Future Stock Price Prediction Based on Bayesian LSTM in CRSP

ZhangQi Wang[#] ,ZiYi Qi[#]*

Beijing University of Technology,Beijing, China

[#] These authors contributed equally to this work.

*qiziyi2002@buaa.edu.cn

Abstract. This paper mainly studies the problem of forecasting future stock prices based on historical data. Taking six companies in the CRSP Daily Stock database as research objects, three models of RNN, LSTM, and Bayesian LSTM are used to predict stock prices, and Gaussian noise is added to the data set to analyze the disturbance of the three models. Finally, the results prove that Bayesian LSTM has better predictive effect and robustness for processing time series data.

Keywords: future stock prices , RNN, LSTM, Bayesian LSTM

1 Introduction

With the continuous development and popularization of artificial intelligence technology, more and more fields begin to apply artificial intelligence technology for data analysis and forecasting, among which the stock market is also an important application field. The volatility of the stock market is very large, but through the analysis of historical data of stock prices, the trend of future stock prices can be determined and the stock market can be predicted. Chen K et al. used deep learning methods to predict China's Shenzhen stock market back in 2015[6] This can not only provide investors with important information to help governments and companies make important economic decisions, but also help analysts and economists better understand market trends and predict future economic developments. In this article, we will predict the stock price based on the Bayesian LSTM model, select 6 companies from the CRSP Daily Stock database as the data set, and compare it with RNN and LSTM stock price predictions to prove the Bayesian LSTM model in stock price prediction. Advantage. We will use Python to perform data processing on three models of RNN, LSTM, and Bayesian LSTM for each company, perform disturbance analysis, and calculate statistical values such as MSE, RMSE, and MAE to evaluate forecasts. It has been proved that the Bayesian LSTM model has very impressive prediction results for stock prices and is not easily affected by disturbances. Also, the Bayesian neural network will give a confidence interval for the predicted data rather than a predicted value. Because

Y. Jiao et al. (eds.), *Proceedings of the 3rd International Conference on Internet Finance and Digital Economy (ICIFDE 2023)*, Atlantis Highlights in Economics, Business and Management 1,
https://doi.org/10.2991/978-94-6463-270-5_23

of the high-risk investment in stocks, it can give investors and economists a more accurate reference value.

2 Background and Theory

2.1 Research Background

At present, research models on stock price prediction include RNN, LSTM, etc. RNN (Recurrent Neural Network) is a neural network model capable of modeling sequence data, which has memory capability and time dependence. In the article "Stock price prediction using the RNN model[2]", the RNN model has been described for stock price prediction. In stock price prediction, RNN processes these data time step by time by taking historical stock price data as an input sequence, and jointly processes the input of the current time step with the hidden state (ie, "memory") of the previous time step, and then Output the prediction result of the current time step, and at the same time pass the hidden state of the current time step to the next time step, so as to predict the future stock price trend.

However, this prediction method has long-term dependence problems, gradient disappearance and explosion problems, etc. LSTM (Long Short-Term Memory) is a special kind of RNN, which can effectively solve the long-term dependence problem of RNN, and can also control the memory and forgetting of information. In the article "Stock price prediction using LSTM, RNN and CNN-sliding window model[1]", the LSTM model predicts the stock price, and the comparison between the LSTM model and the RNN model prediction has been described. When LSTM is processed at each time step, it not only contains the input of the current time step and the hidden state of the previous time step, but also contains an internal state called "cell state". The cell state is controlled by a series of gating units, which determine whether the input of the current time step will be forgotten, whether the input of the current time step will be added to the cell state, and whether the cell state will be output. These gating units can adaptively control the flow of information by learning historical stock price data, so as to better predict future stock price trends. For example, Ghosh A et al.[4] have used LSTM model and companies' net growth calculation algorithm for Indian firms to analyze prediction of future growth of a company.

Although this prediction method can effectively predict stock prices, LSTM's predictions are not very stable for some uncertain results and relatively volatile stock prices, especially for changes in the stock market during the epidemic. This article will use the Bayesian neural network to predict the stock price based on the existing LSTM prediction. This prediction method can improve the accuracy and stability of the prediction. Nowadays Bayesian LSTM models have been used in many fields: for example, technology development[7], pollution control[8], etc., so we would like to try to apply it to stock prediction as well.

Meanwhile, for the stock data in the CRSP database used in this model, the data in this database has also been analyzed using other teams[9], but no one has previously been found to use this dataset for stock prediction, so we would like to train the prediction model in this scenario using a few representative company stock data of the moment.

2.2 Bayesian neural network

(1) Bayes theorem

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\int_H P(D|H')P(H')dH'} = \frac{P(D,H)}{\int_H P(D,H')dH'} \qquad (1)$$

Random variable H: hypothesis; random variable D: data; P(H|D): posterior probability; P(H): prior probability; P(D|H): likelihood probability)

(2) Bayesian Deep Learning

One of the simplest neuron network structures is shown in Fig 1:

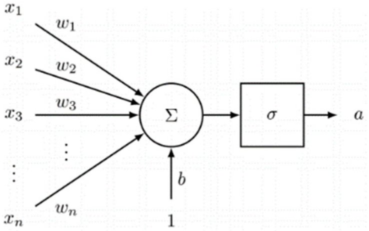

Fig. 1. the simplest neuron network structure

In deep learning, both wi(i = 1,...,n) and b are a certain value, such as w1=0.1, b=0.2. Bayesian deep learning is an interdisciplinary field of deep learning and Bayesian probability theory, and the uncertainty evaluation principle of deep learning architecture is given. In Bayesian deep learning, we change wi and b from certain values into distributions, that is, into the following formula:

$$P(w|D) = \frac{P(Dy|Dx,w)P(w)}{\int_w P(Dy|Dx,w')P(w')dw'} \qquad (2)$$

In this formula, the posterior probability, prior probability, and likelihood probability are no longer a specific number, but a distribution. By giving the input data x, the corresponding posterior distribution P(w|x,D) can be calculated, where w represents the weight parameter of the neural network, and D represents the training data set.

In this work, we utilize variational inference to implement Bayesian neural networks. Specifically, we represent the posterior distribution as a simple distribution family, such as the Gaussian distribution, and attempt to find the distribution parameters that best approximate the true posterior distribution. This allows us to approximate the posterior distribution using these parameters and obtain the posterior probability distribution based on Bayes' theorem. This process is similar to solving for maximum a posteriori estimation (MAP) given the model and data. Its advantage is that it incorporates prior knowledge and observational data, which can provide more robust estimates in the case of less data or more noise. The posterior probability is the probability distribution of parameter values under the condition of given observation data, which transforms the parameter estimation problem into an optimization problem, and selects parameter values by maximizing the posterior probability. This avoids overfitting and provides more accurate estimation results.

(3) LSTM Theory

LSTM is a variant of Recurrent Neural Network (RNN) for processing and predicting sequence data. The traditional RNN model has the problem of gradient disappearance or gradient explosion when dealing with long-term dependencies (long-term interval information), and LSTM solves this problem by introducing memory cells and gate mechanisms.

(4) Bayesian LSTM Theory

Bayesian LSTM is an extension of the Long Short-Term Memory (LSTM) model based on the Bayesian method. The traditional LSTM model uses a point estimation method for parameter learning, while Bayesian LSTM introduces the idea of Bayesian inference. By introducing the prior distribution of the parameters, the posterior distribution of the parameters can be obtained, and then a more comprehensive inference can be obtained.

3 experiments

3.1 Dataset Introduction

The data set selected in this experiment comes from CRSP Daily Stock. This data set is a financial data set managed by the CRSP (Center for Research in Security Prices) organization of the University of Chicago, which contains daily stock prices and transaction information of listed companies in the United States. The data set has been collected since 1925 and covers all listed companies on various exchanges in the United States, including NYSE, AMEX and NASDAQ.

In the CRSP Daily-Stocks dataset, each stock has a unique code called CRSP PERMNO (Permanent Identification Number), and each transaction has a unique identifier. The data set contains the basic information of the company, such as company name, industry classification and geographical location, as well as important indicators such as daily opening price, closing price, highest price, lowest price, trading volume and turnover.

This dataset is very useful for financial market analysis, portfolio construction, and risk management, and has become one of the widely used data sources in academia and the financial industry.

3.2 Data Selection

We selected six companies in the above dataset, including AAPL, FB, JPM, MSFT, PG, WMT. And export the following data from January 2, 2020 to December 30, 2022:

DATE: date

PERMNO: Stock unique identifier, representing each different stock in CRSP.

TICKER: The symbol representing the company's stock code, which can be a combination of numbers or letters.

BIDLO: The lowest buying price of the day, that is, the highest price a buyer is willing to buy.

ASKHI: The highest asking price of the day, that is, the lowest price that the seller is willing to sell.

PRC: The closing price of the day, that is, the final trading price of the stock on that day.

VOL: The trading volume of the stock on the day, that is, the number of stocks bought and sold on the day.

RET: The rate of return of the stock on the day, that is, the rate of change relative to the closing price of the previous trading day.

OPENPRC: The opening price of the day, that is, the price of the first transaction of the day.

vwretx: Market capitalization-weighted daily rate of return, representing the effect of investing in each stock in a portfolio.

Afterwards, we use DATE, OPENPRC, ASKHI, BIDLO, PRC, VOL, and TICKER as our analysis data, corresponding to the data names date, Open, High, Low, Close, Volume, and Name respectively. The collated data is shown in Table 1.

Table 1. The data

date	Open	High	Low	Close	Volume	Name
2019/01/02	91.64	93.650	91.64	93.34	8152733	WMT
2019/01/03	93.21	94.710	92.70	92.86	8277289	WMT
…	…	…	…	…	…	…

3.3 Data preprocessing

We standardize and normalize the five data of Open, High, Low, Close, and Volume according to formula (1), and change the data range to [-1,1], so that the influence of these five data volumes equal.

normalization:

$$\frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{3}$$

standardization:

$$\frac{X_i - \mu}{\sigma} \tag{4}$$

Where μ and σ represent the mean and standard deviation of the sample, X_{max} is the maximum value, and X_{min} is the minimum value.

3.4 Data processing

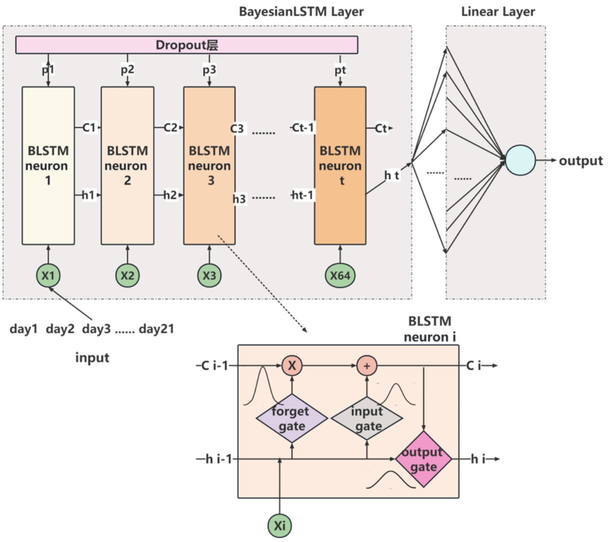

Fig. 2. The structure of Bayesian LSTM

This model uses Bayesian LSTM as the main neural network architecture. The specific structure of the model is shown in Fig 2. First of all, this model uses historical time series data as input data, and predicts the price at a certain time in the future based on the price at several moments in the past. These historical data are organized into a sliding window, and the window size is set to 21. Second, the model uses variational inference to learn parameter uncertainty and generate confidence intervals, that is, at each time step, the model generates multiple samples that are forward-passed using random weights and biases to obtain, and then by averaging the outputs of these samples, the final prediction result is obtained. In addition, considering that the data set is during the 2020-2022 epidemic period, the data volatility is relatively large. To avoid paying too much attention to outliers using mean squared error (MSE), this model uses a quantile loss function to train the network. This loss function can make the model better predict the tail distribution of stock prices, so it is more suitable for the stock market. During training, the Adam optimizer is used for parameter updates. Many thanks to Sen J's team[5] for their detailed explanation of the LSTM approach to stock forecasting. Also thanks to Karasan A's book[3] for mentioning finance related knowledge

3.5 Experimental results

(1) Evaluation indicators

We compare the model's predicted price with the real price quantile loss function, and use mse, rmse, mae to evaluate the prediction effect.

$$MSE = \frac{1}{n}\cdot\Sigma(y_i - \widehat{y_i})^2(5) \tag{5}$$

$$MAE = \frac{1}{n}\cdot\Sigma(y_i - \widehat{y_i}) \tag{6}$$

$$RMSE = (\frac{1}{n}\cdot\Sigma(y_i - \widehat{y_i})^2)^{\frac{1}{2}} \tag{7}$$

Where n represents the number of samples, yi represents the actual value, and ŷi represents the predicted value.

(2) forecast result

The predicted results are shown in Fig 3, where it can be visualized that BayesianLSTM can relatively accurately predict the next daily stock prices with confidence intervals.

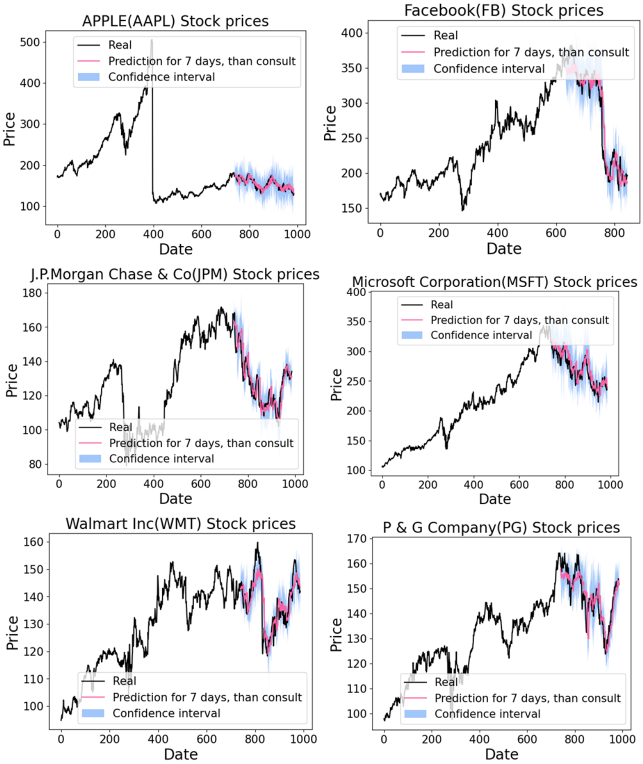

Fig. 3. The visualization of the prediction

3.6 Comparative Experiment

Table 2. The results of the comparative experiment

Company	Evaluation Method	Bayesian LSTM	RNN	LSTM
AAPL	mae	0.42	2.85	4.69
	mse	0.07	13.12	32.88
	rmse	0.26	3.62	5.73
FB	mae	1.02	5.86	13.21
	mse	0.51	86.28	381.81
	rmse	0.72	9.29	19.54
JPM	mae	1.03	1.97	4.30
	mse	0.48	6.47	29.79
	rmse	0.69	2.54	5.46
MSFT	mae	0.70	4.89	12.62
	mse	0.19	37.33	239.82
	rmse	0.43	6.11	15.49
PG	mae	1.00	1.65	3.38
	mse	0.40	4.98	16.88
	rmse	0.63	2.23	4.11
WMT	mae	1.04	1.81	3.53
	mse	0.51	6.64	24.97
	rmse	0.71	2.58	5.00

After prediction, with the data in Table 2, the mae, mse, and rmse of the results predicted by each company using BayseianLSTM perform better overall than the other two methods. Therefore, we concluded that the Bayesian LSTM model has a good effect on predicting future stock prices using historical 21-day data, compared with RNN and LSTM models. Especially for some discontinuous data, this model shows a better prediction effect than the other two models. Moreover, the confidence interval given by the Bayesian LSTM model can almost completely cover the real data, and has better accuracy and reference value.

3.7 Disturbance analysis

(1) The principle of perturbation analysis.

To assess the robustness of our model, we added different levels of Gaussian noise to the dataset. Specifically, we used four sets of Gaussian noise with mean 0 and variances of 0, 0.1, 0.2, and 0.3. The fluctuation in MAE, MSE, and RMSE values under different levels of noise were compared to determine the robustness of this experiment. Generally, as the variance of the added Gaussian noise increases, the disturbance becomes greater, and the performance of the model declines accordingly. However, if the model performs well under Gaussian noise with variances of 0, 0.1, 0.2, and 0.3, it can be concluded that the model has good robustness.

Gaussian noise is a continuous, symmetric probability distribution. Its probability density function can be expressed mathematically as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{8}$$

where μ is the mean and σ is the standard deviation. By adjusting the values for the mean and standard deviation, you can control the shape and magnitude of the Gaussian distribution.

(2) Implementation method

Firstly, a random number generator is needed to generate random numbers that follow the standard normal distribution (with mean 0 and standard deviation set to a series of values).

Secondly, linear transformations are applied to adjust the mean and standard deviation:

$$noise = \mu_n + \sigma_n * random\_number$$

Among them, μ_n is the mean value of the added noise, σ_n is the standard deviation of the added noise, and $random\_number$ is the generated random number of the standard normal distribution.

Finally, the resulting noise is added to the original numerical data to obtain Gaussian-noised data:

$$noisy\_data = original\_data + noise$$

For the sake of simplicity in the experiment, data from Apple Inc. was used for validation in this study.

(3)Comparison of analysis results

The experimental results are shown in Table 3:

Table 3. Results of Apple's data after different Gaussian noise disturbances

Gaussian noise	MAE	MSE	RMSE
Variance=0.0	0.39	0.06	0.25
Variance=0.1	0.42	0.07	0.27
Variance=0.2	0.38	0.06	0.24
Variance=0.3	0.43	0.07	0.27

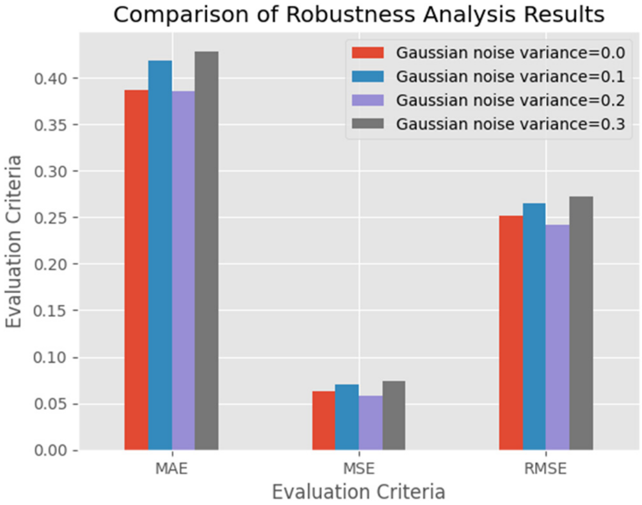

Fig. 4. Comparison of robustness analysis results

In this experiment, we added different levels of Gaussian noise to the input dataset and compared the results of MAE, MSE, and RMSE under different levels of Gaussian noise. As shown in Fig 4, the variation in MAE, MSE, and RMSE was negligible under different variances of Gaussian noise in this experiment. The overall fluctuation was within 0.1, indicating that the model has excellent robustness, which has been verified.

(4) Disturbance analysis for RNN and LSTM

In addition, we use RNN and LSTM to analyze the disturbance of Apple company, and the results are presented in Table 4 and Table 5 respectively.

Table 4. Results of Apple's data after different Gaussian noise disturbances under the RNN model

Gaussian noise	MAE	MSE	RMSE
Variance=0.0	1.05	1.71	1.31
Variance=0.1	1.88	5.30	2.30
Variance=0.2	3.16	12.94	3.60
Variance=0.3	7.97	68.09	8.25

Table 5. Results of Apple's data after different Gaussian noise disturbances under the LSTM model

Gaussian noise	MAE	MSE	RMSE
Variance=0.0	1.22	2.21	1.49
Variance=0.1	10.55	176.34	13.28
Variance=0.2	20.17	659.56	25.68
Variance=0.3	30.30	1485.16	38.54

Since the data we used included data during the COVID-19 pandemic, there was a discontinuity of data, which led to relatively large instability in LSTM. This also illustrates the universality of BayesianLSTM when dealing with different extreme data, further proving that the Bayesian LSTM model has good robustness.

4 Conclusion

To sum up, the Bayesian LSTM model has a good ability to process time series data compared with the RNN and LSTM models. For some discontinuous data, especially the volatility data caused by the 2020-2022 epidemic, Bayesian LSTM The model shows strong learning and prediction ability, gives accurate confidence interval, and has good robustness.

5 Outlook

The Bayesian LSTM model has broad application prospects in the field of time series data processing. With the continuous development of technology and the continuous accumulation of data, researchers may continue to work on improving the structure, algorithm and robustness of the Bayesian LSTM model to improve its modeling ability and prediction accuracy for time series data, and better accurately capture key information in the sequence.

In addition, the Bayesian LSTM model not only has application potential in the financial field, but also can play a role in other fields. For example, in the medical field, the Bayesian LSTM model can be used to process the time series data of diseases to realize the prediction and intervention of patients' conditions.

We believe that future research will continue to promote the development of Bayesian LSTM models to cope with more complex and diverse data requirements, bringing more accurate and reliable prediction and decision support.

References

1. S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1643-1647, doi: 10.1109/ICACCI.2017.8126078.
2. Zhu Y. Stock price prediction using the RNN model[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1650(3): 032103.
3. Karasan A. Machine Learning for Financial Risk Management with Python[M]. " O'Reilly Media, Inc.", 2021.
4. Ghosh A, Bose S, Maji G, et al. Stock price prediction using LSTM on Indian share market[C]//Proceedings of 32nd international conference on. 2019, 63: 101-110.
5. Sen J, Mehtab S, Nath G. Stock price prediction using deep learning models[J]. Lattice: The Machine Learning Journal, 2020, 1(3): 34-40.

6. Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market[C]//2015 IEEE international conference on big data (big data). IEEE, 2015: 2823-2824.
7. Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM[J]. International Journal of Science and Research (IJSR), 2017, 6(4): 1754-1756.
8. Han Y, Lam J C K, Li V O K, et al. A Bayesian LSTM model to evaluate the effects of air pollution control regulations in Beijing, China[J]. Environmental Science & Policy, 2021, 115: 26-34.
9. Evgeniou T, Vermaelen T. Share buybacks and gender diversity[J]. Journal of Corporate Finance, 2017, 45: 669-686.