Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(1): 174–192.
DOI: 10.3934/era.2024009
Received: 02 August 2023
Revised: 08 November 2023
Accepted: 20 November 2023
Published: 18 December 2023

Research article

Application of machine learning in quantitative timing model based on factor stock selection

Yufei Duan[1], **Xian-Ming Gu**[1,*] **and Tingyu Lei**[2]

[1] School of Mathematics, Southwestern University of Finance and Economics, Chengdu 611130, China

[2] School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, China

* **Correspondence:** Email: guxm@swufe.edu.cn; Tel: +862887092087.

Abstract: In this paper, we integrated machine learning into the field of quantitative investment and established a set of automatic stock selection and investment timing models. Based on the validity test of factors, a multi-factor stock selection model was established to select stocks with the highest investment value to create a stock pool. By comparing the cumulative returns and the overall market returns of different timing signals over the same time period, both the decision tree and the long short-term memory (LSTM) models had great results. Finally, empirical research was reported to show that it is a good combination to introduce machine learning algorithms into quantitative timing.

Keywords: quantitative investment; multi-factor model; decision tree model; LSTM model

1. Introduction

With timeliness and accuracy characteristics, quantitative investment has become a new and popular investment method in the global investment field. Based on a vast quantity of data, machine learning algorithms can learn models with great generalization performance. It is what quantitative investment requires [1]. As the effectiveness of traditional multi-factor stock selection strategies gradually declines, the adoption of machine learning algorithms to optimize stock selection strategies has become a popular trend [2].

The multi-factor model is a model developed from the asset pricing model. It takes into account a combination of factors, which is very sensitive to market fluctuations and changes in strategies. There have been many theoretical and empirical studies on the asset pricing model. They have improved from a single-factor model to a five-factor model, from simply considering market risk factors to considering a wider range of factors, such as technical index factors. Zhao et al. [3] carefully sorted out the various factor models and analyzed their advantages and disadvantages in detail. They used the Fama-French five-factor pricing model to analyze China's stock market and found that the regression coefficients of CMA (investment factors in the Fama-French five-factor model) and RMW (profitability factors in

the Fama-French five-factor model) are not significant in China, which means the explanatory power of asset pricing models varies with the level of capital market development. In addition, to fully consider the effect of factors, Wang et al. [4] constructed the factor database using the financial index indicators, technical index indicators and public opinion. They used the neural network to describe the relationship between stock factors and individual stock excess returns, selecting stocks with the largest rise probability to form the portfolio. However, the selection of factors in Wang's article is based solely on the previous research experience, and the research lacks a test of validity of factors.

N. Nguyen and D. Nguyen [5] adopted the hidden Markov model (HMM) to predict regimes of six global economic indicators. Based on this, they analyzed the stock performance in the identified time periods and assigned weight for the stock factors. By selecting the top 10% in the global markets, they traded stocks with the highest composite scores. In addition, it is worth noting that Baykasoğlu and Gölcük used multiple attribute decision making (MADM) to address poorly defined problems with multiple and interrelated criteria [6]. Therefore, according to the above researches, we decide to start in a different direction, focusing on stock factors and selecting more factors to score stocks [7]. In this paper, we use stocks from the Shanghai Stock Exchange and select four characteristic indicators of value factor, growth factor, size factor and trading factor, with a total of 23 factors, then we utilize the ranking method to test the validity of the factors. This method integrates a variety of information and is relatively stable, which means it is a good choice for testing the validity of factors. In addition, considering the different trading concepts of investors and information environments in the Chinese stock market and other mature stock markets, we rebuild a stock selection model using A-share data [8].

At present, there are many studies on stock market forecasting. Jiang [9] has made a comprehensive comparison of the common research methodology, object and process. He pointed out that in the step of collecting data, the common types of data nowadays includes market data, text data, macroeconomic data etc., and that market data is the most frequently used data. This author also summarized that the state-of-the-art predictive models can be categorized into standard, hybrid and other models. Standard models include feed-forward neural network (FFNN), convolutional neural network (CNN) and recurrent neural network (RNN). The long short-term memory (LSTM) model is an RNN model. Hybrid models are the combination of deep learning and traditional models or different deep learning models. Sonkavde et al. [10] analyzed various existing machine learning algorithms, including time series models, deep learning models and integrated learning methods. They deeply compared the three models. The authors mentioned that there is no generalized method to accurately predict stock prices. They also predict that in the future, trend analysis may become the focus of stock market forecasting. As for quantitative timing strategy, it is not difficult to find that the research models can be divided into four types: Traditional timing model, decision tree model, the LSTM exponential quantization model, and implicit HMM. Tenti [11] employed machine learning techniques to mine technical indicators such as the average trend index, movement index and change rate in order to determine the price trend of financial assets. Tay and Cao [12] found that the support vector machine has a higher accuracy than the neural network in future forecasting. The profitability of the timing strategy created by an algorithm for machine learning is greater than that of the market portfolio. Consequently, technical analysis based on machine learning has become a reliable method for predicting the price of financial assets. To improve the generalization ability and meet the demands of dynamic behavior of trading action execution, Deng et al. [13] introduced the contemporary deep learning into a typical direct reinforcement learning framework, but such a research only handled one share of the asset. Another research just considered the price trend of financial assets when building the model, without considering the range of increases or decreases [14]. Therefore, our study develops a quantitative timing system that is capable of manag-

ing a number of assets simultaneously. In addition, it proposes the index construction method, which takes into account the price fluctuation range in order to execute stop profit or stop loss operations. Drawing on the above research results, this paper adopts market data, selects the decision tree model and the LSTM model as the timing model and determines the optimal investment time based on the trend of stock price changes.

The meaning of the technical indicator value is unique. Using the characteristics of technical indicators allows for more accurate forecasting of the future price trend of financial assets. Patel [15] presents a trend deterministic data preparation layer (TDDPL) approach to remedy the aforementioned issues. We use this method to discrete continuous technical index values to highlight the characteristics of each technical index, thereby increasing the accuracy of the machine learning model's predictions. Note that we aim to use a neural network algorithm to optimize and improve the three index parameters of the moving average convergence divergence (MACD), a quantitative timing strategy commonly used in the stock market. In other words, we are establishing the LSTM and MACD timing investment strategy.

It is worth emphasizing that we do short-term forecasting and do more innovative decision tree index screenings as well. Compared with the other state-of-the-art algorithms, we have not only selected state-of-the-art deep learning algorithms for stock prediction, but also linked factor-based stock selection with quantitative timing to build a fully automated stock trading model.

The remainder of this manuscript is organized as follows. Section 2 describes the construction of the multi-factor model and Section 3 describes different quantitative timing models constructed by two algorithms, the decision tree model and the LSTM model. The results of an empirical application are discussed in Section 4. We conclude the paper with some remarks in Section 5.

2. Multi-factor model

This paper first selects 23 candidate factors in four categories based on the company's fundamentals by referring to relevant literature*. We then use the ranking and scoring sorting method to conduct the validity test and eliminate the factors that have little correlation with stock returns or poor stock selection ability. Based on the weight of the selected factors, the top 10 stocks can be selected.

Data used in this paper is based on the Shanghai stock market collected from the Tushare package† and JointQuant platform‡.

2.1. Candidate factors

Based on market experience and economic logic, selecting more effective factors can enhance the ability to capture model information. Thus, 23 candidate factors in total are chosen from four categories in relevant papers including value, growth, size and trading. The factors are proposed as Table 1.

*https://zhuanlan.zhihu.com/p/20634542 (accessed on 10 August 2022)
†http://localhost:8888/edit/tushare%E7%88%AC%E5%8F%96%E6%95%B0%E6%8D%AE(1).py (accessed on 5 September 2022)
‡https://www.joinquant.com (accessed on 29 November 2022)

Table 1. Candidate factors.

Category	Factors
Value Class Factor	price earnings ratio (PE)
	price-to-book ratio (PB)
	price-to-sales ratio (PS)
	basic earnings per share (EPS)
	book-to-market ratio (B/M)
Growth Class Factor	return on equity (ROE)
	return on assets (ROA)
	gross profit margin
	net profit growth year-on-year
	net profit growth rate month-on-month
	operating profit growth rate year-on-year
	operating profit growth rate month-on-month
	main gross profit margin
	net profit margin (P/R)
Size Factor	net profit
	operating income
	total equity
	outstanding share capital
	total market capitalization
	circulating market capitalization
	assets and liabilities (L/A)
	fixed assets ratio (FAP)
Trading Class Factor	turnover rate

2.2. Factor validity test

Using data from 2014 to 2020 within 1489 stocks, we divide stocks into five groups according to the circulating market capitalization (CMC) and set the Shanghai Composite Index as the benchmark group. We calculate the monthly returns of six groups of stocks weighted by CMC and add up the total returns of each group in the past 7 years. Two standards are used to examine the validity of factors.

a) The factor correlation > 0.7 or < -0.7;
b) The winner portfolio wins and the loser portfolio loses for a probability above 0.6.

According to Table 2[§], we get seven effective factors, including EPS, L/A, PE, PS and Gross Profit Margin, then we calculate the total and annualized return of the six stock portfolios based on the effective factors[¶].

2.3. The construction of multi-factor model

In this sector, we use equal weights to sum factor scores and select the stocks with the highest score to trade. The corresponding formula is:

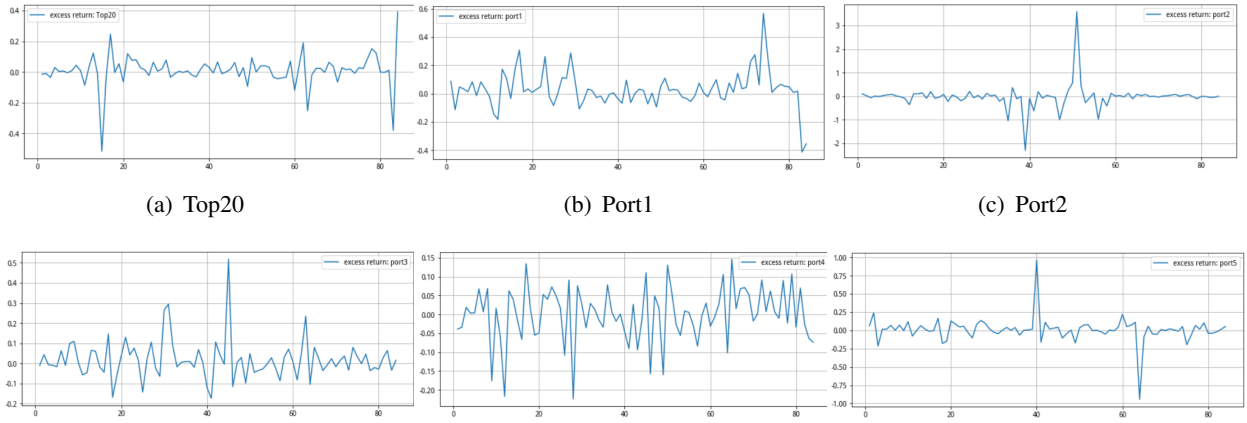
$$E\left[R^e\right] = \alpha + \sum \beta_i \lambda_i, \tag{2.1}$$

[§]The results of the validity tests for all factors are presented in Appendix (i.e., Figure A1).
[¶]The total and annualized return is shown in Appendix (i.e., Figure A2).

Table 2. Validity test.

Factors	Factor Relevance	The Probability of Winning and Losing For Portfolio
EPS	0.711845	[0.678571428571, 0.404761904762]
L/A	-0.868283	[0.702380952381, 0.428571428571]
PE	-0.842341	[0.690476190476, 0.452380952381]
PS	0.857868	[0.678571428571, 0.452380952381]
Gross Profit Margin	0.859684	[0.714285714286, 0.416666666667]

where $E[R^e]$ is stock excess return, α refers to the error term, β_i refers to factor exposure and λ_i refers to factor excess return. i equals $1, 2, \cdots, 7$, then we pick out 10 valuable stocks from the Shanghai stock market ranging from high to low scores: 603040.SH, 688399.SH, 600749.SH, 600865.SH, 603156.SH, 603258.SH, 603087.SH, 603444.SH, 688188.SH, 600674.SH. The time series plots of monthly excess return below illustrate that the top 20 stocks have had more stable earnings in the past 7 years.



(a) Top20 (b) Port1 (c) Port2

(d) Port3 (e) Port4 (f) Port5

Figure 1. Back-test excess return of factor.

2.4. Model correction

After the multi-factor model is built, we successfully select ten stocks out of the pool. However, we find that our long-term holding yield is negative. One possible explanation is that our output based on 7 years of historical data has limitations. We believe that this phenomenon occurs as a result of the overly adequate selection of historical data and the long time interval between the selected data. Therefore, we alter for 2 year historical data from 2019 to 2020, with 1497 stocks in total. Surprisingly, the long-term holding yield becomes positive. Hence, by recalculating the features of each stock, we get a new series of effective factors and stock selection results: 603199.SH, 688366.SH, 600830.SH, 688188.SH, 600052.SH, 688111.SH, 688016.SH, 688020.SH, 688019.SH, 603087.SH. We choose 20 stocks based on the new model and calculate their excess monthly return to prove that the new model performs better than the original model. The results are shown in Figure 2.

The article also does a back-test. The purpose is to verify the feasibility and effectiveness of the trading strategy based on historical data, hoping to use performances after the back-test to evaluate the real future performance, thereby saving the opportunity cost of choice. In fact, the model does report

a higher holding period return compared to the original multi-factor model. The results are shown in Table 3.

Table 3. Return of different methods.

Methods	Annual Return
Origin Multi-factor model	-0.1370
Revised Multi-factor model	-0.0805

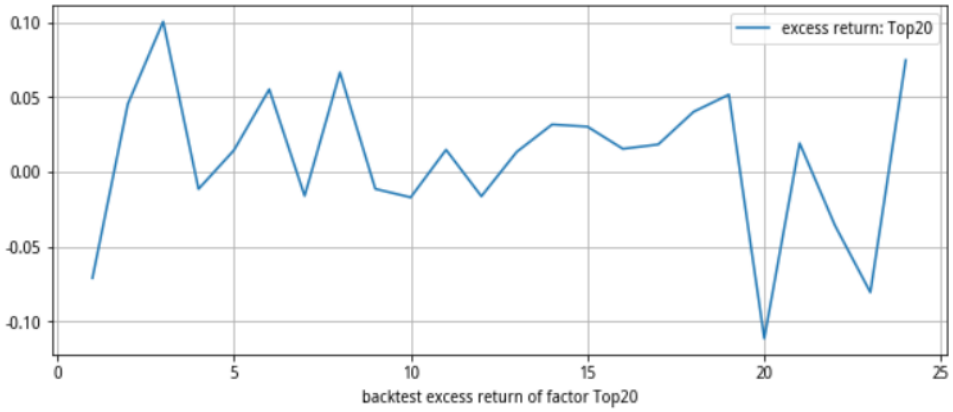


Figure 2. Backtest excess return of factor top 20.

3. Quantitative timing model

Quantitative timing is an important area of research in quantitative trading. By utilizing quantitative approaches, it purchases or sells specified financial assets at a predetermined period [16]. Consequently, technical analysis based on machine learning has become a reliable method for predicting the price of financial assets [17].

After establishing the stock selection model, this research compares two quantitative timing methods: The decision tree model and the LSTM model. After getting the outcomes of the two timing strategies, we will compare the yields of these models by the Sharpe ratio, yearly yield and Sortino ratio.

3.1. Decision tree model

Decision tree is widely applied in many areas, such as classification and recognition [18]. Decision trees operate on the principle of recursively and continuously generating decision trees. One of the disadvantages of this is that the trees generated based on the unknown tested data are not sufficiently accurate, i.e., they can suffer from over-fitting problems. In order to solve this problem, we need to simplify the decision tree. The algorithm of decision tree model is shown in Algorithm 1.

In the study of this problem, we use the ID3 algorithm. The core of the ID3 algorithm is to construct the decision tree recursively by selecting features at each node of the decision tree corresponding to the information gain criterion. The process of the ID3 algorithm is described as follows [19].

1) Starting from the root node, the information gain of all possible features is calculated for the node, and the feature with the largest information gain is selected as the feature of the node;

2) Create child nodes from different values of the feature, then call the above method recursively on the child nodes to build a decision tree until the information gain of all features is small or there are no features to choose;

3) Finally, a decision tree is obtained.

Quantitative indicators utilize historical data such as the price and trading volume of financial assets to represent the current market condition. The meanings of the indicator values are unique. Applying the properties of technical indicators improves the ability to forecast future price trends of financial assets. To construct an effective decision tree model, this paper reselects the following prediction indicators, including not only the indicators to predict the rise and fall but also the indicators to measure the price fluctuation range in order to operate profit stops and loss stops [20].

We choose 20 stocks and select 10 representative quantitative indicators (cf. Table 4), with 1 representing buying in, -1 representing selling out, and 0 representing waiting and holding, so as to evaluate each indicator of each stock. Finally, the decision action for each stock is determined based on the results of the 10 indicators. The decision process simply means that we buy when most of the indicators suggest we buy. The same goes for holding and selling. If there are a number of indicators with conflicting results, we tend to decide on a wait-and-see hold, but when we take these results as input to our decision tree algorithm, the output we get is a simple decision tree with only one level. We begin to reflect on the reasons for this phenomenon. We believe this may be due to the low input data, so we increase the total number of stocks from 10 to 20, and the number of indicators from 10 to 15. By repeating the previous calculation, we get a satisfactory decision tree result. Figure 3 shows an example of decision tree construction.

Finally, we complete the processing of timing signals of the 10 stocks in 2021 based on the decision tree model above. Starting from January 4, 2021, the first trading day of 2021, we make a judgment every 10 trading days and draw a time axis to calculate the timing signal. Figure 3 is the time axis we made. For reasons of space, we only show these[‖]. The timeline is set up so that the X axis is the date of the transaction we are judging, which is converted to 365 days. Points above the X axis mean buying in, points on the X axis are waiting and seeing, and points below the X axis mean selling. The results of top 20 stocks are shown in Figure 4.

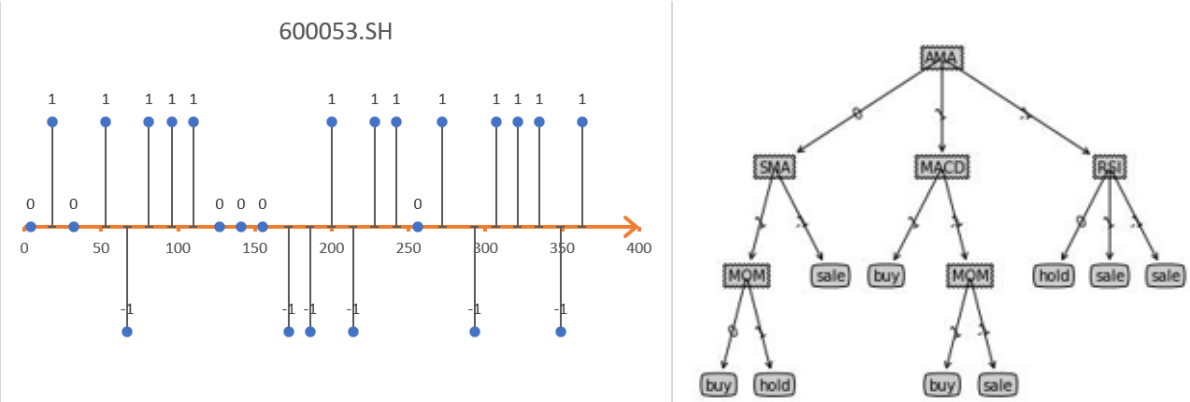


Figure 3. Investment signal of one example stock and the result of decision tree.

[‖]The others are in Appendix (i.e., Figure A4).

ID	Stock Name	SMA	EMA	MACD	MOM	WR	RSI	OBV	VR	MFI	ROC	TRIX	BOLL	ADX	AMA	DMAON	Result
1	Ningde era	-1	-1	1	0	-1	-1	-1	0	1	-1	-1	0	0	0	-1	-1
2	Motion controlling	-1	-1	-1	1	-1	-1	-1	1	0	-1	-1	1	-1	-1	-1	-1
3	Yunnan White Medicinal Powder	1	0	-1	0	1	1	-1	1	0	1	1	0	0	0	-1	0
4	China Merchants Bank	-1	-1	-1	1	-1	-1	-1	1	1	-1	-1	-1	-1	0	1	-1
5	Anhui Conch Cement	1	0	1	0	1	0	-1	0	0	1	-1	-1	-1	-1	1	0
6	Huayou cobalt industry	-1	-1	-1	-1	-1	1	-1	0	0	-1	-1	-1	0	-1	-1	-1
7	Muyuan Foods Co., Ltd.	1	0	-1	1	1	1	-1	0	0	-1	-1	0	-1	0	-1	0
8	Hikvision	-1	0	1	0	-1	-1	-1	0	-1	1	-1	-1	1	-1	1	-1
9	Shenzhen Expressway	-1	-1	1	0	1	-1	-1	0	-1	-1	-1	-1	0	-1	1	-1
10	Kweichow Moutai	1	-1	-1	1	1	-1	-1	0	-1	-1	-1	-1	0	-1	-1	-1
11	The north of rare earth	-1	-1	1	-1	1	1	-1	-1	0	1	1	0	0	1	1	1
12	Nanjing securities	1	1	1	1	-1	0	-1	0	-1	1	1	-1	-1	1	1	1
13	Valin seiko	1	1	1	1	0	0	1	0	0	-1	-1	0	0	1	-1	1
14	Tai long forever	1	1	-1	1	0	0	1	0	0	1	-1	0	0	1	-1	1
15	Concord electronics	1	1	1	1	-1	0	1	0	0	1	1	0	0	1	1	1
16	Silicon energy	1	1	1	1	0	0	1	0	-1	-1	1	-1	0	1	-1	1
17	Fu can science and technology	1	1	-1	-1	0	0	1	0	0	-1	-1	-1	-1	1	-1	-1
18	Zhen Jiang shares	1	1	1	-1	-1	-1	1	0	0	1	-1	0	0	1	-1	1
19	Christie technology	1	1	1	1	-1	-1	1	0	-1	1	-1	-1	1	1	1	1
20	Yingjie electric	1	1	1	-1	-1	0	-1	0	0	-1	1	-1	0	1	1	1

Figure 4. Data establishment of historical data set.

Algorithm 1 Decision Tree Model

Input: Indicator values (0, 1, -1). A trading judgment based on traditional timing.

Output: Timing trading judgments.

1: Calculates the empirical entropy of a given data set
   Storing the number of occurrences of each label
   Statistics for each set of feature vectors, and label counts.
2: Import factor scoring
   Create a test dataset and divide the dataset according to the given features.
3: Calculate the empirical entropy of the given dataset
   Remove the axis features and add the eligible ones to the returned dataset.
4: Count the elements with the most occurrences in the class list
   Arranging them in descending order according to the values of the dictionary.
5: Create a decision tree
   Extract classification labels, and iterate through all features to return the most frequent class labels;
   Select the best features and generate a tree based on the best features;
   Remove used feature labels and remove duplicate attribute values.
   Traversing the features and creating a decision tree.
6: Decision tree visualization
   Obtaining the number of decision tree leaf nodes, obtaining the number of decision tree layers, and drawing the decision tree.
7: **return** Generate decision tree to give transaction judgment

Table 4. Prediction indicators.

Factor	Factor Abbreviation	Factor description
Sample Moving Average	SMA	The SMA is the average price of the given time period, with each period's price given equal weight.
Exponential Moving Average	EMA	EMA is a price average that gives greater weight to recent prices.
The Moving Average Convergence Divergence	MACD	MACD is calculated using the differences between two moving averages of different lengths, a Fast moving average, and a Slow moving average. The change in MACD indicates the shift in market trends.
Momentum Index	MOM	MOM is an indicator that compares the current price to the price from a predetermined number of periods ago.
William Index	WR	WR is a technical indicator for analyzing the price range of fluctuating financial assets.
Relative Strength Index	RSI	RSI is a quantitative statistic used to analyze the price volatility range of financial assets.
On Balance Volume	OBV	OBV keeps a cumulative running average of the volume that occurs during up periods relative to down periods.
Volume Ratio	VR	It can analyze the volume-price relationship. Observing the trading volume might therefore provide insight into the financial market's fluctuations.
Money Flow Index	MFI	MFI uses both price and volume to measure buying and selling pressure.
Rate of Change	ROC	ROC indicator compares the current price with the previous price from a selected number of periods ago.

3.2. Long short-term memory

LSTM is a modified RNN whose model was originally proposed by Hochreiter and Schemidhuber in 1997, mainly to solve the RNN for long-series samples [21]. The model is a variant of the RNN model, which is mainly used to solve the problem of the lack of RNN's ability to learn long-term dependent information. Dynamic investments using the LSTM model can generate significant returns with relatively low risk [22]. We use the LSTM model to build a predictive model to predict the price of stocks so that we can subsequently analyze this data to generate our timing strategy. The detailed steps on how to build an LSTM model are shown in Algorithm 2.

With the data given and the adjustment of the model training data size (number of days without making decisions), the LSTM model is trained and we obtain the following training data, starting with a graph of the results of stock's price prediction. The LSTM closing price prediction results are plotted in Figure 5.

Finally, based on the LSTM price prediction results, we create the following timing signal strategy: If the next day's price is greater than 5 percent of the day's price, a buy signal is given; if the next day's price is less than 5 percent of the day's price, a sell signal is given. Figure 6 shows the results of 2 example stocks**.

4. Empirical analysis and comparison of results

4.1. Comparison of returns of different timing algorithms

The annualized yield and the Sharpe ratio are used to compare the advantages of the two timing techniques [13]. The Sharpe ratio is one of the most widely used methods for calculating risk-adjusted return:

**The others are in Appendix (i.e., Figure A5).

Algorithm 2 LSTM Model

Input: Historical closing price series.

Output: The price prediction result.

1: Calculate the given dataset, divide the dataset

Convert the data in DataFrame format to the format of a two-dimensional array;

Convert the data in the form of time series into the form of a supervised learning set;

Divide the dataset into a training set and a test set;

2: Model training

Create a Sequential model;

Stacking LSTM layers, stacking fully connected layers;

3: Model generalization

Separate the input and output columns of a dataset, and transform the input into the prediction function for single-step prediction;

After getting the predicted values, inverse scaling and inverse differencing are performed to reduce them to the original range of values;

Traversing the entire test set data.

4: Visualization of prediction results

5: **return** Build LSTM prediction model based on historical data

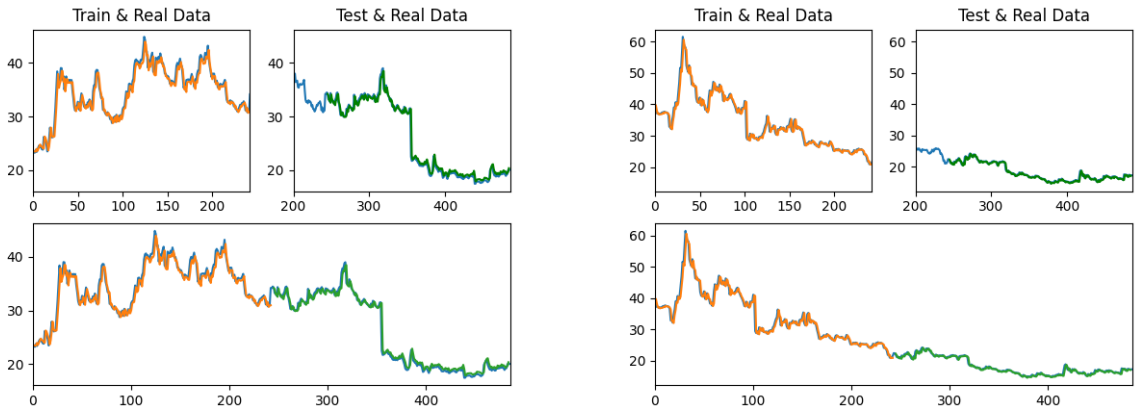


Figure 5. Plot of two example stocks.

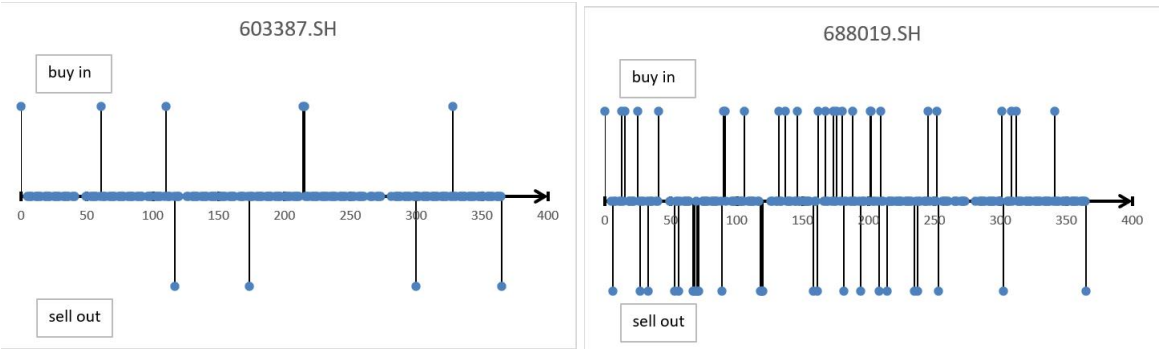


Figure 6. Plot of two example stocks.

$$\text{Sharpe ratio } = \frac{R_p - R_f}{\sigma_p}, \tag{4.1}$$

where R_p, R_f and σ_p mean the expected rate of return on investment portfolio, risk-free rate, and standard deviation of the portfolio, respectively.

Volatility is the degree of volatility in the price of a financial asset, a measure of uncertainty in asset returns, and is used to reflect the level of risk in a financial asset. The higher the volatility, the more violent the fluctuation of financial asset prices and the stronger the uncertainty of asset returns; the lower the volatility, the smoother the fluctuation of financial asset prices and the stronger the certainty of asset returns:

$$Volatility = \sqrt{\frac{250}{days}(x - avr)^2}, \tag{4.2}$$

where *avr* means the average returns of assets. Based on the two different timing signals mentioned above, we calculate their cumulative return, return volatility and Sharpe ratio, respectively. Also, to further evaluate the returns of different timing signals, we introduce, for example, the SSE index[††] and compare it with the overall market returns. Additionally, to better compare the results of the two signals and choose the best one to establish a model, we do a financial evaluation based on traditional signal. The algorithm is as follows:

1) Compute the short and long moving average (MA) of stock price;

2) Use the information of MA to trade the index;

3) Record and compute data of buy and sell, position, return, etc. with daily frequency for later analysis;

4) Do financial evaluation: Sharpe ratio, annual simple return; visualization and output of data.

The comparison results are shown in Table 5 and the cumulative return graph for different timing signals are in Appendix (i.e., Figure A3).

Table 5. Earnings comparison.

	Traditional Signal	Decision Tree Signal	LSTM Signal	Market Performance
Cumulative Return	239123	345458	483151	-9588
Earnings Volatility	0.11	0.13	0.12	0.04
Sharpe ratio	1.92	2.49	3.89	-0.96

From the graph, we can see that the market is in a relatively depressed situation in 2021. The decision tree and LSTM timing signals still get good returns in this market, so we think it is a good combination to introduce machine learning algorithms into quantitative timing.

4.2. Analysis and improvement of decision tree algorithm

The first issue we find with the timing part is that there is a certain correlation between the attributes we initially choose. We obtain poor training results and are unable to screen out effective indications when we utilize the ID3 decision tree method to screen the indicators.

[††]SSE: The sum of the squares of the errors of the corresponding points of the fitted and original data.

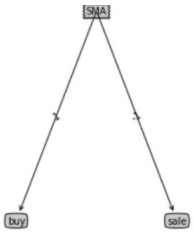


Figure 7. Initial training result of decision tree model.

As shown in Figure 7, a historical data collection of 10 indicators containing 10 stocks is imported, and the ID3 decision tree constructs a single-branch decision tree. We now have modified the historical dataset that is imported into the decision tree. Initially, the number of stocks is increased from 10 to 20 and the dataset is expanded, then we update the decision tree's input indications. We select the indicators with the lowest correlation possible and increased the number of indicators from 10 to 15. The following decision tree in Figure 8 is built by the ID3 decision tree once the new dataset has been imported and the information entropy between features and indicators has been identified. We also continue to use this result to determine the decision tree's timing signal [23].

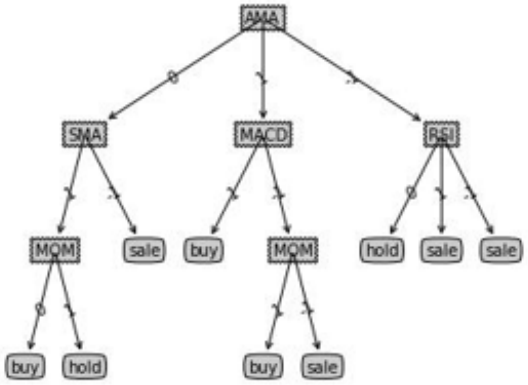


Figure 8. Decision tree training final results.

In addition, the returns of decision tree models for specific stocks are found to be less than those of traditional timing techniques when comparing the yields of timing strategies. To investigate the reasons for the timing process' unsatisfactory returns, we use the 603387.SH as an example (cf. Figure 9).

We find that in the signal processing of the decision tree, we simply process the buy and sell signals as 1 and - 1 and will lose the information about the rise and fall of the stock price in this process. Therefore, for different stock price fluctuations, the same number of shares are bought or sold. This treatment will also cause a decline in revenue, so we will assign a proportional value between -1 and 1 to the buy and sell signals according to the rise and fall of the stock, which may increase the income of the decision tree model.

4.3. LSTM algorithm analysis

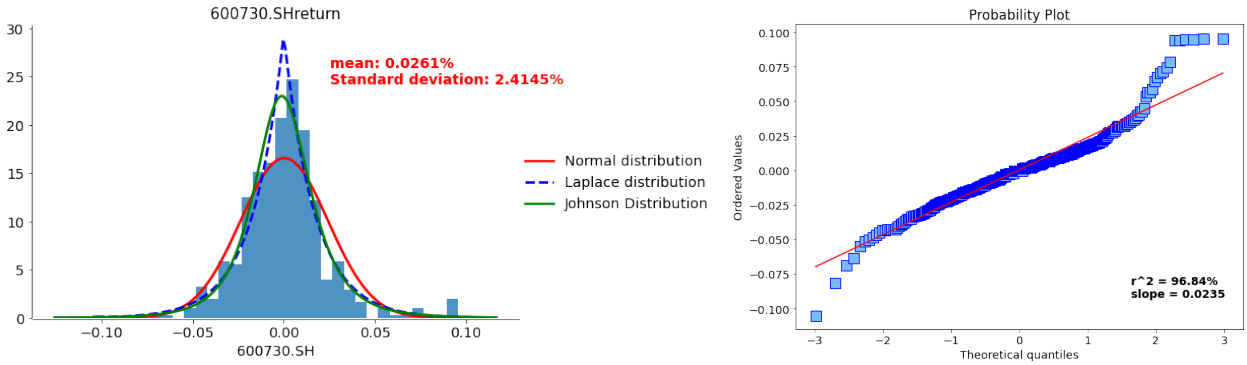
When we obtain the outcomes predicted by the LSTM model, we notice that the yield has reached 130%, which is considerably above the acceptable range. We then analyze this unusual outcome and attempt to apply the LSTM model to forecast the stock return. We use the stock price's logarithmic

Figure 9. 603387.SH price curve.

difference, but the prediction is unsuccessful due to the excessively high data volatility.

Before undertaking analysis and forecasting, we must ensure that the time series is stable [24]. As the closing price of the stock market is a non-stationary series, it is inappropriate to use it as the primary foundation for analysis. Instead, we examine the stationary of stock price data. As seen in Figure 10(a), the normal distribution performs a very poor job of fitting the dataset, whereas the Laplace and Johnson distributions do a good job of doing so. The statistical distribution demonstrates that the stock index return throughout the study period is not normally distributed. The quantile diagram (cf. Figure 10(b)) is then utilized to examine the quantile distribution of parameters and to examine the divergence from the normal distribution.

The yield of the 600730.SH exhibits a "fat-tail," as is seen from Figure 11. This indicates that, compared to what the normal distribution would imply, the frequency of extreme returns is significantly larger.



(a) The stationary of 600730.SH's return

(b) The quantile diagram of 600730.SH

Figure 10. 600730.SH.

Figure 10(a) demonstrates that the stock prices are non-stationary in terms of mean and variance. The possible actual price cannot be seen clearly because the forecast number is so close to the actual price. LSTM model appears to be effective at predicting the following value of the time series under consideration.

The time-series split method of the machine learning software Sklearn is utilized to study the distribution of samples while attempting to forecast the future distribution of returns utilizing statistical

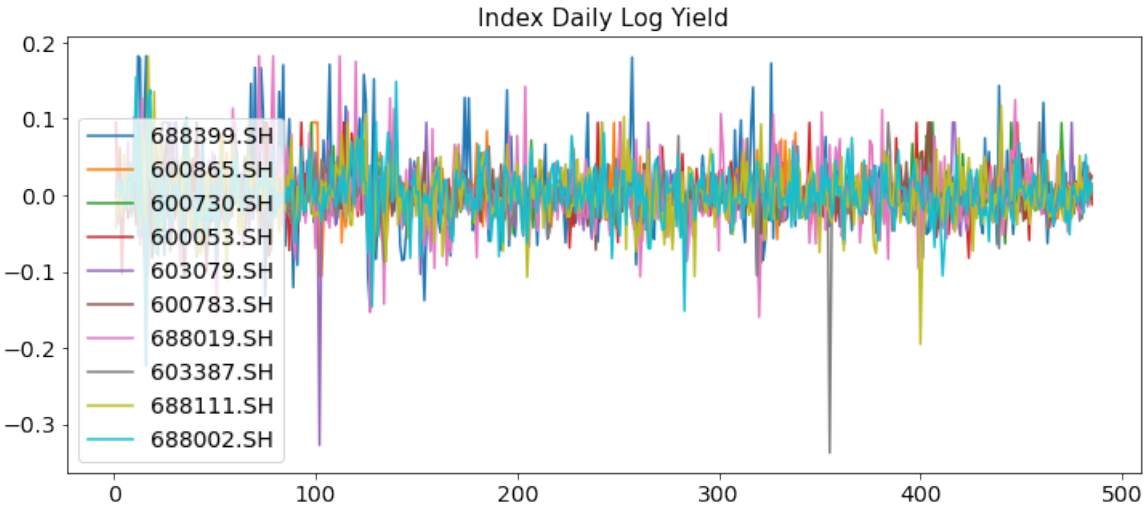


Figure 11. The index daily log yield.

markers (mean and variance) whose prior returns are similar to the normal distribution. This method gives the forward walk a forward version of cross-validation and predicts the next cycle in sequence using the prior data points, thus conserving time related information.

A static time series is a constant whose statistical parameters (such as mean, variance, autocorrelation, etc.) vary over time. Most statistical prediction methods are based on the assumption that time series can be mathematically transformed to be approximately stationary. As a result of this change, we no longer consider the index directly, but instead calculate the difference between subsequent time steps.

We then analyze this unusual outcome and attempt to apply the LSTM model to forecast the stock return. We use the stock price's logarithmic difference, but the prediction is unsuccessful due to the excessively high data volatility. If the data is discriminated once again, it is possible to produce a stable time series. However, such information has lost its economic significance and cannot be utilized to forecast the actual rate of return.

5. Conclusions

In this study, in order to better explain stock returns and fully validate the situation of the Chinese market, we considered 23 factors in terms of the value, growth, size, and transactions of the Shanghai stock market. Therefore, our contributions lie in factor validity tests and trial sorting methods. In terms of quantitative timing, we used the TDDPL method to obtain discrete and continuous technical index values [8]. We optimized the two index parameters of the MACD quantitative timing strategy commonly used in the stock market by using a neural network algorithm; that is, to establish an LSTM and MACD timing investment strategy. By empirical analysis, we found that both the decision tree model and the LSTM model get great results, which means it is a good combination to introduce machine learning algorithms into quantitative timing. It is worth emphasizing that we not only made a short-term prediction but also have many innovations in decision tree index screening. In conclusion, we have made innovations in both multi-factor models and automatic time selection models in order to construct a systematic stock trading strategy.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank Mr. Shengyuan Lu (individual researcher) and Miss Xinya Han (Nanjing University), who helped us worked out some problems during the difficult course of the paper. This study was partially supported by the Undergraduate Research and Learning Program of Southwestern University of Finance and Economics (No. YX220013).

Conflict of interest

The authors declare no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

1. A. McAfee, E. Brynjolfsson, T. H. Patil, D. Barton, Big data: the management revolution, *Harv. Bus. Rev.*, **90** (2012), 60–68.
2. E. A. Gerlein, M. McGinnity, A. Belatreche, S. Coleman, Evaluating machine learning classification for financial trading: An empirical approach, *Expert Syst. Appl.*, **54** (2016), 193–207. https://doi.org/10.1016/j.eswa.2016.01.018
3. S. M. Zhao, H. L. Yan, K. Zhang, Does fama-french five factor model outperform three factor model? Evidence from China's A-share market, *Nankai Econ. Stud.*, **32** (2016), 41–59. https://doi.org/10.14116/j.nkes.2016.02.003
4. J. J. Wang, Z. Z. Zhuang, L. Feng, Intelligent optimization based multi-factor deep learning stock selection model and quantitative trading strategy, *Mathematics*, **10** (2022), 566. https://doi.org/10.3390/math10040566
5. N. Nguyen, D. Nguyen, Global stock selection with hidden Markov model, *Risks*, **9** (2020), 9. https://doi.org/10.3390/risks9010009
6. A. Baykasoğlu, Í. Gölcük, Development of a novel multiple-attribute decision making model via fuzzy cognitive maps and hierarchical fuzzy TOPSIS, *Inf. Sci.*, **301** (2015), 75–98. https://doi.org/10.1016/j.ins.2014.12.048
7. X. Zhong, D. Enke, Forecasting daily stock market return using dimensionality reduction, *Expert. Syst. Appl.*, **67** (2017), 126–139. https://doi.org/10.1016/j.eswa.2016.09.027
8. F. W. Jiang, H. Xue, M. Zhou, Does big data improve multi-factor asset pricing models? Exploration of China's A-share market with machine learning, *Syst. Eng.-Theory Pract.*, **42** (2022), 2037–2048. https://doi.org/10.12011/SETP2021-2552
9. W. W. Jiang, Applications of deep learning in stock market prediction: recent progress, *Expert Syst. Appl.*, **184** (2021), 115537. https://doi.org/10.1016/j.eswa.2021.115537

10. G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, S. K. Bhat, Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications, *Int. J. Financial Stud.*, **11** (2023), 94. https://doi.org/10.3390/ijfs11030094
11. P. Tenti, Forecasting foreign exchange rates using recurrent neural networks, *Appl. Artif. Intell.*, **10** (1996), 567–582. https://doi.org/10.1080/088395196118434
12. F. E. Tay, L. Cao, Application of support vector machines in financial time series forecasting, *Omega*, **29** (2001), 309–317. https://doi.org/10.1016/S0305-0483(01)00026-3
13. Y. Deng, F. Bao, Y. Kong, Z. Ren, Q. Dai, Deep direct reinforcement learning for financial signal representation and trading, *IEEE Trans. Neural Netw. Learn. Syst.*, **28** (2016), 653–664. https://doi.org/10.1109/TNNLS.2016.2522401
14. J. Kamruzzaman, R. Sarker, Comparing ANN based models with ARIMA for prediction of forex rates, *Asor Bulletin*, **22** (2003), 2–11.
15. J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, *Expert Syst. Appl.*, **42** (2015), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040
16. G. J. Jiang, G. R. Zaynutdinova, H. Zhang, Stock-selection timing, *J. Bank. Finance*, **125** (2021), 106089. https://doi.org/10.1016/j.jbankfin.2021.106089
17. K. C. Rasekhschaffe, R. C. Jones, Machine learning for stock selection, *Financ. Anal. J.*, **75** (2019), 70–88. https://doi.org/10.1080/0015198X.2019.1596678
18. M. Li, H. Xu, Y. Deng, Evidential decision tree based on belief entropy, *Entropy*, **21** (2019), 897. https://doi.org/10.3390/e21090897
19. S. G. Deb, A. Banerjee, B. B. Chakrabarti, Market timing and stock selection ability of mutual funds in India: an empirical investigation, *Vikalpa*, **32** (2007), 39–52. https://doi.org/10.1177/0256090920070204
20. M. J. Zhang, H. C. Rao, J. X. Nan, G. D. Wang, Quantitative trading timing strategy based on decision tree, *Syst. Eng.*, **40** (2022), 118–130.
21. S. Hochreiter, J. Schmidhuber, LSTM can solve hard long time lag problems, in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, MA, (1996), 473–479.
22. H. Yao, S. Xia, H. Liu, Six-factor asset pricing and portfolio investment via deep learning: Evidence from Chinese stock market, *Pac. Basin. Finance J.*, **76** (2022), 101886. https://doi.org/10.1016/j.pacfin.2022.101886
23. A. Suáez, J. F. Lutsko, Globally optimal fuzzy decision trees for classification and regression, *IEEE Trans. Pattern Anal. Mach. Intell.*, **21** (1999), 1297–1311. https://doi.org/10.1109/34.817409
24. C. Ma, G. Dai, J. Zhou, Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method, *IEEE Trans. Intell. Transp. Syst.*, **23** (2021), 5615–5624. https://doi.org/10.1109/tits.2021.3055258

Appendix

	1	2	3
B/M	0.439341	[26.1856453574, 19.9930369335]	[0.738095238095, 0.392857142857]
EPS	0.711845	[27.7733159328, 11.8529826427]	[0.678571428571, 0.404761904762]
FAP	-0.101805	[16.8813416468, 10.8231132872]	[0.714285714286, 0.428571428571]
GP/R	0.33395	[22.4171421831, 21.5378319764]	[0.630952380952, 0.297619047619]
L/A	-0.868283	[28.5016128577, 10.7668241109]	[0.702380952381, 0.428571428571]
P/R	0.381557	[21.7347534606, 21.6563578468]	[0.654761904762, 0.297619047619]
PB	0.490657	[16.2011771306, 10.3105732135]	[0.607142857143, 0.404761904762]
PE	-0.842341	[33.6082285064, 7.04649162945]	[0.690476190476, 0.452380952381]
PS	0.857868	[33.2924976345, 7.16144670626]	[0.678571428571, 0.452380952381]
ROA	0.53566	[37.760142035, 21.5616981796]	[0.785714285714, 0.357142857143]
ROE	0.516171	[29.9136141357, 17.7402268006]	[0.75, 0.392857142857]
capitalization	-0.515472	[27.4456647809, 12.3236581934]	[0.654761904762, 0.392857142857]
circulating_cap	-0.0133969	[17.6989981122, 13.3780223251]	[0.619047619048, 0.357142857143]
circulating_market_cap	-0.603429	[24.1174049049, 19.379870653]	[0.666666666667, 0.25]
gross_profit_margin	0.859684	[22.7941651935, 14.1611150282]	[0.714285714286, 0.416666666667]
inc_net_profit_annual	0.404461	[20.3641496408, 20.1183915697]	[0.678571428571, 0.25]
inc_net_profit_year_on_year	0.858423	[24.7104431148, 12.4408302076]	[0.690476190476, 0.380952380952]
inc_operation_profit_annual	0.658584	[21.3398937296, 18.1188599747]	[0.833333333333, 0.345238095238]
inc_operation_profit_year_on_year	0.876961	[25.4550160321, 10.0889937803]	[0.702380952381, 0.392857142857]
market_cap	-0.653921	[35.6014814842, 18.9710502035]	[0.678571428571, 0.261904761905]
net_profit	0.25151	[20.2644520059, 19.4442113667]	[0.630952380952, 0.297619047619]
operating_revenue	0.491532	[18.5889978424, 17.6976512068]	[0.690476190476, 0.392857142857]
turnover_ratio	0.292739	[25.2666116387, 20.3568742797]	[0.630952380952, 0.261904761905]

Figure A1. Validity test.

	port1	port2	port3	port4	port5	benchmark
EPS	0.206645	0.258821	0.154189	0.351434	0.365935	0.088211
L/A	0.373226	0.418582	0.251828	0.275774	0.195902	0.088211
PE	0.424276	0.232296	0.278372	0.240802	0.158614	0.088211
PS	0.159776	0.236854	0.260613	0.236294	0.421129	0.088211
gross_profit_margin	0.229863	0.235950	0.227385	0.313811	0.316158	0.088211
inc_net_profit_year_on_year	0.212544	0.193793	0.267625	0.381285	0.336264	0.088211
inc_operation_profit_year_on_year	0.188977	0.173078	0.303381	0.383502	0.342753	0.088211

(a) Effective Factor Annualized Return

	port1	port2	port3	port4	port5	benchmark
EPS	2.086571	2.979088	1.364071	5.092131	5.495012	0.660651
L/A	5.705828	7.149406	2.848285	3.311631	1.925316	0.660651
PE	7.347648	2.501797	3.364585	2.649353	1.418987	0.660651
PS	1.433579	2.580221	3.013204	2.570521	7.237589	0.660651
gross_profit_margin	2.460507	2.564557	2.418881	4.142766	4.198139	0.660651
inc_net_profit_year_on_year	2.178229	1.894504	3.149016	5.945432	4.693167	0.660651
inc_operation_profit_year_on_year	1.825138	1.605916	3.902615	6.012585	4.861056	0.660651

(b) Effective Factor Seven Year Return

Figure A2. Effective factor return.

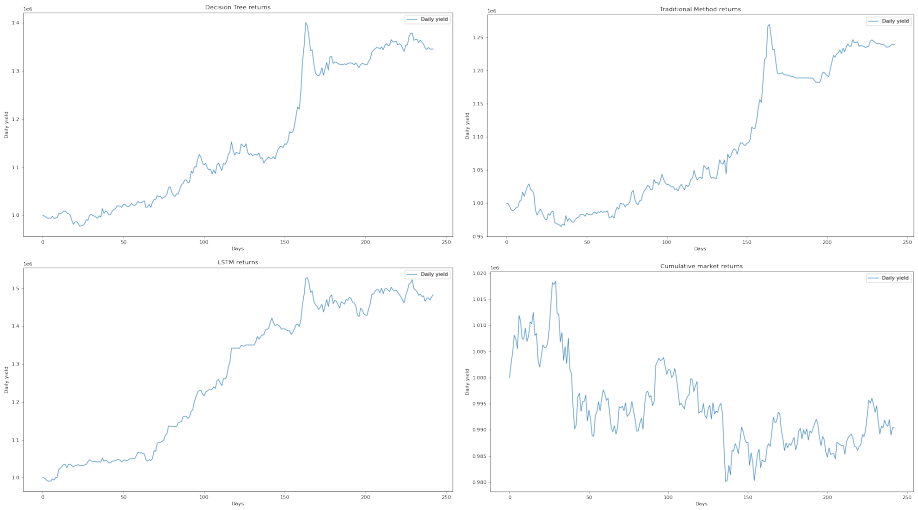


Figure A3. Cumulative return chart for different timing signals.

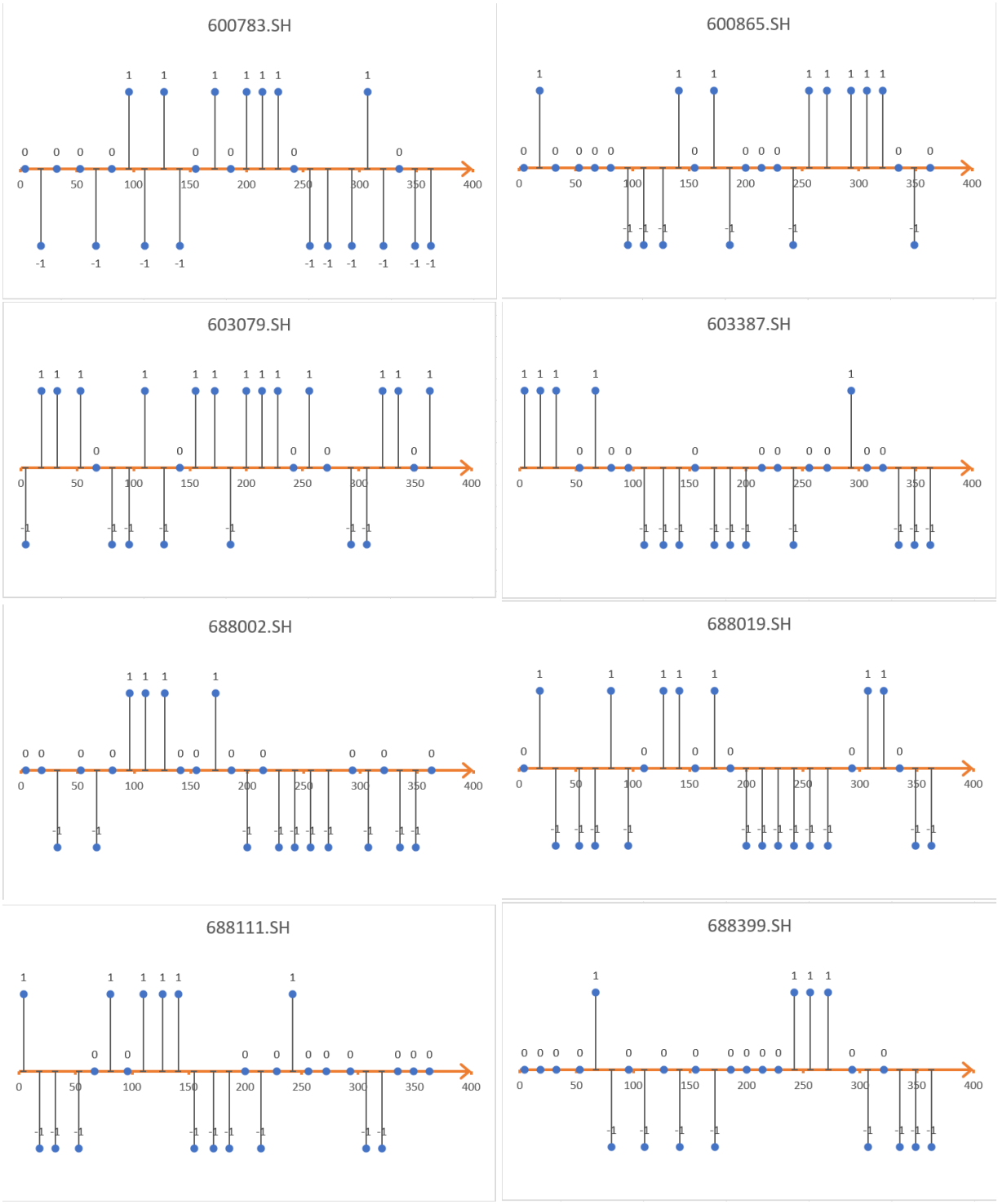


Figure A4. Time signal diagrams of decision tree model.

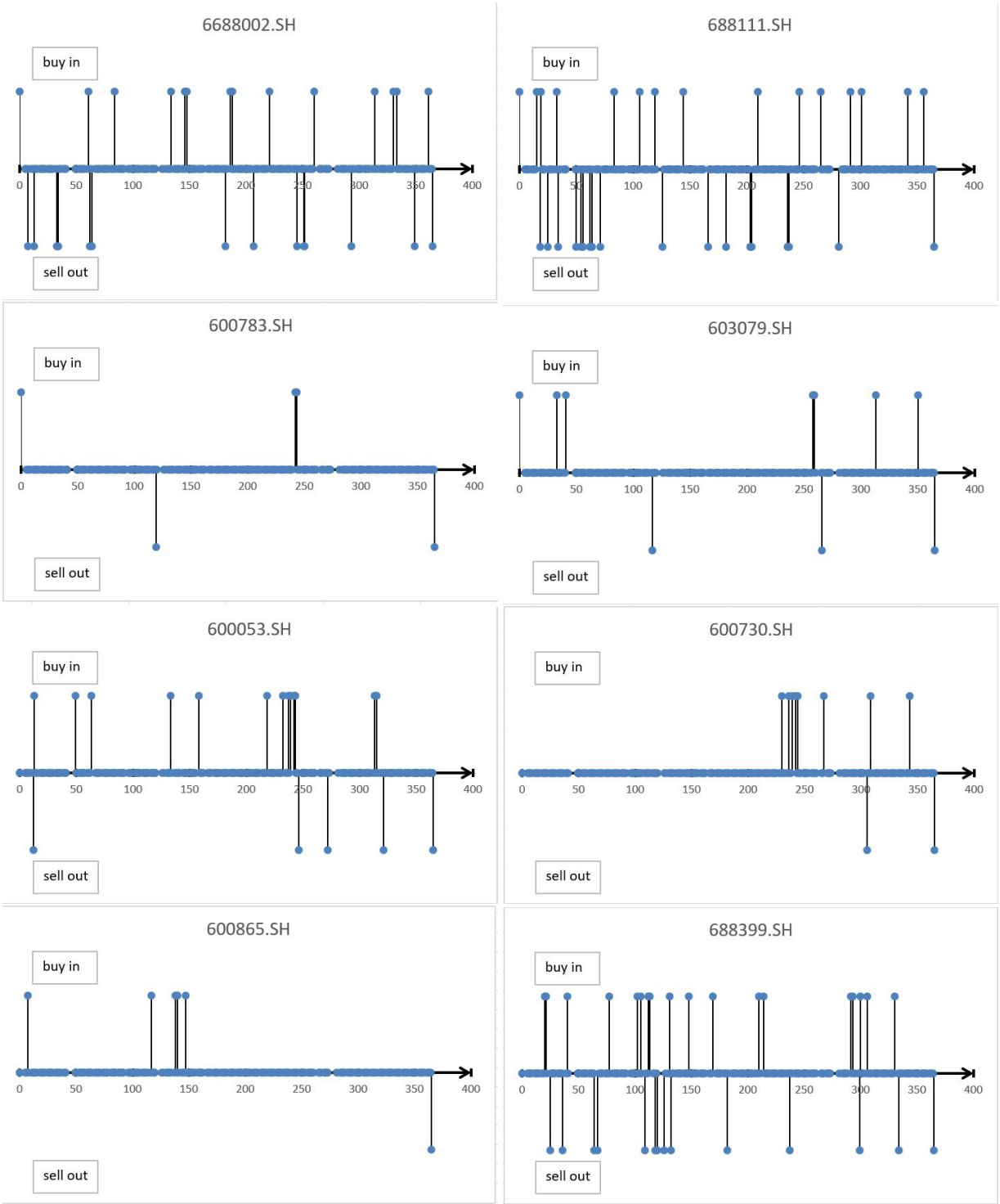


Figure A5. Time signal diagrams of LSTM model.