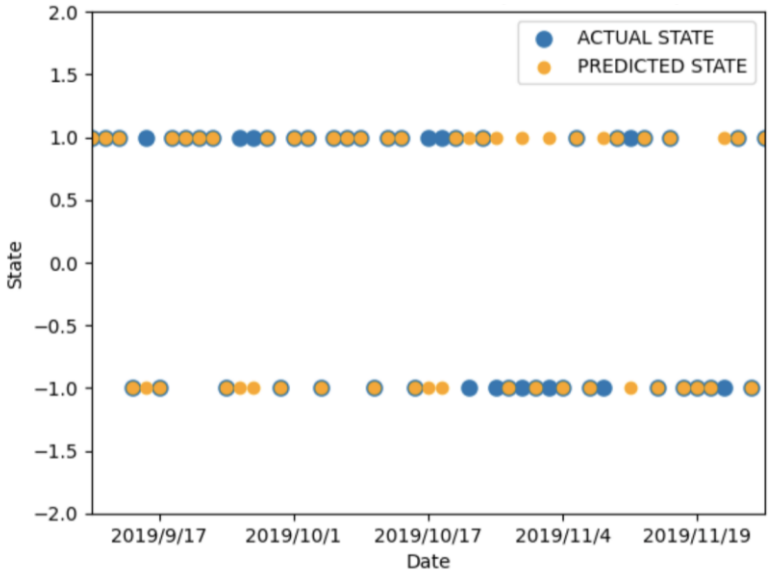

Figure 4. INFOSYS STOCK PREDICTION (DECISION TREE).

4.3. The experimental result of LSTM

Table 3 shows the information on the input and output of the model. The closing price was selected as the benchmark for stock forecasting. First, the closing price column is extracted, and then the data is subjected to feature scaling to achieve data normalization, aiming at involving the transformation of features in a common range [10]. This research's goal is to predict the rise and fall of stocks on a certain day after this period based on historical data, that is, data within a period, so the input is a time series. The output is the closing price for a certain day, and the input is chosen to be the closing price for the previous 60 days. Hyper-Parameters are summarized in Table 4.

DOI: 10.54254/2755-2721/37/20230512

Table 3. The Input and Output of the LSTM Model.

Input	Output
Closing prices for the previous 60 days	Closing prices

Table 4. Hyper-Parameters in the LSTM Model.

Epoch	Batch Size	Verbose	Optimizer	Loss	Validation Split
15	16	2	Adam	Mean Squared Error	0.2(test)+0.8(train)

Table 5 gives the structure of the model, which contains four LSTM layers and a dense layer. Each LSTM layer is subjected to dropout processing, and the dropout rate is set to 0.2 to prevent overfitting [11].

Table 5. Structure of the LSTM Model.

Layer (Type)	**Output Shape**
LSTM	(None,60,100)
Dropout	(None,60,100)
LSTM	(None,60,100)
Dropout	(None,60,100)
LSTM	(None,60,100)
Dropout	(None,60,100)
LSTM	(None,100)
Dropout	(None,100)
Dense	(None,1)

The Change of training Model loss with the increase of epoch is shown in Figure 5. The epoch of this training is 20. As the epoch increases, the loss gradually decreases, and the final loss drops below 0.003.

DOI: 10.54254/2755-2721/37/20230512

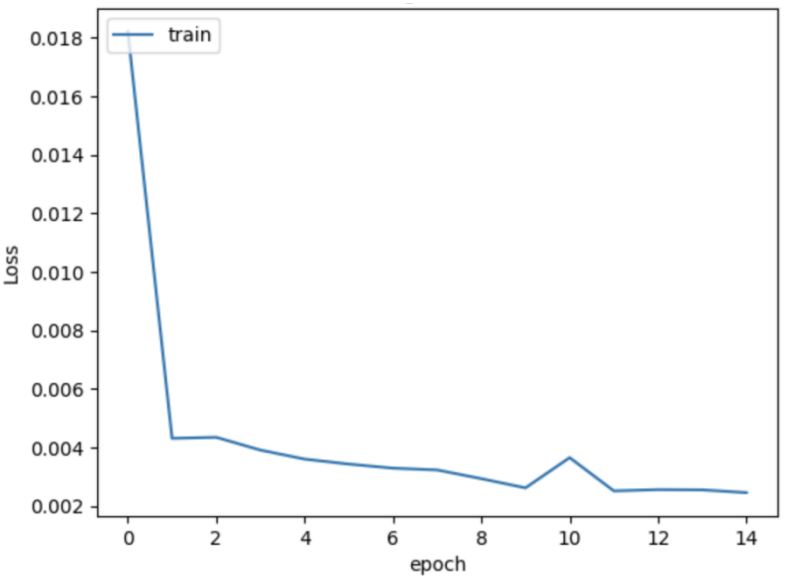

Figure 5. Training Model Loss.

Figure 6 shows the predicted stock movement versus the actual stock movement. The horizontal axis represents time, the vertical axis represents the stock price, the green curve represents the prices predicted by the model. As a contract, the orange curve shows actual prices. The overlapping rate shows the performance of this model.

Figure 6. INFOSYS Stock Price Prediction.

DOI: 10.54254/2755-2721/37/20230512

5. Conclusions

This research evaluates the efficiency of two deep learning models for stock predictions. The decision tree classifier is used to judge every day's stock price trend. Two inputs are selected, which are close price minus open price and the difference between the two extreme prices every day. The final accuracy can reach a percentage of 74%. Another model uses LSTM networks. Compared to the first model, it can predict precise closing prices instead of price trends. It still lacks the details of price changes to accurately predict daily stock prices, but the prediction fits the actual stock price in general.

References

[1] Khan W, Ghazanfar M A, Azam M A, Karami A, Alyoubi K H and Alfakeeh A S 2022 Stock market prediction using machine learning classifiers and social media, news J. Ambient Intell Human Comput 3433–3456(2022)013

[2] Lu W, Li J, Wang J 2021 A CNN-BiLSTM-AM method for stock price prediction Neural Comput & Applic 4741–4753(2021)033

[3] Ampomah E K, Qin Z and Nyame G 2020 Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement Information 11(6)(2020)332

[4] Beyaz E, Tekiner F, Zeng X-j and Keane J 2018 IEEE 20th Int. Conf. on High Performance Computing and Communications; IEEE 16th Int. Conf. on Smart City; IEEE 4th Int. Conf. on Data Science and Systems (HPCC/SmartCity/DSS) (Exeter, UK) p 1607-1613

[5] Patil P, Wu C-S M, Potika K and Orang M 2020 Proc. of 3rd Int. Conf. on Software Engineering and Information Management (ICSIM '20) (Sydney, NSW, Australia) p 85–92

[6] Luo Y-X and Ji Y 2022 Int. Conf. on Machine Learning and Cybernetics (ICMLC) (Japan) p 43-48

[7] Althelaya K A, El-Alfy E S M and Mohammed S 2018 9th Int. Conf. on information and communication systems (ICICS) (Irbid, Jordan) p 151-156

[8] Karim R, Alam M K and Hossain M R 2021 1st Int. Conf. on Emerging Smart Technologies and Applications (eSmarTA) (Sana'a, Yemen) p 1-6

[9] Mehtab S, Sen J and Dutta A 2020 Machine Learning and Metaheuristics Algorithms, and Applications: 2nd Symp.(SoMMA 2020) (Chennai, India) p 88-106.

[10] Singh D and Singh B 2020 Investigating the impact of data normalization on classification performance Applied Soft Computing 97(2020)105524

[11] Garbin C, Zhu X and Marques O 2020 Dropout vs. batch normalization: an empirical study of their impact to deep learning Multimedia Tools and Applications 12777-12815(2020)079