Comparison of Stock Price Prediction Based on Different Machine Learning Approaches

Qianqiao Hu[1], Songshan Qin[2], and Shuai Zhang[3](✉)

[1] Department of Mathematics and Statistics, Zhongnan University of Economics and Law, Wuhan, Hubei, China
[2] Adam Smith Business School, University of Glasgow, Glasgow, UK
2538008q@student.gla.ac.uk
[3] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China
zhangshauino.1@sjtu.edu.cn

Abstract. The advantages of machine learning model for fuzzy nonlinear data modeling enable it to be well applied to predict complex nonlinear stock price of low signal-to-noise ratio. Most of the research on stock price prediction with machine learning focuses on the effect evaluation or improvement of a single algorithm, while the comparative research on algorithms has little attention. For investors, the first confusion before predicting stock price trend is to choose the appropriate model instead of optimizing. In this paper, we compare the up-or-down classification performance on a series of prediction windows of LightGBM, Random Forest and Logistic Regression on three stocks to verify the consistency of results. Some technical indicators, e.g., Relative Strength Index (RSI), Simple Moving Averages (SMA) etc. are selected as factors to train our models. Encouragingly, the comparison results show that the prediction performance of the models are significantly different in short and long time window. This finding has some guiding significance for the improvement of long-term and short-term forecasting performance. In addition, some useful suggestions based on the conclusion can be put forward to instruct investors to make better quantitative investment.

Keywords: Logistic Regression · Random Forest · LightGBM · Model Comparison · Stock Price Forecast

1 Introduction

Machine learning has its advantages in the low signal-to-noise ratios and complex nonlinear stock markets. As for the field of quantitative investment, the investigation of the future price movement trend of stocks prediction based on machine learning algorithms has also gained more and more extensive attention. However, most of the machine learning stock price prediction approaches focus on the evaluation or improvement of the

Q. Hu, S. Qin and S. Zhang—Contributed equally.

© The Author(s) 2023
D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 215–231, 2023.
https://doi.org/10.2991/978-94-6463-030-5_24

effectiveness of individual algorithms, while less attention has been paid to the comparative study of algorithms. Meanwhile, current research has limited practical significance due to insufficient consideration of technical factors in algorithms. To fill the gap in this area, this paper will use multiple machine learning algorithms for up and down trend prediction for selected stocks. This has important practical implications for guiding investors in choosing algorithms, weighing risk-return, and making more rational investments.

Many factors affect stock prices, among which fundamental indicators are important factors that influence stock prices [7, 8, 12, 15, 21]. In addition, technical indicators such as momentum technical indicators have a more flexible impact on stock prices [4]. To be more consistent with the real stock market situation and the large amount of non-linear data in the market, scholars tend to combine fundamental factors with technical factors [9, 18]. There are many technical indicators for stock price research, and each type of research has different methods of indicator selection, choosing between fundamental factors and technical indicators.

For stock price analysis of a single industry, random forest models can effectively predict stock prices, which have double randomness and can overcome subjective empirical judgments and emotional factors interference [2, 3, 5, 10]. Logistic regression is also effective when it is applied to the Shanghai and Shenzhen markets, which can successfully perform its predictive function on the probability of stock price increases [16]. Besides, support vector machine regression forecasting models can reflect stock price changes more comprehensively, being better suited for the non-linear time-varying pattern of stock prices [13, 19]. Moreover, as time-series data, many scholars have used ARIMA models to forecast stock prices, having better short-term forecasting effects [1, 11, 17] Contemporarily, the LightGBM algorithm developed by one of the Microsoft groups has also started to be applied to the financial field, and studies have proved that the stock price prediction model based on LightGBM has better prediction ability and higher returns [14, 20].

In this paper, we will select the raw data of three different stocks, calculate their technical indicators and screen them by correlation analysis and significance test. Forecasting is performed based on random forest, logistic regression, and lightGBM models. With the rolling prediction, the consistency and validity of comparing the three models are explored in four aspects (accuracy, recall, precision, and f1 value, respectively).

The rest part of this paper is organized as follows. Section 2 will describe the data and meanings of indicators chosen for this paper and introduce the three machine learning models. Section 3 will give the specific results of the three models and compare them. Section 4 will analyze and discuss the results. Finally, Sect. 5 will discuss the practical implications of this study.

2 Data and Methods

2.1 Stock Data

The time series stock data which comes from Wind is the high, low, trading volume, open and close price of minute frequency for 2015–2019 in Shanghai Stock Exchange.

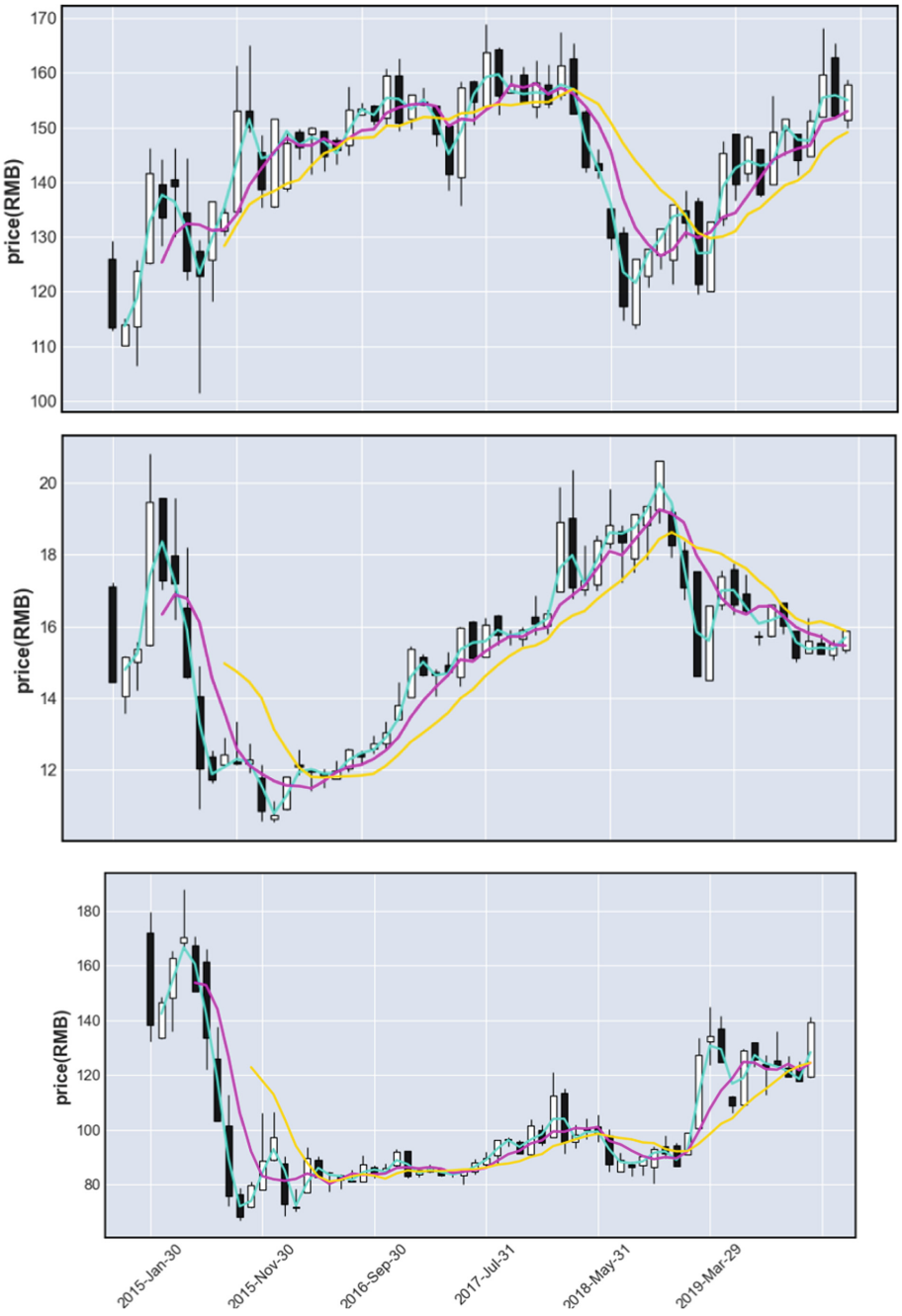

Fig. 1. The candlestick chart of SH600000, SH600028 and SH600030 (from the upper to the lower)

Their ticker symbols are SH600000, SH600028 and SH600030. The stock data we get has been weighted and the candlestick charts of them are illustrated in Fig. 1.

2.2 Data Preprocessing

The time series stock data is firstly Exponentially Smoothed, which strengthens the effect of the recent observation value on the predicted value while preserves the memory of the history, i.e., the predicted value can quickly reflect the actual change of the market. The exponential smoothing statistic of a series Y can be calculated recursively as:

$$S_t = \begin{cases} Y_0 \ t = 0 \\ a * Y_t + (1 - a) * S_{t-1} \ t > 0 \end{cases} \tag{1}$$

where a is the smoothing constant which is between 0 and 1. Different a values can be selected to adjust the uniformity of time series observation values. Exponential smoothing method can remove randomness and noise in time series data and make data easier to predict. In this paper, the a is set as 0.6.

Based on the exponentially smoothed data, the label in different time windows to be predicted are calculated as followed:

$$label_t = Sign(close_{t+window} - close_t) \tag{2}$$

Here, window is the prediction window. In this paper, the window is set as [1, 3, 5, 10, 20, 30].

2.3 Feature Extraction

The Technical Indicators are calculated from the data of three stocks with Python library Stockstats. The technical indicators used are introduced as follows:

2.3.1 Simple Moving Averages (SMA)

The formula to calculate SMA is:

$$SMA_n = \frac{(C_1+C_2+C_3+...+C_n)}{n} \tag{3}$$

where C_i is Closing price on day i, n is Moving average period.

The theory of moving averages is one of the most widely used technical indicators today. It helps traders identify existing trends, identify upcoming trends, and spot overextended trends that are about to reverse. In this paper, the n is set as 5.

2.3.2 Moving Average Convergence/Divergence (MACD)

The formula to calculate MACD is:

$$MACD = EMA_{12}(C) - EMA_{26}(C) \tag{4}$$

$$SignalLine = EMA_9(MACD) \tag{5}$$

where C is Closing Price series and EMA_n is n day Exponential Moving Average.

MACD is a technical indicator that uses the convergence and separation between the short-term (usually 12-day) and long-term (usually 26-day) closing price exponential moving averages to make buy and sell decisions.

2.3.3 Stochastic Indicator

The formula to calculate Stochastic Indicator is:

$$RSV_n = 100 \times \frac{(C_n - L_n)}{(H_n - L_n)} \tag{6}$$

$$K_0 = 50 \tag{7}$$

$$D_0 = 50 \tag{8}$$

$$K_n = \frac{2}{3} \times K_{n-1} + \frac{1}{3} \times RSV_n \tag{9}$$

$$D_n = \frac{2}{3} \times D_{n-1} + \frac{1}{3} \times K_n \tag{10}$$

$$J_n = 3 \times D_n - 2 \times K_n \tag{11}$$

where RSV_n is day n Raw Stochastic Value, C_n is day n Closing Price, L_n is Lowest Low over past n days, H_n is Highest High over past n days, K_n is day n Stochastic Indicator K, D_n is day n Stochastic Indicator D and J_n is day n Stochastic Indicator J.

The Stochastic Indicator on the chart is three curves, namely line K, line D and line J. The relationship between these three curves can be used to study the trend of stock prices. Stochastic index is mainly used to reflect the phenomenon of overbought and oversold in the stock market, the phenomenon of trend backtracking and the cross breakthrough of K line and D line. In this paper, the n is set as 14.

2.3.4 Relative Strength Index (RSI)

The formula to calculate RSI is:

$$RSI = 100 - \frac{100}{1 + RS} \tag{12}$$

$$RS = \frac{Average\ Gain\ ver\ past\ 6\ day}{Average\ Loss\ ver\ past\ 6\ day} \tag{13}$$

RSI is first used in futures trading. Subsequently, investors find that it is also very effective to guide stock market investment, i.e., the characteristics of this index is constantly summarized. Contemporarily, RSI has become one of the most widely used technical indicators by investors. It is based on the principle of supply and demand balance, by measuring the percentage of the total rise of stock prices in the average of the total change of stock prices in a certain period, to evaluate the strength of long and short forces, and then suggest specific investment operations.

2.3.5 BIAS

The formula to calculate BIAS is:

$$BIAS(n) = \frac{C - MA_n(C)}{MA_n(C)} \times 100 \tag{14}$$

where C is Current Closing Price, $MA_n(C)$ is Moving Average of n day Closing price.

BIAS is a technical indicator which reflects the deviation between closing price and MA by calculating the percentage difference in a certain period. It indicates the possibility of price retracting or rebounding due to deviating from the moving average trend in severe volatility and the credibility of price moving within the normal range of volatility to continue the original trend. In this paper, we set n as 5.

2.3.6 On Balance Volume (OBV)

The formula to calculate OBV is:

$$OBV(t) = \begin{cases} OBV(t-1) + Vol(t) & if \ \ C(t) > C(t-1) \\ OBV(t-1) - Vol(t) & if \ \ C(t) < C(t-1) \\ OBV(t-1) & if \ \ C(t) = C(t-1) \end{cases} \tag{15}$$

where $OBV(t)$ is On Balance Volume at time t, $Vol(t)$ is Trading Volume at time t and $C(t)$ is Closing Price at time t.

OBV is a technical index that takes volume as a breakthrough to find hot stocks and analyze stock price movement trend.

2.3.7 Price Momentum Index (CR) and CR-MA

The formula to calculate Price Momentum Index is:

$$CR = \frac{Bull\ Strength}{Bear\ Strength} \times 100 \tag{16}$$

$$Bull\ Strength\ =\ the\ sum\ of\ Gain\ over\ 26\ days \tag{17}$$

$$Bear\ Strength\ =\ the\ sum\ of\ Loss\ over\ 26\ days \tag{18}$$

$$Gain(t) = High(t) - Low(t-1) \tag{19}$$

$$Loss = SP(t-1) - Low(t) \tag{20}$$

$$SP(t) = \frac{High(t) + Low(t)}{2} \tag{21}$$

where $SP(t)$ is Standard Price at day t, $High(t)$ is the Highest Price at day t and $Low(t)$ is the Lowest Price at day t.

Price momentum index (CR) can roughly reflect the pressure band and support band of stock prices, making up for the inadequacy of AR and BR.CR-MA is the moving average of CR which can be calculated as followed:

$$CR - MA(n) = \frac{CR_1 + CR_2 + ... + CR_n}{n} \tag{22}$$

where n is the moving average period.

2.3.8 Volatility Ratio (VR)

$$VR = \frac{AVS + \frac{1}{2}CVS}{BVS + \frac{1}{2}CVS} \tag{23}$$

$$AVS = \sum_{n=1}^{p} AV_n, BVS = \sum_{n=1}^{q} BV_n, CVS = \sum_{n=1}^{r} CV_n \tag{24}$$

where $p + q + r = N$, AV_n is the volume when price rises during N days, BV_n is the volume when price falls during N days and CV_n is the volume when price stays during N days.

Volatility Ratio (VR) is a strength index which mainly measures the trend of stock price from the perspective of volume.

2.3.9 Directional Movement Index (DMI)

DMI includes four indexes. They are positive direction index (+DI), negative direction index (−DI), Average Directional Indicator (ADX) and ADXR in which in this paper we use +DI, −DI and DX which is calculated from +DI and −DI. The formulas to calculate them are:

$$+DI = \frac{+DM}{TR} * 100 \tag{25}$$

$$-DI = \frac{-DM}{TR} * 100 \tag{26}$$

$$DX = \frac{(+DI) - (-DI)}{(+DI) + (-DI)} * 100 \tag{27}$$

$$+DM = \begin{cases} High_n - High_{n-1} & if\ High_n > High_{n-1} \\ 0 & if\ High_n < High_{n-1} \end{cases} \tag{28}$$

$$-DM = \begin{cases} Low_{n-1} - Low_n & if\ Low_{n-1} > Low_n \\ 0 & if\ Low_{n-1} < Low_n \end{cases} \tag{29}$$

$$TR = \max\begin{pmatrix} |High_n - Low_n|, |High_n - Close_{n-1}|, \\ |Low_n - Close_{n-1}| \end{pmatrix} \tag{30}$$

The DMI indicator can be used as a buy or sell signal and can also tell if a move has started. However, it must be noted that when the market's upward (downward) trend is very clear, the use of this indicator for buying and selling guidance is better, otherwise it can be distorted.

2.4 Model

There are many types of classification models for machine learning. In this paper, we select three well-known models: Logistic Regression, Random Forest and LightGBM.

2.4.1 Logistic Regression

Logistic regression is a linear classifier. Although its name contains the word regression, it is a classification algorithm. Logistic regression can be used to solve binary classification and multiple classification problems, and the response results of both problems can be expressed as conditional probability $P(Y|X)$.

However, Logistic regression is more commonly used in solving binary classification problems. For binary classification problems, the label Y is usually set as 0 and 1, hence, the respond results can be expressed as followed:

$$\begin{aligned} \pi &= P\big(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\big) \\ &= \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} \end{aligned} \tag{31}$$

$$\begin{aligned} 1-\pi &= P\big(Y = 0|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\big) \\ &= \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} \end{aligned} \tag{32}$$

where π is the probability of something happening, $x_1, x_2, \ldots, x_p$ are the value of each predictive variables, and $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters needed to be estimated. The right side of the Eq. (32) is called Logistic Function.

The expression $\frac{\pi}{1-\pi}$ is the odds of the event, i.e.:

$$odds = \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p} \tag{33}$$

Taking the logarithm of both sides, we derive Logit Function:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{34}$$

One knows that the Logit Function is a linear function of X. The parameter β of Logistic regression model can be obtained by maximum likelihood estimation. Supposing that $P(Y = 1|X = x) = \pi(x), P(Y = 0|X = x) = 1-\pi(x)$, the likelihood function is:

$$L(\beta|x, y) = \prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{35}$$

The log- likelihood function is:

$$lnL(\beta|x, y) = \sum_{i=1}^{N}[y_i(\beta x_i) - ln\big(1 + e^{\beta x_i}\big)] \tag{36}$$

At this point, the Logistic regression problem becomes the optimization problem of maximizing $L(\beta|x, y)$, which can be solved by using gradient descent method.

At present, Logistic regression has been widely used in data mining, automatic diagnosis of diseases, economic forecasting and other fields. It is favored by quantitative investors because of its simple model and strong interpretability.

2.4.2 Random Forest

Random Forest is a decision tree model based on Bagging framework, which can construct multiple decision trees. Decision tree is a classification method of tree structure. The specific implementation method is a recursive classification method from the feature root node to the leaf node of the tree. Each node in the tree represents an object, each bifurcated path represents a possible attribute value, and each leaf represents a value for an object along the path from the root to the leaf. For decision tree model, the three common algorithms are ID3, C4.5 and CART. In practice, Decision trees are easy to understand and implement. The preparation of data is often simple or unnecessary, and the ability to handle both data and conventional attributes can produce viable and effective results for large data sources in a relatively short period of time. Generally, decision trees often produce overfitting problems, but Random Forest can inhibit it at some extent.

Based on decision tree model, Random Forest constructs a set of decision trees to become a forest. When we need to predict a certain data, all the different trees in the random forest will spontaneously generate a prediction result. Then, one selects all the results by voting to produce the final prediction result. This enables the random forest model to have a good performance and a fast running speed, which is not easy to overfit.

2.4.3 LightGBM

LightGBM is a project of Distributed Machine Learning Toolkit (DMKT) owned by Microsoft [6]. It is also a decision tree algorithm based on Boosting framework, which is similar in principle to XGBoost and GBDT mainly used to solve the problems encountered by GDBT in mass data. It can be considered as a lightweight version of XGboost. In order to speed up the training speed of GBDT model without influencing the accuracy, it makes the following improvements on the traditional GBDT model:

- Decision tree algorithm based on Histogram
- Gradient-based One-Side Sampling (GOSS)
- Exclusive Feature Bundling (EFB)
- Leaf growth strategies with depth constraints
- Categorical Feature
- Support parallelism
- Cache-hit rate optimization

These developments enable LightGBM to use less cache but speed up noticeably while perform well.

2.5 Method

2.5.1 Correlation Analysis

In order to avoid the problem of collinearity between independent variables and observe the relevance between independent variables and dependent variables, correlation analysis and significance test of technical indicators is carried out.

2.5.2 Model Training

For each model, the input of the model is the 12 technical indicators, while the output is the predicted labels which stands for whether the stock price is going up or down in different prediction windows. When model tuning, grid search method is used to obtain the optimal model. The same processes are applied to all three models.

2.5.3 Model Evaluation

In order to avoid the model in the long-term forecast failure caused by the stock market changing, we adopted the rolling test. In other words, we use the data of a year to train models while the next year to test, thus the training and testing process are carried out for four times. For every time, the accuracy, precision, recall and f1 score are calculated and stored. The introduction of these evaluation indexes of model performance are as followed:

- Accuracy
  Accuracy is the ratio of the number of samples correctly classified by the classifier to the total number of samples for the current data set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{37}$$

- Precision
  Precision is the ratio of the total number of positive samples correctly classified to the total number of positive samples discriminated by the classifier for the current data set.

$$Precision = \frac{TP}{TP+FP} \tag{38}$$

- Recall
  Recall is the ratio of the total number of positive samples correctly classified by the classifier to the total number of real positive samples for the current data set.

$$Recall = \frac{TP}{TP+FN} \tag{39}$$

- F1 score
  F1 score is the weighted average of Precision and Recall.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{40}$$

  The average of each kind of these indexes in rolling test are regarded as the final indicators to evaluation the performance of each model. Finally, we paint figures between window and indicators to observe all three models' performance in different prediction windows.

3 Results

This section can be encompassed with two components. Primarily, the tables will show the results of indicators selection. Subsequently, the following results of stock prediction models will be illustrated by a series of figures.

3.1 Descriptive Analysis of Indicator Selection Results

This paper refers to the traditional exchange rate determination theory and the mainstream variables of influential factors in previous literature. Twenty-eight common technical indicators are initially selected to exam the correlations in this study. The specific indicators are shown in the Table 1.

Table 1. INDICATORS DESCRIPTION

Indicators Abbreviation	Name of Indicators
close_5_sma	Simple Moving Average on 5 Days Closing Price
macd	Moving Average Convergence Divergence
kdjd	Stochastic Indicator - D
rsi_6	Relative Strength Index
BIAS_5	Bias Ratio on 5 days
OBV	On Balance Volume
cr	Price Momentum Index
cr-ma3	Price Momentum Index in Moving Average of 20 Days
pdi	Positive Directional Movement Index
mdi	Negative Directional Movement Index
dx	Movement Index
vr	Volatility Ratio

Table 2. CORRELATIONS OF INDEPENDENT AND DEPENDENT VARIABLES

Indicators	close_5_sma	macd	kdjd	rsi_6	BIAS_5	OBV	cr	cr-ma3	pdi	mdi	dx	vr
label1	0.003	0.081	0.159	0.635	0.565	0.005	0.014	−0.011	0.214	−0.208	0.020	0.021
label3	−0.002	0.035	0.038	0.335	0.352	0.002	0.011	−0.006	0.154	−0.144	0.019	0.014
label5	−0.004	0.021	0.015	0.243	0.271	0.003	0.007	−0.008	0.122	−0.111	0.017	0.009
label10	−0.010	0.011	0.007	0.168	0.194	0.000	0.004	−0.013	0.089	−0.078	0.013	0.005
label20	−0.019	0.002	0.009	0.124	0.145	0.003	0.003	−0.019	0.066	−0.061	0.011	0.004
label30	−0.026	−0.003	0.012	0.105	0.120	0.003	0.004	−0.025	0.054	−0.054	0.009	0.003
Average Correlation	−0.010	0.025	0.040	0.268	0.275	0.003	0.007	−0.014	0.116	−0.109	0.015	0.009

As summarized in Table 2, after the selection from twenty-eight different technical analysis indicators through the Pearson correlation coefficient, this study remains twelve indicators that have relatively weak correlations among each other. Further, this study found the insignificance by t-test within the twelve indicators. Therefore, it might be less likely to exist linear relationships. Additionally, the significant correlation coefficient between independent and dependent variables can prove that the expected forecasting results are reasonable and adequate. Generally, the twelve technical indicators can be evidenced as valid, practical, and stable to predict A-share stock price.

3.2 Descriptive Analysis of Prediction Model Results

In this paper, the twelve technical indicators are applied as the classification features to train the three types of machine learning models. Combining with historical data from the three fluctuated shares, these predictive models can make forecasts on the next stock price rise and fall trends. In the meantime, the predictive accuracy of these models can be evaluated in the time windows from one to thirty minutes. The model accuracy of the training and testing sets are shown in the Fig. 2.

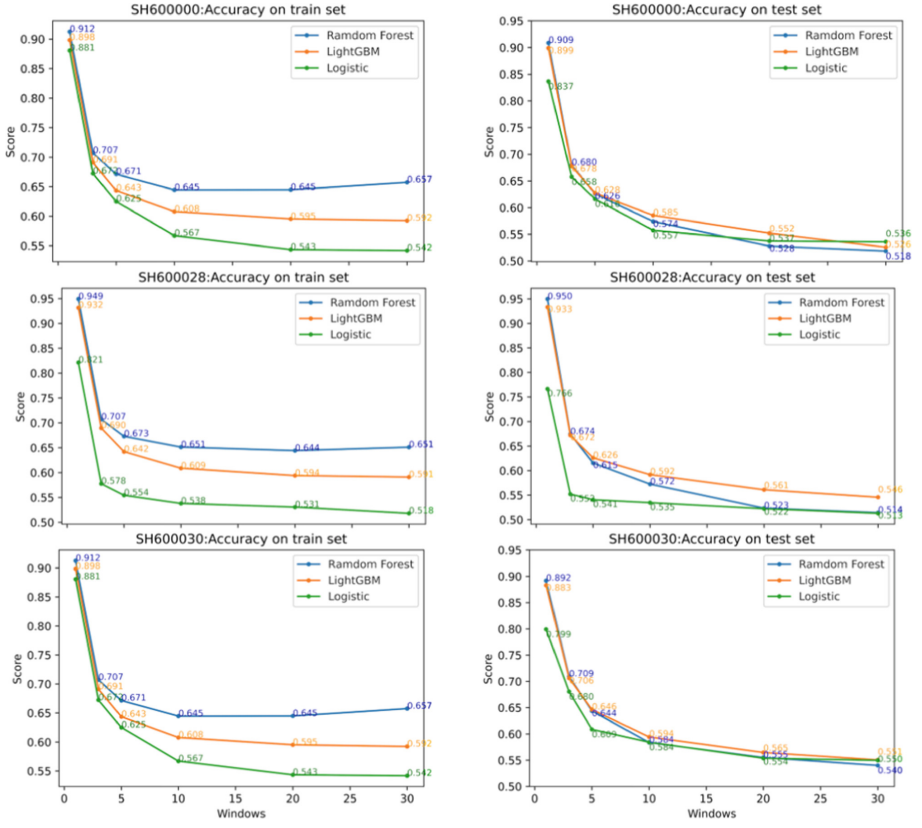

Fig. 2. Models Accuracy on Training and Testing Sets

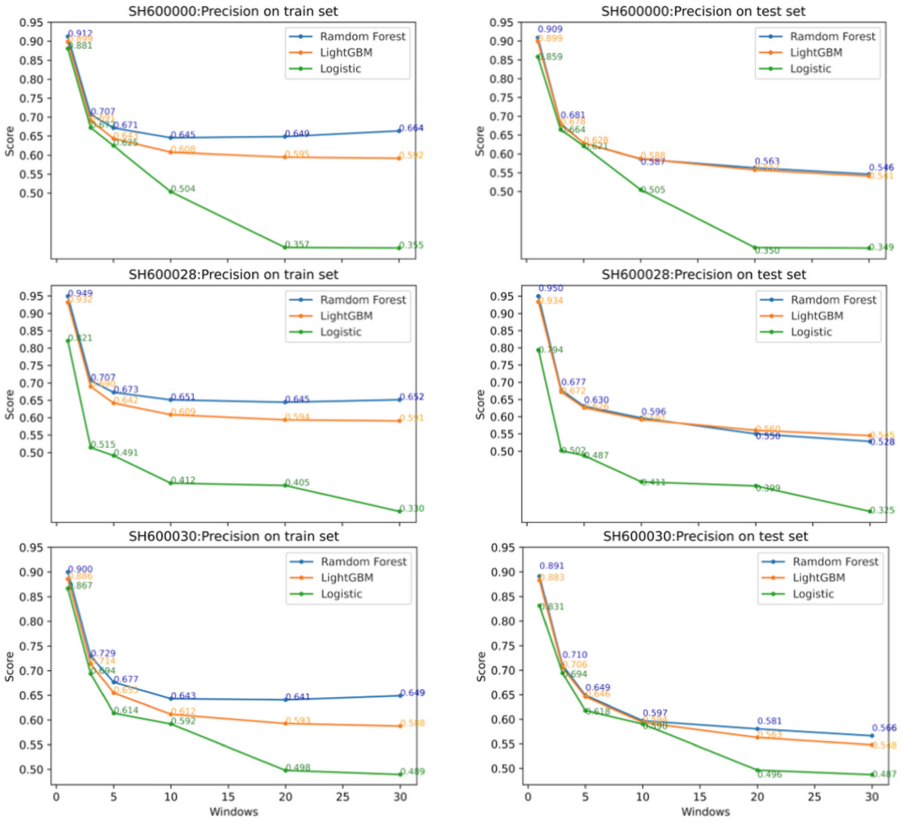

Fig. 3. Models Precision on Training and Testing Sets

For further comparative analysis, the precision and recall of the three models are compared and evaluated, respectively. The results of the three models in the training and testing sets are demonstrated in the Figs. 3 and 4.

In order to evaluate the advantages and disadvantages of different algorithms, F1 score is introduced to estimate the precision and recall of the model as a whole. The F1 scores of the above three models in the training and testing sets are shown in the Fig. 5.

As seen in the Figs. 3, 4 and 5, the time windows are distributed in one, three, five, ten, twenty, and thirty minutes. From the perspective of training sets, the random forest prediction models can generally outperform the LightGBM and Logistic Regression algorithms in predictive accuracy, f1 scores, precision, and recall rate. However, in the short time window (less than three minutes), the performance of the LightGBM model has slight differences from the random forest model.

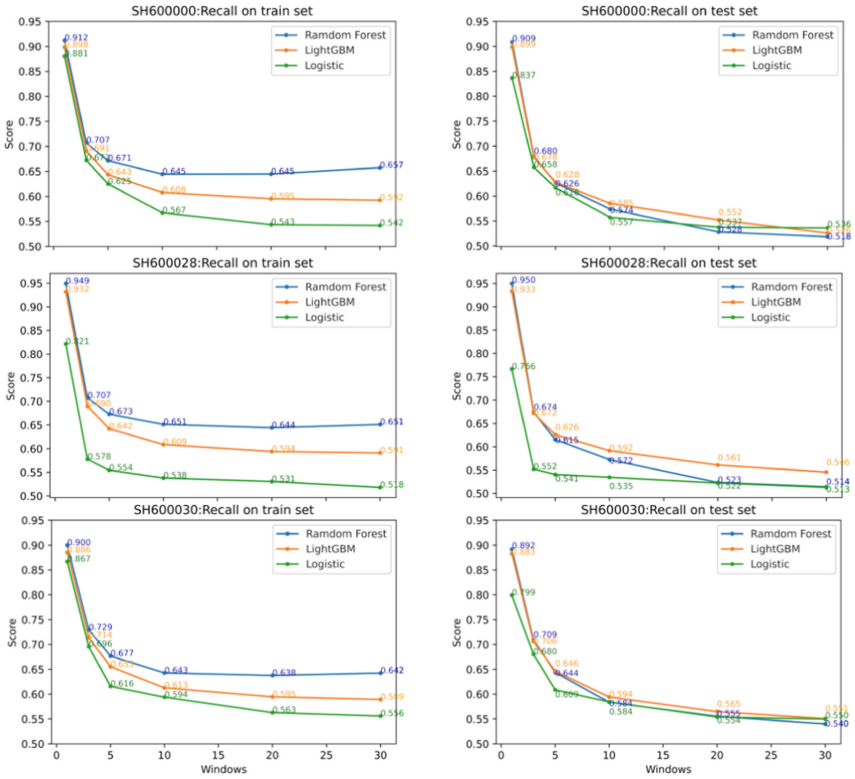

Fig. 4. Models Recall on Training and Testing Sets

From the perspective of testing sets on the three shares, the three prediction models have similar results on accuracy in the long-time window. On the contrary, the logistic regression model significantly decreased the forecasting accuracy in the short term, which has less than ten-time windows. According to the results of f1 scores, LightGBM can be the optimal model due to the significant stability. Specifically, based on the precision results, random forest and LightGBM have similar results all the time. At the same time, the logistic regression dramatically performs worse than other models, especially in the long-time window. Regarding the recall rate, the scores on these models have minor differences in the three shares after ten-time windows, but the logistic regression model obtained much lower scores within ten-time windows.

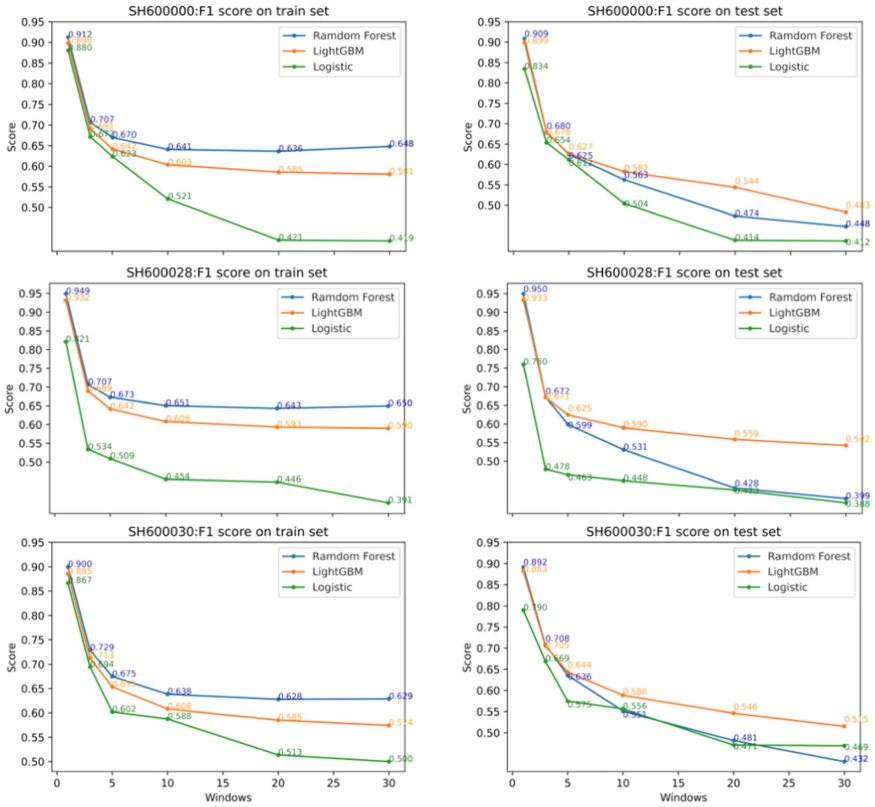

Fig. 5. Models F1 Scores on Training and Testing Sets

4 Discussion

Through comparing three different prediction models, we found consistent results in the three shares. This study suggests that investors can make decisions on machine learning model selections based on the length of time windows. This paper recommends the random forest algorithm for predicting short-term share price tendency because of its relatively high-accuracy and stable performance on precision and recall. For the medium and long-term share price forecasting, this study suggests LightGBM as its outstanding performance on f1 scores. In other words, these two models can be practical to provide investors with valuable information on share trading strategy, as their accurate and stable predictive abilities.

In contrast, the predictive model based on logistic regression might be less likely to provide reliable information for investors because the stock price rise and fall prediction were less precise than the other two models. On this basis, investors may not make appropriate decisions on selling share opportunities and miss signal on the buying shares opportunities. Therefore, the performance logistic regression in this paper model might not fulfill the expectations of application in forecasting the share price movement.

This study may have limitations on optimizing parameters for the most accurate predictions. Additionally, the influential factors of the share price can be complex, and other non-technical indicators should be considered in further studies, e.g., market sentiment, fundamental indicators, and operational conditions. Moreover, ignorant risk management should also be concerned because it might provide more accurate and comprehensive results with investment values for shareholders and investors.

5 Conclusion

In summary, 12 technical indicators are screened as factors for stock price trend prediction based on correlation analysis and significance test, i.e., these indicators have some correlation with the results obtained from the model and do not have high linear correlation with other indicators. Three models, random forest, logistic regression, and lightGBM, were used to forecast the selected three stocks separately to ensure the general applicability and consistency of the results. Based on the rolling forecast method, the performance of the test and training sets of different models in four aspects of accuracy, recall, precision, and f1 value are compared in time windows of 1, 3, 5, 10, 20, and 30 min, respectively, to provide investors with more accurate references.

According to the results, the parameters will be more finely tuned in the future to make them generally applicable and achieve better performance. In addition, fundamental factors and other extraneous factors will be taken into account in future studies to make the prediction results more consistent with the real stock market conditions, and then provide investors with more reliable references.

In brief, several technical factors were initially selected, and based on the results of the screening, 12 of them were finally chosen to fill the gaps in the studies where few or no technical factors were used, making the forecasting effect closer to the real market. The multifaceted comparison of the three models in this paper provides investors with data on the degree of accuracy, the degree of misjudgment, and the overall forecasting level of different models in different time windows. These results offer a guideline for investors to choose appropriate models according to their preferences.

References

1. Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model [C]// 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014:106–112.
2. Breiman, L. (2001) Random Forests, Machine Learning, 45(1), 5–32.
3. Gao ZY. Research on stock price trend prediction based on random forest [D]. China University of Political Science and Law, 2021.
4. Han Yu-Feng, Wang Xiong-Jian, Zhou Guo-Fu, Zou Heng-Fu. Is there a trend in the Chinese stock market? [J]. Financial Research, 2014, 3: 152–163
5. Khaidem L, Saha S, Dey S R. Predicting the direction of stock market prices using random forest [J]. 2016.
6. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30: 3146–3154.

7. Lo A W, Mackinlay A C. A non-random walk down wall street [M]. USA: Princeton University Press, 1999
8. Li JJ. Correlation analysis of stock prices and financial indicators of listed companies [J]. Journal of Huazhong Agricultural University (Social Science Edition), 2014, 5: 138–143
9. Li X.Y. Comparative analysis of neural networks and multi-factor models in the field of quantitative investment [J]. Market Forum, 2018, 8: 65–70
10. Lin Nana, Qin Jiangtao. Research on A-share stock rise and fall prediction based on random forest[J]. Journal of Shanghai University of Technology, 2018(3): 267–273,301.
11. Liu Hongmei. Application of ARIMA model in stock price forecasting [D]. 2008.
12. Richard Frankel, Charles M. C. Lee. Accounting Valuation. Market Expectation, and Cross-Sectional Stock Returns [J]. Journal of Accounting Economics, 2004, 25(3): 283–319
13. Ran Yangfan, Jiang Hongxun. Research on stock price prediction based on BPNN and SVR[J]. Journal of Shanxi University (Natural Science Edition),2018,41(01):1–14.
14. Tailor V M. Exploiting LightGBM Ensemble Method for Stock Prediction[J]. International Journal of Scientific and Engineering Research, 2020, 11(10):648–650.
15. Wang, H. S., Zhang, H. Y., He, T. Y., et al. Exploring the relationship between financial parameters of listed companies and their share price volatility [J]. Securities Market Herald, 2010, 2: 74–77
16. Wang WX, Cai WH. A study on prediction of stock price increase probability based on logistic regression[J]. China Market,2020(06):7–8.
17. Wu Yuxia, Wen Xin. Short-term stock price forecasting based on ARIMA model [J]. Statistics and Decision Making, 2016, 23:83–86.
18. Xu Jingzhao. Analysis of quantitative stock selection based on multi-factor models [J]. Exploration of financial theory, 2017(3):9.
19. Xie G. Stock price prediction based on support vector regression machine[J]. Computer Simulation,2012,29(04):379–382.
20. Ye F, Wang J, Li Z, et al. Jane Street Stock prediction model based on LightGBM [C]// 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP). 2021.
21. Zheng Ruixi. An empirical study on the impact of financial performance on stock prices of listed companies in China [J]. Seeking, 2009, 8: 39–41

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.