

Application of LSTM and portfolio optimization in Chinese stock market

Zichun Zheng

Herbert Business School, University of Miami, Hui Zhong Bei Li 117-3-602, Beijing, China
zhengzc@stu.cpu.edu.cn

Abstract. The purpose of this paper is to examine the application of LSTM and mean variance portfolio optimization in Chinese stock market. 20 stocks are selected from CSI 300 components, we collect their High, Low, Open, Adjust Close and trade volume from June 16th 2020 to June 16th 2022. Then we use LSTM model to forecast the stock price. The forecast results are used to construct 2 portfolios. One portfolio maximize Sharpe ratio, the other portfolio minimize variance. From April 6th to June 16th 2022, the Maximize Sharpe Ratio portfolio outperformed CSI 300 index, the Minimize Variance Portfolio did not beat the market but the return was very close to CSI. Therefore, the combination of LSTM and Mean Variance Portfolio Optimization theory is effective in Chinese stock market.

Keywords: Chinese stock market, Machine learning, deep Learning, LSTM, Portfolio Optimization.

1. Introduction

Chinese stock market has grown at an astonishing rate in recent years. With the development of Chinese stock market, research on Chinese market gradually increase. A study by Lim, Huang, Yun and Zhao in 2013 indicates that China's stock market is weak-form efficiency [1]. Therefore, China's stock market appears to be more volatile and riskier than the developed market. A study by Wang and Jia suggests that Chinese investors' judgment on Chinese stock market valuation is greatly influenced by their confidence variation in domestic economic fundamentals [2]. There are also researches on application of machine learning/ deep learning in investment. Prasad and Seetharaman discussed the importance of machine learning in making investment decision in Stock Market [3]. Akhtar, Zamani, Khan, Shatat, Dilshad and Samdani made an attempt to predict the stock market based on statistical data using machine learning algorithms [4]. Agrawal, Shukla, Nair, Nayyar and Masud tried to predict stock market trend using technical indicators [5]. As the classical theory of investment field, Markowitz's portfolio optimization, needless to say there are huge amount of research on this field. However, research on the combination of LSTM and portfolio optimization in Chinese stock market is not yet seen, this paper fills the gap of this field.

Chinese stock market is known for its volatility. The standard deviation of the Nasdaq index, S&P 500 and CSI 300 for the last 10 years are 2.44, 2.16 and 2.97. Volatility of CSI 300 is significantly higher than that on the NASDAQ and S&P 500CSI 300. From July 2014 to June 2015, Shanghai Securities Composite Index (000001.SH) went from around 2000 to 5178, went up for more than 150% in less than a year. Then immediately crashed to 2600 in less than 6 months.

Considering the risk level and the market efficiency, the purpose of this study is to explore the combination of using machine learning in predicting stock price movement and using portfolio optimization theories to construct portfolios that can potentially beat the benchmark in China's stock market.

This paper uses the CSI300 index to represent China's stock market and as the benchmark, then from the components of the CSI300 index we select 20 stocks based on their market capitalizations. LSTM neural network is used to forecast the stock price movements, then we use the forecast results to construct the portfolios. Monte Carlo simulation is used to create the efficient frontier, from which we get the maximize Sharpe Ratio portfolio and minimize variance portfolio. During the testing period, the maximize Sharpe ratio portfolio outperform the CSI300 index but the minimize variance portfolio did not.

2. Method

2.1 LSTM

In recent years, financial time series forecasting has attracted substantial attention [6]. Financial time series forecasting has many applications in securities investment and risk management. There are several traditional models to forecast the time series data, such as Autoregressive Integration Moving Average (ARIMA), Simple Moving Average (SMA) and Linear Regression Models. Many complex models such as Neural Networks Autoregression (NNAR) and RNN (Recurrent Neural Network) may be useful in certain cases to forecast time series data [7].

LSTM is an advanced recurrent neural network (RNNs) that are capable of learning long-term dependencies and is very useful in time series forecasting problems. LSTM units include a 'memory cell' that can maintain information in memory for long periods of time. This memory cell lets them learn longer-term dependencies. LSTM has feedback connections, which means it is capable of processing the entire sequence of data, apart from single data points, which shows outstanding performance on a large variety of problems.

There are many attempts to incorporate LSTM with investment. Preeti, R. Bala and R. P. Singh's work in 2019 confirms that LSTM is effective for time series prediction, even in some instances where the data is non-stationary [8]. Vora, Shaikh and Bhanushali's research using LSTM models with features including open, closed, lowest, highest, date, and everyday transaction size of stock data to forecast stock prices in India market suggests that although LSTM model has few restraints including a forecast time lag, but can still use the attention level to foretell stock prices [9]. A work by Rather proposed a new method of predicting time-series-based stock prices and a new model of an investment portfolio based on forecast results from LSTM model. Their results show that their model outperforms various standard predictive models [10].

In this paper, for stock returns forecasting process, the idea is to incorporate machine learning/deep learning in predicting stock price movement. The chosen model is LSTM neural networks. The following are the steps of the model.

The first step is to define a simple function to process raw data. This function is supposed to return the prepared data with different memory days and different predict days. By creating a new feature named 'Label' to store the real adjusted close price after 50 days, the model is able to compare the real value with our results later. The train set and the test set are split at the proportion of 8 to 1.

The second step is to prepare some parameters to find an optimal model for each stock. To run the code more efficiently, we suppose the memory days is 5, the number of the LSTM layers is either 1 or 2, the number of dense layers is either 1 or 2 and the number of units of each layer is either 16 or 32. The parameters of model have 8 possible permutations so that each stock needs to run the code 8 times for the model to learn.

In the third step, use for-loop to build LSTM model. The first loop is for the LSTM layers, and the second loop is for the dense layers. To compile the model, the code will calculate mean absolute percentage error (mape) to measure the model's accuracy and the model with the smallest error will be the final choice.

The last part is the execution of the best model. the best model with smallest mape is chosen and saved from the third step, then the best model is loaded and used to predict the adjusted close price of the test set.

2.2 Mean Variance Portfolio Optimization

The goal of the portfolio optimization process is to select the best portfolio considering maximizing or minimizing the factors such as expected return or volatility. In Markowitz's Modern portfolio theory, it assumes that for a certain amount of risk, the investors want to maximize a portfolio's expected return. In this paper, the forecast results from LSTM are used in the portfolio optimization process.

The process of finding the efficient frontier is simplified to two steps for a stock-only portfolio in this case. Firstly, construct the efficient frontier using Monte Carlo simulation which generates a series of random weights for each asset used for calculating the portfolio features. Then pick out the targeted portfolio. There are several optimal portfolios such as risk parity portfolio or inverse volatility portfolio. Among all the portfolios, Maximum Sharpe Ratio portfolio and Minimum Volatility portfolio are chosen.

In the mean variance portfolio, the mean return and the variance are two important performance measures. For higher mean return, an investor must take on more risk, in this case it means more variance, and vice-versa. They define a fundamental mean-variance trade-off. The risk tolerance and risk appetite translate to the choice of a specific point in this trade-off curve. Maximizing the Sharpe ratio leads to a better risk-adjusted performance, and minimum volatility strategies helps investors stay invested in the market by seeking to minimize the equity risk while providing equity market exposure.

Some important details are shown below:

$$E_p = \omega^T R \quad (1)$$

$$V = \omega^T \Omega \omega \quad (2)$$

$$SR = \frac{E_p - R_f}{\sqrt{V}} \quad (3)$$

Where ω represents the weights of assets in the portfolio, R and Ω refer to the expected return and variance-covariance matrix of the selected asset. R_f is the risk-free rate.

3. RESULTS

From the stocks prices forecasted using the LSTM model, the daily returns of the 20 considered stocks can be calculated. The average return and the covariance matrix are then available. Applying the Monte Carlo simulation, random weights are generated for the chosen stocks in 100000 simulated portfolios, after which the portfolio return, volatility and Sharpe ratio can be calculated. By plotting the portfolios by their return and volatility, the efficient frontier is formed. Specifically, the empirical results in this paper are shown below.

The index that is chosen is The Shanghai Shenzhen CSI 300 index. CSI 300 is one of China's most closely followed stock market indices. It is a capitalization-weighted index consist of the top 300 stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange. CSI 300 is also seen as indicative of trends in both those markets. Therefore, CSI 300 index can be used to represent the Chinese stock market and as the benchmark for the portfolios. The components of the index formed the stock pool.

20 stocks were selected from the stock pool based on their market capitalizations (China Mobile was excluded due to lack of historical data). Table 1 shows the 20 stocks that were chosen, the time period of the data is from June 16th 2020 to June 16th 2022. The next step is to predict the price movement of these 20 stocks, then construct a portfolio.

Table 1: 20 chosen stocks

Ticker	Name	Ticker	Name
600519.SH	KWEICHOW MOUTAI	601628.SH	CHINA LIFE
601398.SH	ICBC	000858.SZ	WULIANGYE
601939.SH	CCB	601088.SH	CHINA SHENHUA
601857.SH	PETROCHINA	600900.SH	CYPC
300750.SZ	CATL	600028.SH	SINOPEC CORP.
601288.SH	AGRICULTURAL BANK OF CHINA	601658.SH	POSTAL SAVINGS BANK OF CHINA
600036.SH	CMB	601012.SH	LONGI
002594.SZ	BYD	601166.SH	INDUSTRIAL BANK
601988.SH	BANK OF CHINA	000333.SZ	MIDEA GROUP
601318.SH	PING AN OF CHINA	603288.SH	HAI TIAN

After the stocks are selected, the LSTM neural network is built using the 5 features which are HIGH, LOW, OPEN, adj CLOSE and TRADE VOLUME. As mentioned above, the best model is selected based on its accuracy which is measured by mape. The mape of the best model of each stock and their ranks are shown in Table 2.

Table 2: rank of each stock's mape

Rank	Ticker	Name	mape(%)
1	601288.SH	AGRICULTURAL BANK OF CHINA	0.9958
2	601398.SH	ICBC	1.4662
3	601939.SH	CCB	1.7748
4	600028.SH	SINOPEC CORP.	2.1575
5	601988.SH	BANK OF CHINA	2.3195
6	600900.SH	CYPC	2.5093
7	600519.SH	KWEICHOW MOUTAI	3.1626
8	601658.SH	POSTAL SAVINGS BANK OF CHINA	3.2853
9	601166.SH	INDUSTRIAL BANK	3.5138
10	601318.SH	PING AN OF CHINA	3.559
11	601857.SH	PETROCHINA	3.5711
12	600036.SH	CMB	7.3307
13	601088.SH	CHINA SHENHUA	8.2626
14	601012.SH	LONGI	9.3658
15	002594.SZ	BYD	11.8689
16	300750.SZ	CATL	12.0419
17	601628.SH	CHINA LIFE	15.1025
18	000333.SZ	MIDEA GROUP	18.3712
19	603288.SH	HAI TIAN	18.8436
20	000858.SZ	WULIANGYE	26.2562

The average return and the covariance matrix are then available. Applying the Monte Carlo simulation, random weights are generated for the chosen stocks in 100000 simulated portfolios, after which the portfolio return, volatility and Sharpe ratio can be calculated. The efficient frontier is gained by plotting the portfolios by their return and volatility. The efficient frontier is gained by plotting the portfolios by their return and volatility, as shown in Figure 2. The black markers on the plot are the portfolios which consist of only one stock.

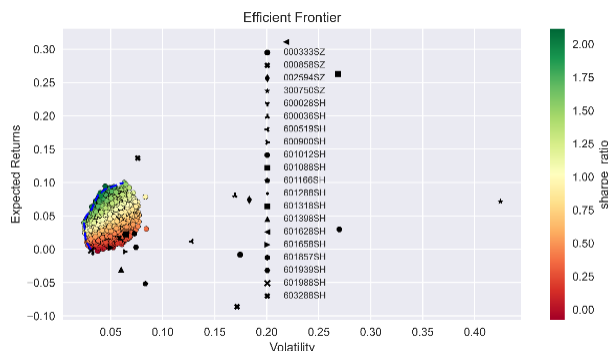


Figure 1: Efficient frontier

The two targeted portfolios, as mentioned before, are the Maximum Sharpe Ratio portfolio and the Minimum Volatility portfolio both of which have been marked on Figure 2. The portfolios would give the specific weights of the 20 stocks which allows us to calculate the real portfolio return in the next step.

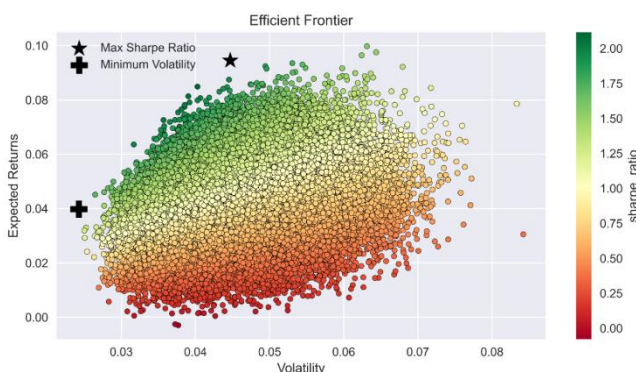


Figure 2: Maximum Sharpe Ratio and Minimum Volatility portfolio

Table 3: The specific weights for the Maximum Sharpe Ratio Portfolio

Ticker	Weight	Ticker	Weight
000333SZ	1.23%	601166SH	2.50%
000858SZ	0.37%	601288SH	0.94%
002594SZ	3.39%	601318SH	5.02%
300750SZ	3.94%	601398SH	1.72%
600028SH	4.57%	601628SH	12.06%
600036SH	6.94%	601658SH	9.10%
600519SH	10.11%	601857SH	8.06%
600900SH	0.83%	601939SH	1.17%
601012SH	0.63%	601988SH	4.82%
601088SH	9.65%	603288SH	12.94%

The Maximum Sharpe Ratio Portfolio using the predicted prices has a return of 9.46%, the volatility is 4.47%, the Sharpe ratio is 211.63%. In this portfolio, KWEICHOW MOUTAI (600519SH), CHINA SHENHUA (601088SH), CHINA LIFE (601628SH), POSTAL SAVINGS BANK OF CHINA (601658SH), PETROCHINA (601857SH), and HAI TIAN (603288SH) each has a weight of more than 8%, altogether they made up 61.92% of the total weight. KWEICHOW MOUTAI and HAI TIAN are from the food and beverage sector, CHINA LIFE and POSTAL SAVINGS BANK OF CHINA are from the bank and insurance sector, PETROCHINA are from the energy sector. The possible reason is, the banking sector and energy sector's valuation during the testing period is at a historic low, so the stocks are cheaper than other sectors. The food and beverage

sector's valuation is at the 10-year average level, during that time China experienced a hardcore COVID lockdown in April and slowly open up in May and June, the demand of food and beverage increase as the lockdown restriction lifted.

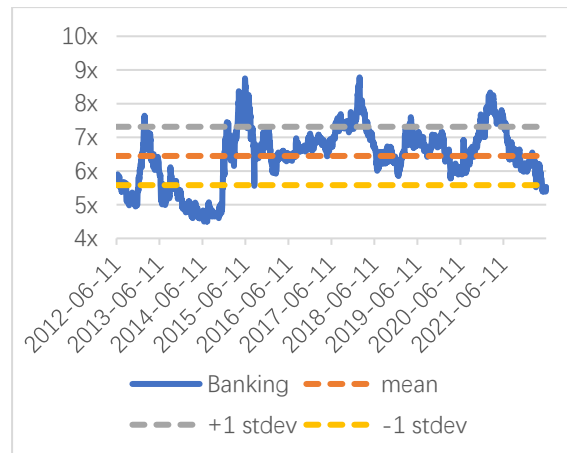


Figure 3: PE ratio of banking sector

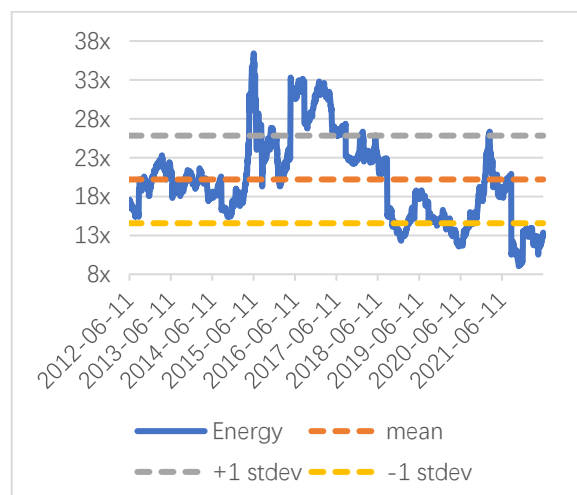


Figure 4: PE ratio of energy sector

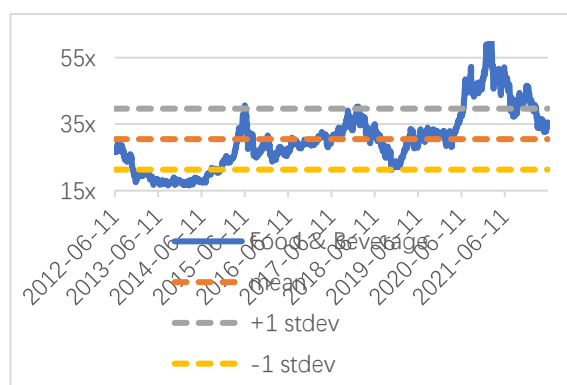


Figure 5: PE ratio of food and beverage sector

Table 4: The specific weights for the Minimum Volatility portfolio

Ticker	Weight	Ticker	Weight
000333SZ	0.86%	601166SH	2.63%
000858SZ	1.39%	601288SH	12.01%
002594SZ	1.56%	601318SH	4.84%
300750SZ	0.83%	601398SH	1.42%
600028SH	1.04%	601628SH	3.71%
600036SH	3.40%	601658SH	10.36%
600519SH	2.70%	601857SH	8.93%
600900SH	10.81%	601939SH	11.32%
601012SH	0.49%	601988SH	8.37%
601088SH	4.34%	603288SH	8.99%

The Minimum Volatility portfolio using the predicted prices has a return of 3.98%, the volatility is 2.43%, the Sharpe ratio is 164.06%. In this portfolio, CYPC (600900SH), AGRICULTURAL BANK OF CHINA (601288SH), POSTAL SAVINGS BANK OF CHINA (601658SH), PETROCHINA (601857SH), CCB (601939SH), BANK OF CHINA (601988SH), and HAI TIAN (603288SH) each has a weight of more than 8%, altogether they made up 70.79% of the total weight. CYPC and PETROCHINA are from the energy sector; AGRICULTURAL BANK OF CHINA, AGRICULTURAL BANK OF CHINA and BANK OF CHINA are from the banking sector; HAI TIAN is from the food and beverage sector. For the minimum variance portfolio, the possible reason for this outcome is the banking sector is considered as a traditional defensive sector, volatility of large cap blue chip companies is much smaller than the small cap companies, therefore it can benefit the purpose of minimizing risk. The energy sector is at a low value, so the stock price is less likely to go down a lot.

Applying the real market data of the 20 stocks from April 6th to June 16th, 2022, using the weights above for the two portfolios, the real returns of the two portfolios can be calculated. The cumulative returns of the two portfolios along with the cumulative return of CSI 300 are shown in Figure 4. During this period, the Maximum Sharpe Ratio Portfolio yield a return of 1.45%, and the Minimum Volatility portfolio yield a return of -0.43%, both are lower than the predicted returns. Comparing to benchmark, during the same period, the CSI 300 yield a return of -0.32%. The Maximum Sharpe Ratio Portfolio successfully outperformed the benchmark by 1.77%. While the Minimum Volatility portfolio didn't beat the benchmark, the result is very close, also the volatility of the portfolio is much smaller than the CSI300 index, so the purpose of minimize the risk is achieved.

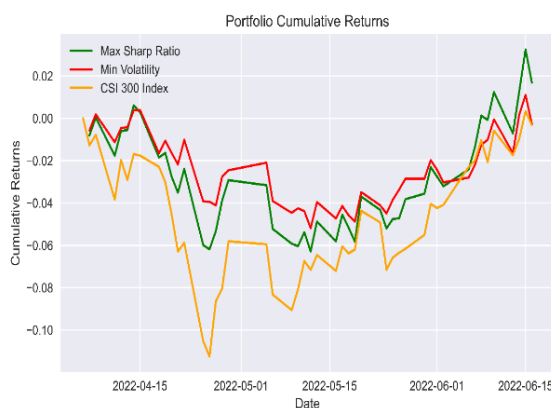


Figure 6: Portfolio Cumulative Returns

4. Conclusion

From our study, we use LSTM model to forecast stock price using empirical data including High, Low, Open, Adjusted Close and Volume. The forecast results are then used to construct two portfolio using Mean Variance Portfolio theory, one portfolio maximize Sharpe Ratio and the other portfolio minimize variance. The test results using real stock price data from April 6th to June 16th 2022 show that the Maximize Sharpe Ratio Portfolio outperform CSI 300 index, the minimize variance portfolio did not beat the CSI 300 index but the return is very close, also the volatility of the minimum variance portfolio is much smaller than the CSI300 index. This study shows that although China's stock market is relatively immature and riskier than the developed market, the combination of LSTM model and Mean Variance Portfolio Optimization theory is still effective. Although the forecast of stock price using LSTM has few restraints including a forecast time lag, we can still use the attention level to foretell stock prices. Mean Variance Portfolio Optimization theory can be used by institutional investors to construct their portfolio based on their risk appetite.

References

- [1] Lim, Thian. (2013). Has Stock Market Efficiency Improved? Evidence from China. *Journal of Finance & Economics*. 1. 1-9.
- [2] Dao-ping Wang, Yu-ning Jia. Investor Sentiment and Excess Volatility of Chinese Stock Markets. *Finance Forum*, 2019, 24(7):14.
- [3] Prasad, Akhilesh & Seetharaman, Arumugam. (2021). Importance of Machine Learning in Making Investment Decision in Stock Market. *Vikalpa: The Journal for Decision Makers*. 46. 025609092110599.
- [4] Akhtar, Md & Zamani, Abu & Khan, Shakir & Shatat, Abdallah & Dilshad, Sara & Samdani, Faizan. (2022). Stock market prediction based on statistical data using machine learning algorithms. *Journal of King Saud University - Science*. 34. 101940.
- [5] Agrawal, Manish, Asif Ullah Khan, and Piyush Kumar Shukla. "Stock price prediction using technical indicators: a predictive model using optimal deep learning." *Learning 6.2* (2019): 7.
- [6] Y. S. Abu-Mostafa and A. F. Atiya, "Introduction to financial forecasting," *Applied Intelligence*, vol. 6, no. 3, pp. 205–213, 1996.
- [7] Li S, Li W, Cook C, et al. Independently recurrent neural network (indrnn): Building a longer and deeper rnn[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 5457-5466.
- [8] Preeti, R. Bala and R. P. Singh, "Financial and Non-Stationary Time Series Forecasting using LSTM Recurrent Neural Network for Short and Long Horizon," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-7
- [9] Vora, Sakshi & Shaikh, Rayees & Bhanushali, Kartik & Patil, Prof. (2022). Stock Price Prediction using LSTM. *Indian Journal of Artificial Intelligence and Neural Networking*. 2. 1-5.
- [10] Rather, Akhter. (2021). LSTM-based Deep Learning Model for Stock Prediction and Predictive Optimization Model. *EURO Journal on Decision Processes*. 9. 100001.