# Study on the application of LSTM-LightGBM Model in stock rise and fall prediction

*Yuankai* Guo[1,*], *Yangyang* Li[2], and *Yuan* Xu[1]

[1]Ankang Vocational Technical College, College of Engineering, Ankang 725000, China
[2]Xunyang Second Middle School, History Teaching Group, Ankang 725741, China

**Abstract.** This paper proposes a hybrid financial time series forecast model based on LSTM and LightGBM, namely LSTM_LightGBM model. Use the LightGBM model to train the processed stock historical data set, and save the training results. Then the opening price, closing price, highest price, lowest price, trading volume and adjusted closing price are separately input into the LSTM model for prediction. The prediction result of each attribute is used as the test set of the prediction after LightGBM training. Constantly adjust the parameters of each model, and finally get the optimal stock price forecast model. The model is validated with the rise and fall of AAPL stock. Through the comparison of evaluation index root mean square error RMSE, mean absolute error MAE, prediction accuracy Accuracy and f1_score. It is found that the LSTM_LightGBM model exhibits stable and better prediction performance in the stock prediction. That is to say, the LSTM_LightGBM model proposed in this paper is stable and feasible in the stock price fluctuation forecast.

## 1 Introduction

In recent years, many new forecasting techniques can be used to build efficient and accurate deep learning forecasting models, but stock forecasting is still a hot topic [1]. In order to overcome some shortcomings in the traditional time series model, Long Short-Term Memory (LSTM) was proposed [2-3]. At the same time, some ensemble learning models have also been applied to the application and research of time series data forecasting [4-5]. In the research of many scholars, it is not difficult to find that the use of a single LSTM model to predict time series data shows better model prediction performance [7-8]. Although a single LSTM model has shown strong momentum in the prediction of nonlinear time series data, stock time series data has highly nonlinear characteristics, and there are still some shortcomings in predicting only a single LSTM model. Therefore, the LSTM combined model is widely used.

In this paper, the combined prediction method of LSTM and LightGBM is used [8]. Construct LSTM_LightGBM combined model. First, the collected time-series stock historical data is processed for missing values, and the attributes in the stock data are used as input features, input into the LightGBM model for training, and the training results are

* Corresponding author: 646288897@qq.com

saved. Then, Open, High, Low, Close, Volume, Adj Close were separately input into the LSTM model for prediction. The prediction results of each attribute are used as the test set for the prediction after the LightGBM training, and the parameters of each model are continuously adjusted to finally obtain the optimal stock fluctuation prediction model. Use LSTM_LightGBM prediction model to predict stock prices. Comparing LSTM_LightGBM with LSTM_XGBoost, LSTM_AdaBoost, a single LSTM network model and RNN network model, it is found that the LSTM_LightGBM model proposed in this paper is stable and feasible in the stock fluctuation prediction of time series data.

## 2 Related work

The LSTM_LightGBM model is constructed by using the combined prediction method of LSTM and LightGBM. First, the missing values in the collected stock historical data are removed, and Open, High, Low, Close, Volume, and Adj Close are respectively input into the LSTM prediction model for prediction. The prediction results and time components of each attribute are reconstructed, and the reconstructed data set is used as the prediction test set. There are input gates, forget gates and output gates in the LSTM unit. When the stock history information passes through the input gate of LSTM, the stock history information that meets the requirements of the model will be retained, and the stock history information that does not meet the requirements will be forgotten through the forget gate. LSTM uses Sigmoid neural network layer and pair multiplication to determine the correctness of stock history information. At the same time, LSTM also includes a tanh activation layer. In order to prevent overfitting in the training process of the LSTM model, a Dropout layer is added, and the forgetting rate is set to 0.2.

When the LSTM unit inputs the information $x_t$ and the output $h^{t-1}$ value at the previous moment, the forgetting gate will use the Sigmoid function to calculate the value of $f_t$, as shown in formula 1.

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{1}$$

After determining the forgetting rate $f_t$, the input gate determines the information that needs to be updated through the Sigmoid layer, and the *tanh* function generates a new vector $\hat{C}_t$, which multiplies the forgetting rate it and the candidate value $\hat{C}_t$ to obtain the updated data $C_t$. As shown in formulas 2, 3 and 4.

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{4}$$

The unit state is processed by the *tanh* function to obtain a value between [-1, 1], and it is multiplied with the output of the Sigmoid layer, and the result is the final output result.

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{5}$$

After the LSTM predicts the data, the Open, High, Low, Close, Volume, Adj Close, converted time information and stock fluctuations in the stock data are used as input

features and input to the LightGBM model for training. And save the training results. Use the newly constructed test set as the test set for prediction after LightGBM training, adjust the parameters of each model, and finally obtain the optimal prediction model.

LigthGBM uses the negative gradient of the loss function as the residual approximation of the current decision tree to adapt to the new decision tree. The residual error is approximated by the Taylor expansion of the loss function, and the complexity of the model is controlled by the regular term. LightGBM uses a leaf-by-leaf split strategy, and only selects the node with the largest split gain for splitting, thus avoiding the overhead caused by the small node gain. In addition, LightGBM uses a histogram-based decision tree algorithm to save only the discrete values of features, thereby reducing memory usage and improving model training speed. After the LSTM_LightGBM model prediction ends, the evaluation indicators are used for model evaluation. The prediction performance evaluation indicators of stock prediction models use root mean square error (RMSE), mean absolute error (MAE), accuracy (Accuracy) and F1 score (f1-score). Four evaluation indicators evaluate the performance of the model. As shown in formulas 6-9.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - \hat{y}^{(i)}\right| \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$f1\_score = \frac{2 * pre * recall}{pre + recall} \tag{9}$$

The specific construction flow chart of the LSTM_LightGBM model is shown in Figure 1.

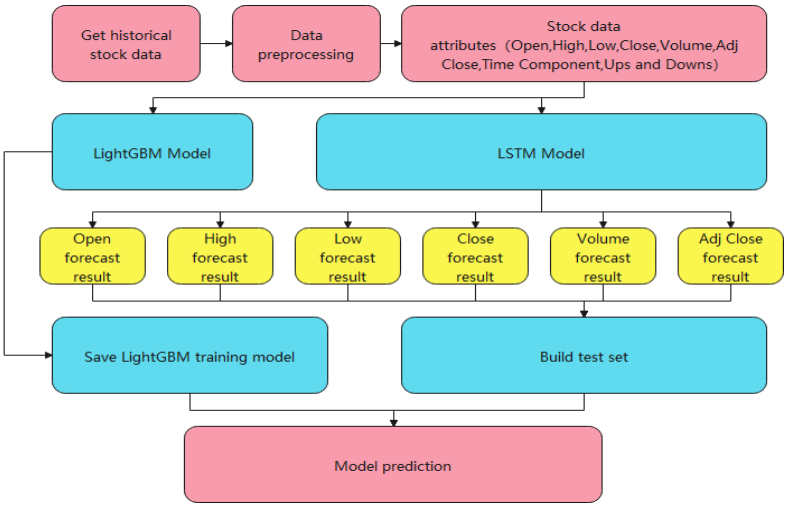


**Fig. 1.** LSTM_LightGBM model construction flowchart.

## 3 Simulation experiment

The experiment in this article uses python language for programming. Use sklearn, pandas, numpy and other python packages. The experimental data set is obtained through the website https://finance.yahoo.com, and the historical stock data of AAPL from January 2002 to August 2020 is selected, with a total of 4678 data. Among them, 80% of the AAPL stock data is used as the training set, and 20% of the data is used as the test set.

During the experiment, in order to better show the performance of the LSTM_LightGBM model, LSTM uses single-step prediction, that is, the prediction sliding window length W=1, epochs is set to 10, and batch_size is set to 64. The six attributes of Open, High, Low, Close, Volume and Adj Close in AAPL stocks are experimental data. After determining the LSTM model parameters, each attribute is trained separately. The comparison result of each attribute prediction is shown in Fig. 3. According to Fig. 2, it can be clearly seen that the prediction fit of each attribute of the LSTM model is better.

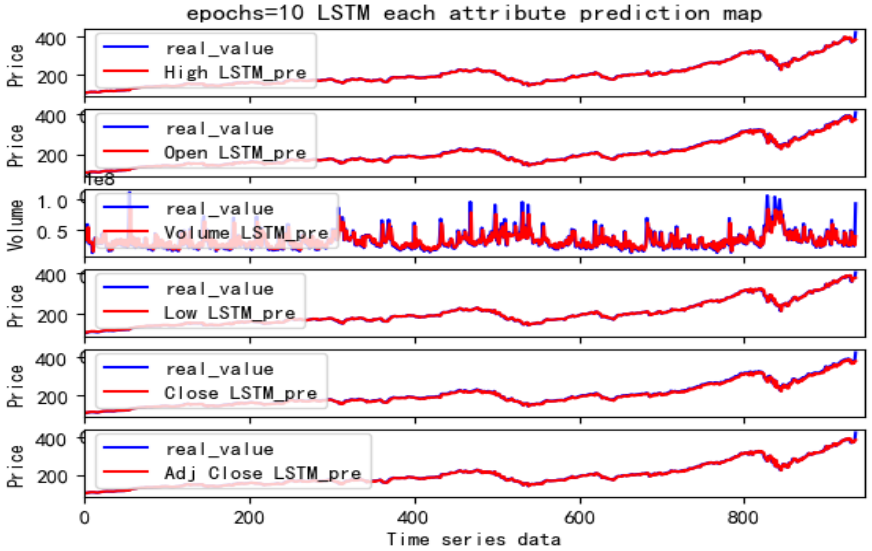


**Fig.2.** LSTM model attribute prediction diagram.

After the parameters of the LSTM model are determined, the LightGBM model is trained. The LightGBM parameters use the default parameters, and the training set uses the top 80% of the AAPL stock data. Save the model trained by LightGBM, and integrate the prediction results of each attribute of the LSTM model in Experiment with the time attribute to form a test training set. The model trained by LightGBM is used to predict the new test set, and the final up and down prediction results of AAPL stocks are obtained. The LSTM_LightGBM model is compared with the five prediction models of LSTM_XGBoost, LSTM_AdaBoost, LSTM and RNN. When the Epochs are 1, 60, 120, the experimental data are shown in Tables 1 to 3.

**Table 1.** Comparison of model prediction results when Epochs=1.

| | LSTM-LightGBM | LSTM-XGBoost | LSTM-AdaBoost | LSTM | RNN |
|---|---|---|---|---|---|
| RMSE | **9.449058** | 9.725710 | 10.450413 | 10.043927 | 14.207721 |
| MAE | **1.992912** | 2.049674 | 2.221587 | 2.090631 | 2.876477 |
| Accuracy | **0.548076** | 0.451923 | 0.451923 | 0.445512 | 0.454059 |
| f1_score | **0.708074** | 0.0 | 0.0 | 0.207633 | 0.015414 |

**Table 2.** Comparison of model prediction results when Epochs=60.

| | LSTM-LightGBM | LSTM-XGBoost | LSTM-AdaBoost | LSTM | RNN |
|---|---|---|---|---|---|
| RMSE | **9.44807** | 9.82260 | 9.53899 | 12.73173 | 13.42140 |
| MAE | **1.99048** | 2.08007 | 2.01092 | 2.65647 | 2.78044 |
| Accuracy | **0.54594** | 0.48290 | 0.52350 | 0.45192 | 0.45512 |
| f1_score | **0.70628** | 0.16262 | 0.38904 | 0.0 | 0.01544 |

**Table 3.** Comparison of model prediction results when Epochs=120.

| | LSTM-LightGBM | LSTM-XGBoost | LSTM-AdaBoost | LSTM | RNN |
|---|---|---|---|---|---|
| RMSE | **9.45091** | 9.73520 | 9.95047 | 12.79745 | 13.74464 |
| MAE | **1.99235** | 2.05050 | 2.09751 | 2.67847 | 2.83618 |
| Accuracy | **0.54059** | 0.48076 | 0.51388 | 0.45085 | 0.45299 |
| f1_score | **0.70138** | 0.15625 | 0.35092 | 0.0 | 0.00389 |

It can be seen from Tables 1 to 3 that when the Epochs are 1, 60, 120, the LSTM_LightGBM model is better than the LSTM_XGBoost model, LSTM_AdaBoost model, LSTM and RNN models in RMSE, MAE, Accuracy and f1-score. This shows that the LSTM_LightGBM model is more stable and practical.

After the experimental comparison of AAPL stock rise and fall forecasts. It is not difficult to find that the LSTM_LightGBM model is better than the LSTM_XGBoost model, the LSTM_AdaBoost model, the single model LSTM model and the RNN model that performs well in the stock price prediction in the stock price prediction. Furthermore, it verifies that the LSTM_LightGBM model proposed in this paper has certain feasibility and stability in the prediction of stock rise and fall trends.

## 4 Conclusion

This paper proposes a hybrid financial time series model based on LSTM and LightGBM, namely LSTM_LightGBM model. Use the LightGBM model to train the processed stock historical data set, and save the training results. Then the opening price, closing price, highest price, lowest price, trading volume and adjusted closing price are separately input into the LSTM model for prediction. The prediction result of each attribute is used as the test set of the prediction after LightGBM training. Constantly adjust the parameters of each model, and finally get the optimal stock price forecast model. The model is validated with the rise and fall of AAPL stock. Through the comparison of evaluation index root mean square error RMSE, mean absolute error MAE, prediction accuracy Accuracy and f1_score. It is found that the LSTM_LightGBM model exhibits stable and better prediction performance in the stock prediction. That is to say, the LSTM_LightGBM model proposed in this paper is stable and feasible in the stock price fluctuation forecast.

## References

1. Wang J , Hou R , Wang C , et al. Improved v -Support vector regression model based on variable selection and brain storm optimization for stock price forecasting[J]. Applied Soft Computing, 2016, 49:164-178.
2. De-Bao D , Yu-Sen L , Ti-Jun F , et al. Stock Forecast with Investors Sentiment by Text Mining and Machine Learning[J]. China Soft ence, 2019.
3. Gers F A , Schmidhuber, Jürgen, Cummins F . Learning to Forget: Continual Prediction with LSTM[J]. Neural Computation, 2000, 12(10):2451-2471.

4. Zhang J , Li L , Chen W . Predicting Stock Price Using Two-Stage Machine Learning Techniques[J]. Computational Economics, 2020.
5. A Y D , A Y Z , A J F , et al. Interpretable spatio-temporal attention LSTM model for flood forecasting[J]. Neurocomputing, 2020, 403:348-359.
6. Yubo N , Yujun Z . LSTM-Adaboost's stock price forecasting model[J]. Journal of University of ence and Technology Liaoning, 2019.
7. Jigen L , Jianqiang D , Bin N , et al. TCM text relationship extraction model based on bidirectional LSTM and GBDT[J]. Application Research of Computers, 2019.
8. Tingyu Weng,Wenyang Liu,Jun Xiao. Supply chain sales forecasting based on lightGBM and LSTM combination model[J]. Industrial Management & Data Systems,2019,120(2).