Predicting the SP500 Index Trend Based on GBDT and LightGBM Methods

Ma shangchen[1,a]

[1]College of Electronic and Information Engineering, Tongji University, Shanghai China

Abstract: Algorithms that are previously difficult to implement have been successfully applied in different fields because of hardware development. Quantitative investment has the characteristics of rationality and efficiency and has obvious advantages over traditional methods. Based on the SP500 index data for 4936 trading days, 10 characteristics such as PSY, MACD, STOCHK and STOCHD were generated. Based on those features, quantitative investment strategies for the GBDT and LightGBM models were constructed. Validation showed that the annualized returns of the two strategies exceeded the direct purchase and holding of the SP500 index, with the annualized returns of 43.4% and 50.7%. The performance of risk control of the two models was also better than the benchmark strategy. The GBDT model had less risk than the LightGBM model when the same benefits were obtained. The accuracy of the LightGBM model was higher than that of the GBDT, and its F1 score was 0.814, while the GBDT model was 0.805. For the different selected components, the results of the principal component analysis showed that the PSY feature weight in the GBDT model was much higher than other features, and a single feature can be applied for straightforward prediction. In the LightGBM model, the seven feature weights such as STOCHK were relatively balanced, and more features can be balanced at the same time to obtain more accurate results. The article designs investment strategies based on the LightGBM model for the first time and provides new ideas for providing a framework for index investment.

1 Introduction

Quantitative investment refers to a transaction method that uses quantitative methods and computerized programmatic orders to obtain stable returns [1]. The investment strategy generated through quantitative calculation can optimize the expected return within an acceptable risk range [2]. In accordance with the efficient market hypothesis, the fluctuation of stocks is caused by the random spread of information [3]. However, in the actual market operation, people will make emotional decisions, and the linear theory and information equilibrium assumptions do not always hold [4], so the market is in a nonlinear chaotic system [5].

The machine learning method was used to fit the non-linear relationship between the moving average index and the stock return, and a better solution than the traditional method was obtained [6]. The use of big data, machine learning and artificial intelligence methods showed that technology-driven investment solutions were also of great significance to the development of finance [7]. If deep learning was further applied to portfolio investment and risk management, more innovative results could be obtained compared to traditional standard analysis [8]. The specialties of quantitative investment are discipline, systemic, timeliness and accuracy, which can avoid the influence of analysts' subjective emotions, and analyze the market from multiple angles to accurately and timely judge the market.

The method of machine learning starts from a large amount of data information and performs feature extraction. Mathematical models could be used to make market predictions, having investors making a more rational decision. When choosing a stock prediction model, one is to build a classification model, which predicts the probability of change at the next moment based on historical data. The other is a regression model, forecasting the price trend curve. Lo AW et al. Sampled data from the past thirty years and found that there were technical indicators that could be used as reference analysis indicators for quantitative investment [9]. Using SVM, NN, AdaBoost, and linear regression to analyze and predict multiple technical indicators, and confirmed the effectiveness of these model strategies for technical analysis [10]. Using ensemble operators in neural network models can improve the accuracy of multi-layer time-series predictions has been proved by the methods of ensemble learning and the new model of ensemble operator.[11]. By constructing the Gradient Boosted Random Forest to make predictions in a situation existing three different behaviours: buy, sell or stand by,

[a]shangchenma1998@163.com

it can adapt to market conditions where a model does not produce a strong signal of buy or sell [12]. When using LightGBM to make price predictions for 42 cryptocurrency markets, it had a certain reference value for the trend of the cryptocurrency market [13].

The SP500 index forecasting based on GBDT and LightGBM methods will be discussed in this article. The main work is building a model based on the GBDT and LightGBM methods to predict the SP500 index ups and downs. In the second section of the paper, the principles of GBDT and LightGBM will be introduced. Data research is in section three. Section four is the introduction of model parameters and the discussion of model results. Finally, the summary and suggestions for improvement are presented.

2 Models

2.1 GBDT

Decision tree divided data into many nonoverlapping regions. A binary tree consists of non-leaf nodes and leaf nodes. Non-leaf nodes contain a simple decision rule, noting as split (feature, threshold), dividing the current region into two regions. In leaf nodes, each sample x_i belongs to one node. The data in the same leaf has a similar label. Commonly used leaf node splitting criterion has the principle of maximum information entropy. The main idea of GBDT is to use weak classifiers such as decision trees to iteratively train to obtain the optimal model. This model has the advantages of good training effect and not easy to overfit. GBDT is widely used in industry and is usually used for tasks such as click-through rate prediction and search ranking.

The main idea of GBDT is to gradually determine each model and superimpose it on the model collection to ensure the minimum loss function. After training a sub-model, the existing fitting conditions of the model are counted, so as to adjust the setting of the next learning task. Sample labels are modified by gradient boosting, and the new sample label becomes the original label and the residual of the model prediction. The basic formulation is to complete model F(X), which is a composite of base learner f(X). Sub-models are added in the learning to change the compound function so that the loss function decreases along the gradient direction relative to f.

Supposing $F(X)=\sum_{m=0}^{M} f_m(X)$, where $f(X)$ is a basic learner, and then we have

$$F_m(X)=F_{m-1}(X)+f_m(X).$$

Assuming using L2-loss-function $L(y,F_m(X))$, $L=[y_i-F_{m-1}(x_i)]^2$, then we have $\hat{y}_i=y_i-F_{m-1}(x_i)$ .We need to minimize $J=\sum_i L(y,F_m(X))$, which $\frac{\partial J}{\partial F(x_i)}=F(x_i)-y_i$, and to learn $f_m(X)$ to fit $\hat{y}$ by using L2 loss:

$$f_m(X)=\arg\min_{f(X)}\sum_{i=1}^{n}(f(x_i)-\hat{y}_i)^2.$$

2.2 LightGBM

LightGBM is a new Boosting algorithm designed by Microsoft Research Asia. It is based on a decision tree algorithm and is suitable for common tasks such as classification. GOSS, EFB, Histogram optimization and leaf-wise decision tree growth methods on the basis of traditional GBDT are applied in LightGBM algorithm, thereby achieving faster speed without loss Accuracy. When training the model, the training error obtained from samples with small data gradients is also relatively small. The GOSS algorithm removes most of the data with very small gradients. Only a part of the data with small gradients and all the data with large gradients are used to estimate the information gain to avoid the effect of the low-tail long tail. At the same time, the algorithm retains part of the small gradient data to ensure the consistency of the data distribution with the original data and improve the accuracy of the trained model.

In the EFB algorithm, the total conflicts between features are considered as weights, and thus a weight graph is constructed. Then a list is constructed sorting in descending order by weight. Finally, the features are checked in the ordered list and assign them to the bunding with the least conflict. The basic idea of the histogram algorithm is to first discretize continuous floating-point eigenvalues into k integers and at the same time construct a histogram with a width of k, as shown in Figure 1. When traversing the data, the statistics are accumulated in the histogram according to the discretized value as an index. Once the data having been traversed, the histogram accumulates statistics of the data. According to the discrete values of the histogram, iterates to find the optimal segmentation point.

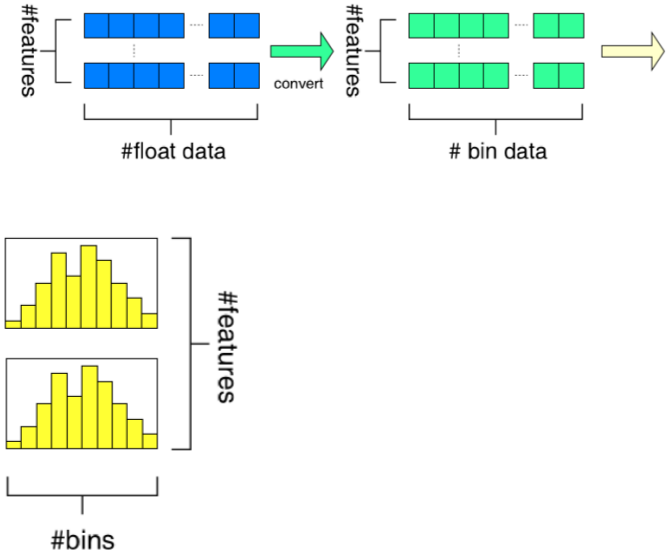

Figure 1 Histogram Algorithm

Figure 2 shows that Decision trees are generated using Leaf-wise principles, which is continuously searching for the leaf node that has the biggest gain after

splitting, and further splits it. This method is fast and effective by continuously searching for the leaf node that has the largest return after splitting, but it is not convenient to speed up the calculation. Limiting the maximum depth of the decision tree can reduce the number of calculations and prevent overfitting.

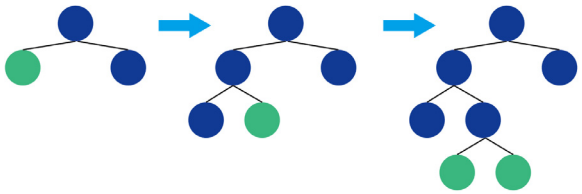

Figure 2 Leaf-wise tree growth

In general, to make LightGBM get better prediction results, the objective function will consist of an error function and a regular term. The result of MT iterations is $F_m(X)=\sum_{m=0}^{M} f_m(X)$. Loss function is recorded as L and the regular term is marked as Ω.The objective function J of LightGBM is:

$$J=\sum_i L(y_i,F_m(X))+\sum_k \Omega(f_k).$$

If the regular term is not considered, then the objective function J

$$J=\sum_i L(y_i,F_m(X))=\sum_i L(y_i,F_{m-1}(X)+f_m(X))$$

learn $f_m(X)$ to fit $\hat{y}$ by using L2 loss: $f_m(X)=\arg\min_{f(X)} J$. In regression tree, supposing threshold is μ, data sample is $(features_i, y_i)$,the total error when splitting is *S,*

$$S=\sum_i (y_i-\mu)^2=(\sum_i y_i^2)-\frac{sum^2}{\|total\|},$$

The split error is S_j :

$$S_j=\sum_{i\in L}(y_i-\mu_L)^2+\sum_{i\in R}(y_i-\mu_R)^2$$

$$=(\sum_i y_i^2)-(\frac{sum_L{}^2}{\|L\|}+\frac{sum_R{}^2}{\|R\|}).$$

$$Split\_gain=S-S_j=(\frac{sum_L^2}{\|L\|}+\frac{sum_R^2}{\|R\|})-\frac{sum^2}{\|total\|}.$$

3 Data Research

The SP500 index is broadly considered the best measure of large US stocks. By recording the index includes about 80% of the 500 leading companies in the available market capitalization, the SP 500 index is considered a measure of the US stock market's average record.

The overall trend of the SP500 index from the 1950s to the present is upward. The growth was slow before 1984, and a wave of rapid growth ushered in from 1984 to 1987. The SP500 index exceeded 300 points, but it fell significantly on Monday, October 19, 1987, and then resumed growth in 1988. Although the SP index fluctuated slightly between 1990 and 1991, it remained relatively stable until 1995. The SP500 index increased rapidly from less than 500 points to about 1500 points during 1995-2000, reaching a maximum value. In 2000, the dot-com bubble began to appear in crisis. Then, the points fell from more than 1,500 to around 700 in 2003. From 2003 to 2008, the SP index ushered in a new wave of rapid growth and exceeded the peak reached in 2000 in the second half of 2007. During 2008 to 2009, a new round of sharp declines fell to less than 700 points. Since 2009, the SP500 index has been increasing, reaching 3337.75 points today (24-2-2020). 10 technical indicators, such as MA, PSY and Aroon based on the closing prices are showed in table 1.

TABLE 1 INDICATORS OF CLOSING PRICES

	Mean	Std	Min	25%	50%	75%	Max
Up_Down	0.5	0.5	0	0	1	1	1
MA	-1.4	25.1	-96.8	-15.9	-4.3	11.0	186.2
PSY	-0.6	17.0	-67.4	-10.3	-2.2	7.7	126.9
Aroon_Up	43.8	36.6	0.0	7.1	35.7	78.6	100.0
Aroon_Down	59.3	36.5	0.0	21.4	71.4	92.9	100.0
CCI	16.7	106.6	-355.7	-69.0	38.7	101.6	318.6
CMO	6.8	22.4	-72.7	-9.3	8.5	23.8	73.4
MACD	0.0	4.9	-30.6	-2.7	0.0	2.7	21.9
RSI	53.4	11.2	13.6	45.3	54.3	61.9	86.7
STOCHK	1.1	0.5	0.1	0.8	1.0	1.2	10.2
STOCHD	1.0	0.3	0.3	0.9	1.0	1.1	2.8

4 Results and Discussion

This data set was derived from Yahoo SP500 index data since 2000, a total of 4936 days. In order to avoid the effects of dividends in the data, Adjusted closing price data was used for training. The training and test set

partition ratio was 8 to 2. Considering the actual situation, the transaction cost was set to a fixed 2%. In the feature indicators, the period parameter of the characteristic indicator Moving Average was 10 days. The Bollinger Band was set to 5 days. Moving Average Convergence / Divergence sets the fastperiod parameter to 12 trading days, slowperiod parameter to 26 trading days and signalperiod to 9 days. Aroon, CCI, CMO and RSI indicators were all set for a period of 14 trading days.

4.1 Parameters of GBDT algorithm and LightGBM

The primary parameters of the GBDT algorithm including the maximum number of iterations of the weak learner, the weight reduction coefficient step size learning_rate of the weak learner, and subsample. In this paper, n_estimators was set to 1000 and learning_rate was set to 0.1 to get the model fit better. **In addition, the important parameters of the GBDT class library weak learner are the maximum feature number, the maximum depth of the decision tree, and the minimum number of samples required for internal node re-division.** Since the number of features is less than 50, the max depth was set to 20.

LightGBM's main parameters are similar to GBDT, including n_estimators, learning rate, etc. The parameters n_estimators and learning rate both affect the fitting effect. **Too small n_estimators parameter will cause underfitting.** The step size is adjusted by the learning_rate parameter. There is a higher probability of a model to converge near the extremes with a low learn rate. And a larger number of iterations need to be set to compensate for a low learning rate parameter. In this model, they were set at 4000 and 0.1 respectively. The model in this article sets num_leaves to 64 and max_depth to 20. The Min_child_weight parameter determines the sum of leaf node sample weights, and it was set to 0.4.

4.2 Results

TABLE 2 EVALUATION OF RESULTS

Parameter	SP500	GDBT	LightGBM
Annualized Returns	0.113	0.434	0.507
Sharpe Ratio	1.56	8.259	7.971
Max Drawdown	0.237	0.124	0.156
Information Ratio	1.77	8.757	8.380
F_1 score	-	0.805	0.814
Accuracy	-	0.784	0.792
Precision	-	0.788	0.790
Recall	-	0.834	0.840

Table 2 is an evaluation of the SP500 index and the results of the two models. The total investment return, average daily return and annualized return of the two models are significantly higher than the SP500 index. The annualized returns of the two machine learning models are 43.6% and 50.7%, respectively and the LightGBM model is better than the GBDT model. The annualized revenue of LightGBM model is 1.16 times that of GBDT model and 4.49 times that of SP500. Considering risk and reporting, the GBDT has a sharp rate and information rate of 8.259 and 8.757 respectively, and a LightGBM sharp rate and information rate of 7.971 and 8.380, respectively. Relatively speaking, GBDT has a higher profit on unit risk. In addition, the maximum retracement rate of GBDT is 0.124 and the maximum retracement rate of LightGBM model is 0.156. Generally, with the same expected return, the GBDT model needs to bear less risk, but its average annualized return is not as impressive as the LightGBM model.

Furthermore, two models are evaluated from the perspective of model accuracy. The Precision and Recall of the GBDT model are 0.788 and 0.834. The Precision of the two is almost the same, and the Recall of LightGBM is relatively better. The Accuracy of GBDT model is 0.784 and LightGBM is 0.792. The F1 score of the LightGBM model is slightly higher than that of the GBDT, so the prediction of the LightGBM model can be considered to be more accurate. Especially when encountering large market fluctuations and relatively high risks, the return gap due to the difference in accuracy will be more obvious. In general, LightGBM model has better prediction accuracy indicators than GBDT. Therefore, the annualized return rate of LightGBM model is higher than GBDT.

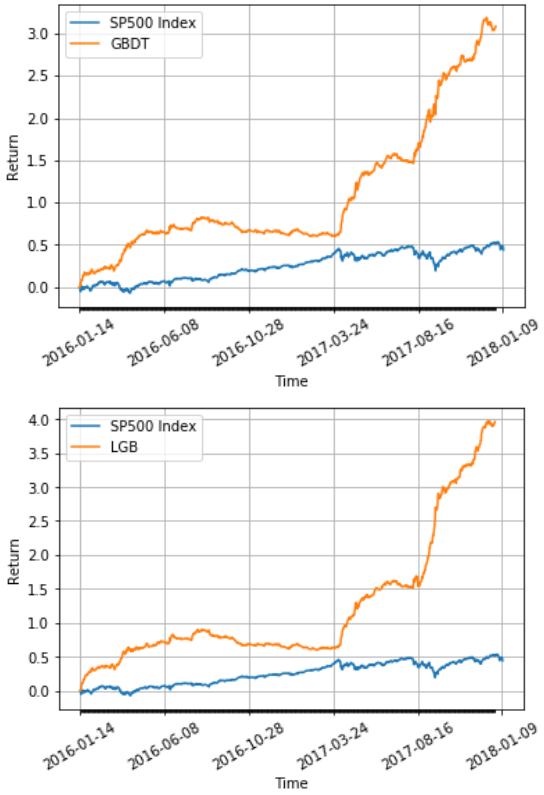

Figure 3 Returns of SP500 and strategies constructed by GBDT and LightGBM

The cumulative returns of the strategies of the SP500 index, GBDT, and LightGBM models are plotted in Figure 3. At most time points after the start of the

investment, the returns of GDBT and LightGBM are higher than SP500. From 2016 to 2018, the total return of LightGBM reached 394%, GDBT was 312%, and SP500 was 48%. When faced with risk, the LightGBM model may experience greater volatility. Combining the above conclusions, the GBDT and LightGBM models are significantly better than the SP500 direct purchase. Although the risk of the GBDT model is relatively lower than that of the LightGBM, in general, the LightGBM model is more accurate than the GBDT strategy, and thus the LightGBM has better return results.

4.3 Discussion

In total, ten indicators were selected as features. In order to obtain the weights of the features on the models, a PCA analysis of the number of features from 1 to 10 was performed on the two models, and the results were recorded in figure 6. It can be found that the F1 score of both models is increasing as the number of features used increases. With the same amount of features, the F1 score of the GBDT model is generally not lower than the value of the LightGBM model, only slightly lower when n is equal to 3 or 4. When n is equal to 1, 9, or 10, F1 score of the two models is equal. Therefore, it can be considered that GBDT is easier to obtain a higher accuracy rate with a smaller number of features, and with enough features, the prediction accuracy rate of LightGBM is not significantly different from the LightGBM model.

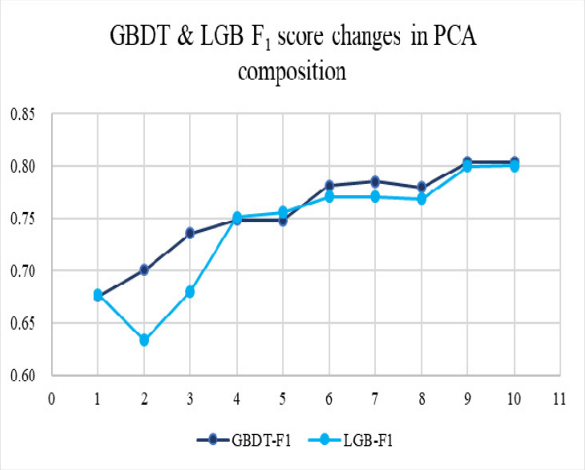

Figure 4 F1 score of GBDT and LightGBM changes in PCA

Figure 5 shows the two model feature indicators in descending order of weight. We can more directly see the contribution of different feature indicators to the forecast of fluctuations In the GBDT model, the PSY feature weight is significantly larger than other indicators, accounting for an absolute large proportion. Features STOCHD and STOCHK account for the second and third largest weights, respectively, but only 23% and 15% of PSY. The weights of other features in the GBDT model are all less than 10%, so only the three features can be used for simple prediction.

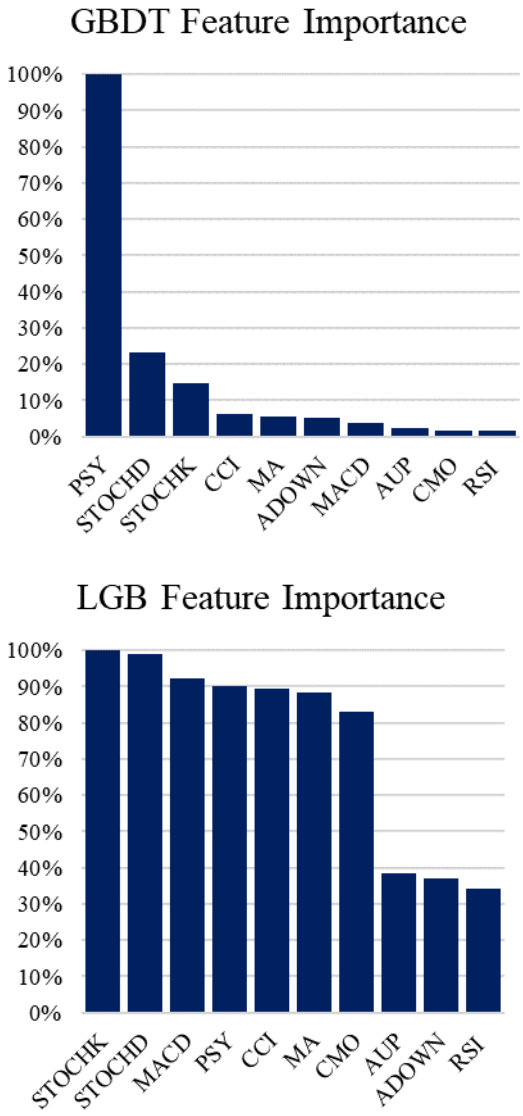

Figure 5 Featuer importance of GBDT and LightGBM

In the LightGBM model, seven feature weights are significantly higher than the other three. The five feature weights of STOCHK, STOCHD, MACD, PSY, and CCI all exceed 90%, the MA and CMO feature weights exceed 80%, and the feature weights of Aroon_Up, Aroon_Down, and RSI are less than 40%. The weight difference between STOCHK and STOCHD is not large. The weights of MACD, PSY, and CCI are basically the same. This weight distribution indicates that the 10 selected features all have a large prediction contribution to the LightGBM model. The LightGBM model can utilize each feature in a balanced manner. If the accuracy of the model needs to be improved, adding new features is a solution.

5 Conclusion

In this paper, the statistics of the SP500 index over the past 4936 days were used as input, and a LightGBM and GBDT algorithm model was established to predict the future index change. The results of this paper showed

that both model strategies sere better than the baseline strategy. The annualized return of LightGBM model reached 50.7%, which was better than 43.4% of GBDT model. The risk control ability of the GBDT model was a bit greater than the LightGBM model. In terms of model accuracy, the F1 score of the LightGBM model was 0.814, which was a little higher than the 0.805 of the GBDT model, indicating that the LightGBM model had higher accuracy. After analyzing 10 features by PCA, it was found that the LightGBM model could use the ten features more evenly, so it is more suitable for complex predictions when there are multiple features.

If the model needs to be further optimized, it needs to be re-screened from multiple experiments with a wider range of technical indicators, or a fusion model that can use SVM, NN, and Boost methods together. Similar models to the LightGBM model include XGBoost and CatBoost. The XGBoost model has universal applicability, but it has the disadvantages of complex parameter adjustment and slow operation. CatBoost is suitable when categorical variables are included in the data. If investment strategies need to be further optimized, investment methods can be improved. The way of portfolio investment or considering the impact of macroeconomic indicators and real-time events and current affairs policies can be used as an optimization means.

References

1. DeFusco R A, McLeavey D W, Pinto J E, et al. Quantitative investment analysis[M]. John Wiley & Sons, 2015.
2. Lakonishok J, Shleifer A, Vishny R W. Contrarian investment, extrapolation, and risk[J]. The journal of finance, 1994, 49(5): 1541-1578.
3. Malkiel B G, Fama E F. Efficient capital markets: A review of theory and empirical work[J]. The journal of Finance, 1970, 25(2): 383-417.
4. Grossman S J, Stiglitz J E. On the impossibility of informationally efficient markets[J]. The American economic review, 1980, 70(3): 393-408.
5. Goodwin R M. Chaotic economic dynamics[M]. Oxford University Press: Oxford, 1990.
6. encay R. Non-linear pprediction of security returns with moving average rules[J]. Journal of Forecasting, 1996, 15(3): 165-174.
7. Zetsche D A, Buckley R P, Arner D W, et al. From FinTech to TechFin: The regulatory challenges of data-driven finance[J]. NYUJL & Bus., 2017, 14: 393.
8. Heaton J B, Polson N G, Witte J H. Deep learning for finance: deep portfolios[J]. Applied Stochastic Models in Business and Industry, 2017, 33(1): 3-12.
9. Lo A W, Mamaysky H, Wang J. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation[J]. The journal of finance, 2000, 55(4): 1705-1765.
10. Thira Chavarnakul, David Enke. Intelligent technical analysis based equivolume charting for stock trading using neural networks[J]. Expert Systems with Applications, 34(2):1004-1017.
11. Kourentzes N, Barrow D K, Crone S F. Neural network ensemble operators for time series forecasting[J]. Expert Systems with Applications, 2014, 41(9): 4235-4244.
12. Qin Q,Wang Q G, Li J, et al. Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market[J]. Journal of Intelligent Learning Systems and Applications, 2013, 5(01): 1.
13. Sun X, Liu M, Sima Z. A novel cryptocurrency price trend forecasting model based on LightGBM[J]. Finance Research Letters, 2018.