ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

Quantitative investment model based on LightGBM algorithm

Yangyang Guo[1], Huihuang Ma[1], Shiyu Tian[2,*]

[1]School of Electronic Science and Engineering, Hunan University of Information Technology, Changsha, Hunan, 410100, China
[2]College of General Education, Hunan University of Information Technology, Changsha, Hunan, 410100, China
**Corresponding author*

Abstract: *Quantitative investment refers to the trading method through quantitative way and computer program issued orders, in order to obtain stable income. Firstly, this paper completes Pearson correlation analysis. Finally, the correlation coefficients of 50 indicators were calculated according to relevant data, and 8 indicators with the highest correlation were obtained. Secondly, according to the LightGBM model, multiple linear regression model and BP neural network model, the three prediction models are used to forecast the trading volume of the "digital economy" plate from January 4, 2022 to January 28, 2022 respectively. Thirdly, the mean absolute value error, mean square error, R and other model evaluation indexes of the three prediction models are analyzed. Finally, the accuracy of LightGBM model is better, followed by BP neural network model, and multiple linear regression is the worst.*

Keywords: *LightGBM algorithm, correlation analysis, BP neural network*

1. Introduction

Quantitative investing is a trading method that uses quantitative methods and computer programming to buy and sell orders in order to achieve steady profits[1].Investors use data analysis to learn the rules of the market and predict market movements to make trading decisions.With the development of big data technology, quantitative investment is becoming more and more important in global financial markets. However, given the heterogeneity of market data and many other factors affecting commodity prices, it is difficult to draw effective indicators and develop trading strategies from a large amount of market information[2].

In this paper, the "digital economy" index of a quantitative investment sector every 5 minutes from July 14, 2021 to December 31, 2021 is used as the training set[3], and the "digital economy" index of a quantitative investment sector every 5 minutes from January 4, 2022 to January 28, 2022 is used as the test set.Each index is extracted to predict the trading volume of "digital economy" index every 5 minutes[4]. The "digital economy" index every 5 minutes from July 14, 2021 to December 31, 2021 is used as the training set, and the "digital economy" index every 5 minutes from January 4, 2022 to January 28, 2022 is used as the test set.According to (1) and (2), the model is established to forecast the "digital economy" sector index (closing price) every 5 minutes.

2. Correlation analysis

2.1 Correlation analysis formula

Analyze the 50 indicators and extract the main indicators related to the "digital economy" plate.First of all, we take the closing price of the digital economy plate at 15:00 every day as y, and take the corresponding index parameter of the day as x_i (i =1,2,3... 50), we use Pearson correlation coefficient method[5], also known as product difference correlation coefficient, which is specially used to measure the degree of linear correlation between, and the letter R represents the correlation coefficient.$x_i y$Will, and to the Pearson correlation coefficient in the formula to calculate, then calculated after 50 value, according to the size of the values for sorting, |r | more hasten is in January, the stronger the correlation, if |r|≥0.6, then on the basis of selected eight "digital economy" plate on the main index as evaluation

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

index of forecasting model. Pearson correlation coefficient formula:

$$pearsonr = \frac{\sum_{i=1}^{n}(x_{ij}-\overline{x_i})(y_j-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\overline{x_i})^2 \sum_{i=1}^{n}(y_j-\bar{y})^2}} \quad (1)$$

Wherex_{ij}, $\overline{x_i}$, $\bar{y}$, $y_j$and are respectively the ith data of the jth indicator, The average value of the ith indicator, the average value of the closing price of the digital economy sector at 15:00 every day, and the closing price of the digital economy sector at 15:00 on day J.

After the model is established, follow the following steps to solve it.

2.2 Correlation analysis process

The flow chart of correlation analysis is shown in Figure 1.

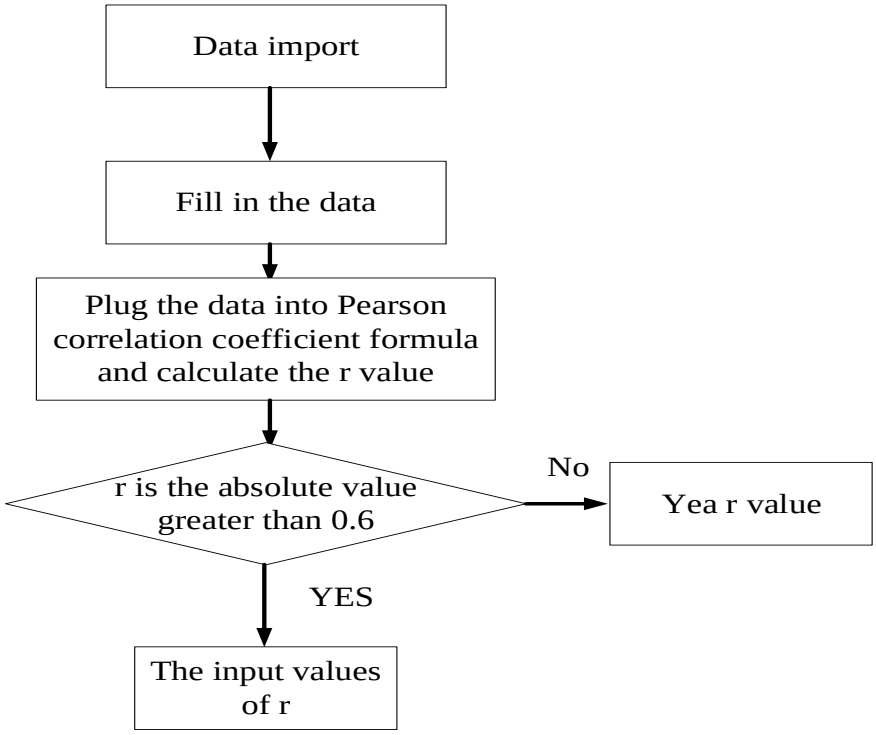

Figure 1: Flowchart of correlation analysis

According to the Pearson correlation coefficient model above, correlation coefficient R of 50 indicators can be finally worked out (see the attached table of correlation coefficient for specific values) [7], and 8 main indicators of | r |≥ 0.6 are taken, namely gem index, Shenzhen Component index, OBV, BBI, DMA, MA, EXPMA and MACD, as shown in Table 1 below.

Table 1: Correlation coefficients of main indicators

	The closing price	Gem index	Shenzhen component index	OBV	BBI	DMA	MA	EXPMA	MACD
The closing price	1.000	0.794	0.802	0.805	0.912	0.704	0.940	0.963	0.732
Gem index	0.794	1.000	0.871	0.479	0.634	0.759	0.712	0.740	0.821
Shenzhen component index	0.802	0.871	1.000	0.503	0.667	0.697	0.718	0.753	0.745
OBV	0.805	0.479	0.503	1.000	0.845	0.550	0.812	0.829	0.491
BBI	0.912	0.634	0.667	0.845	1.000	0.643	0.977	0.979	0.582
DMA	0.704	0.759	0.697	0.550	0.643	1.000	0.726	0.722	0.954
MA	0.940	0.712	0.718	0.812	0.977	0.726	1.000	0.994	0.710
EXPMA	0.963	0.740	0.753	0.829	0.979	0.722	0.994	1.000	0.708
MACD	0.732	0.821	0.745	0.491	0.582	0.954	0.710	0.708	1.000

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

0.732	0.821	0.745	0.491	0.582	0.954	0.710	0.708	1.000
0.963	0.740	0.753	0.829	0.979	0.722	0.994	1.000	0.708
0.940	0.712	0.718	0.812	0.977	0.726	1.000	0.994	0.710
0.704	0.759	0.697	0.550	0.643	1.000	0.726	0.722	0.954
0.912	0.634	0.667	0.845	1.000	0.643	0.977	0.979	0.582
0.805	0.479	0.503	1.000	0.845	0.550	0.812	0.829	0.491
0.802	0.871	1.000	0.503	0.667	0.697	0.718	0.753	0.745
0.794	1.000	0.871	0.479	0.634	0.759	0.712	0.740	0.821
1.000	0.794	0.802	0.805	0.912	0.704	0.940	0.963	0.732

Figure 2: Thermal diagram of correlation coefficient of main indexes

3. Volume prediction algorithm

3.1 LightGBM model

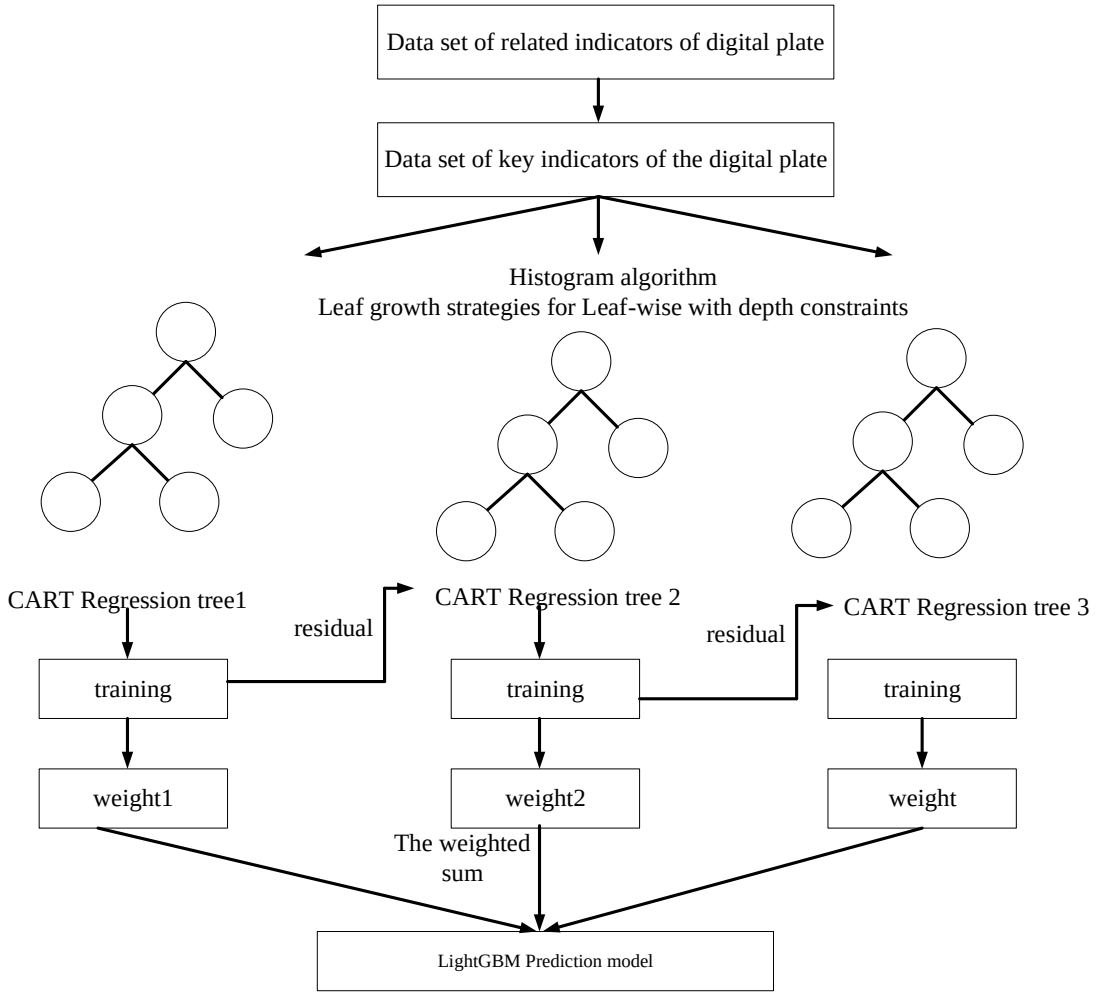

Figure 3: LightGBM model diagram

LightGBM algorithm is an improved algorithm of gradient lifting decision tree, which is also an additive model in essence, and can solve the prediction problem.LightGBM algorithm mainly optimizes the optimal segmentation point search strategy and leaf growth strategy of the decision tree, and further

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

reduces the training time and memory consumption of the model while ensuring the prediction accuracy of the model.Based on the data of 20 main indicators, this paper establishes the prediction model of "digital plate" trading volume based on LightGBM algorithm.The model is shown in Figure 3 below[8].

Volume prediction algorithm of "Digital Plate" based on LightGBM model:

In the first step, input the main index data set of "digital plate" after feature selection, use histogram algorithm to find the optimal segmentation point of feature, and use leaf-Wise Leaf growth strategy with depth limitation to generate CART regression tree[9].

The second step is to calculate the residuals of the first CART regression tree and take the residuals of the previous round as the training samples of the next CART regression tree to continuously fit residuals for repeated training.

The third step is to weighted sum the CART regression tree generated by each round of training to obtain the final "digital plate" trading volume prediction model, and predict the "digital plate" trading volume according to the model.

3.2 Multiple linear regression model

Multiple linear regression is the most widely used mathematical model in practical life.Regression analysis is to analyze the functional relationship between explanatory variables and explained variables according to the experimental results, establish approximate expressions of the relationship between variables, and predict the explained variables.Linear regression is a mathematical model often used in solving practical problems and is also applied to stock market prediction.The input variable of the linear regression valuation model is 20 main indicators, and the output variable is the volume of the "digital economy" sector index every 5 minutes.

Multiple linear regression model can be expressed

$$P = \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4 + \cdots + \theta_{20} z_{20} + \theta_0 \tag{2}$$

Where P is the predicted value of the predicted trading volume,.$\theta_1\theta_2\theta_3\theta_4\theta_{20}$Is the weight before each index,$z_1z_2z_3z_4z_{20}$are the main indicators.$\theta_1$and...$\theta_2\theta_{20}$The estimated value of...$b_0b_1b_{20}$.

$$P_i = +\ldots b_0 + b_1 z_{i1} + b_{20} z_{i20} \tag{3}$$

$$Z=\begin{pmatrix} 1z_{11}z_{12}\ldots z_{120} \\ 1z_{21}z_{22}\ldots z_{220} \\ 1z_{31}z_{32}\ldots z_{320} \\ 1z_{41}z_{42}\ldots z_{420} \\ \ldots \\ 1z_{n1}z_{n2}\ldots z_{n20} \end{pmatrix} \tag{4}$$

Convert it to a matrix for computational convenience, and get

$$P=B=\begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ \ldots \\ P_n \end{pmatrix}\begin{pmatrix} b_0 \\ b_1 \\ \ldots \\ b_{20} \end{pmatrix} \tag{5}$$

Can be converted to

$$Z^T ZB = Z^T P \tag{6}$$

Where, is the transpose matrix, assuming it exists, then$Z^TZ(Z^TZ)^{-1}Z^T$

$$B\hat{}==P\begin{pmatrix} b_0 \\ b_1 \\ \ldots \\ b_{20} \end{pmatrix}(Z^TZ)^{-1}Z^T \tag{7}$$

The regression equation can be obtained:

$$P\hat{}=b_0 + b_1 Z_1 + \cdots + b_{20} Z_{20} \tag{8}$$

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

3.3 BP neural network model

The network structure of BP neural network algorithm is divided into three layers, each layer contains N neurons, the neurons of the same layer are independent of each other, and the neurons of different layers are connected with each other.The input and output of each neuron depend on the function and threshold value, and its structure is shown in Figure 4.

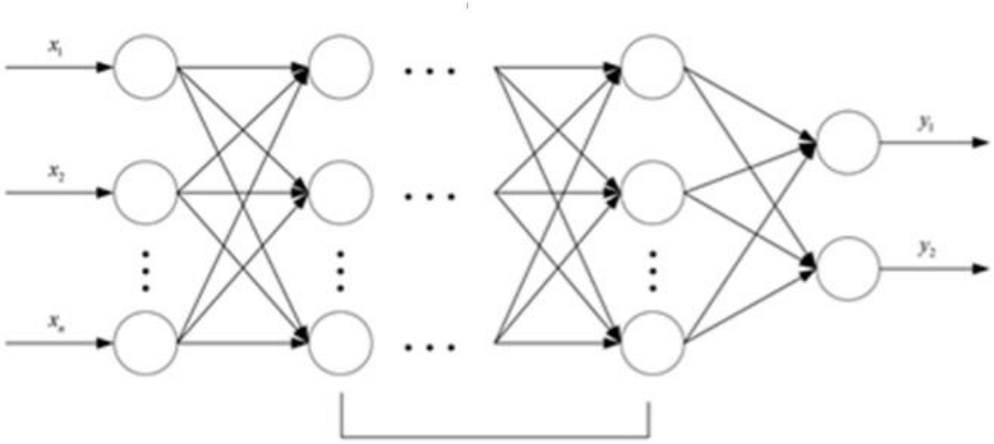

Figure 4: BP neural network algorithm structure

Volume prediction algorithm of "digital plate" based on BP neural network model:

When establishing BP neural network prediction model, network structure is required, including the number of input layer neurons, output layer neurons and hidden layer neurons.

(I) Number of neurons in the input layer.The number of neurons at the input layer depends on the dimension of input variables. The input variables in this study are 20 major indicators, so the number of neurons at the input layer is 20.

(II) Number of neurons in the output layer.The number of neurons in the output layer depends on the dimension of sub-output variables. The output variable in this study is the trading volume of "digital economy" plate index every 5 minutes, so the number of neurons in the output layer is 1.

(III) Number of hidden layer neurons. As for the number of hidden layer, the neural network of single hidden layer can basically meet the research requirements, so the number of hidden layer is set as 1.

3.4 Solving the model

LightGBM model, multiple linear regression model and BP neural network model are used to predict the trading volume of "digital economy" plate every 5 minutes from January 4, 2022 to January 28, 2022, respectively, and analyze the mean absolute value error, mean square error, R and other model evaluation indicators.The errors of the three models are shown in Table 2.Among the three models, LightGBM model has the smallest mean square error and the highest R correlation coefficient. The smaller mean square error, the better accuracy of the prediction model.As can be seen from the table, among the three models, the Accuracy of LightGBM model is better, followed by BP neural network model, and multiple linear regression is the worst.The solution diagrams of the three models are shown in Figure 5, Figure 6 and Figure 7.

Table 2: Volume error comparison of each model

	Multiple linear regression	BP neural network	LightGBM
Mean absolute value error	0.00111	0.01182	0.0090
Mean square error	2.92756	0.00019	0.00016
R^2score	0.96164	1.18714	1.22509

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

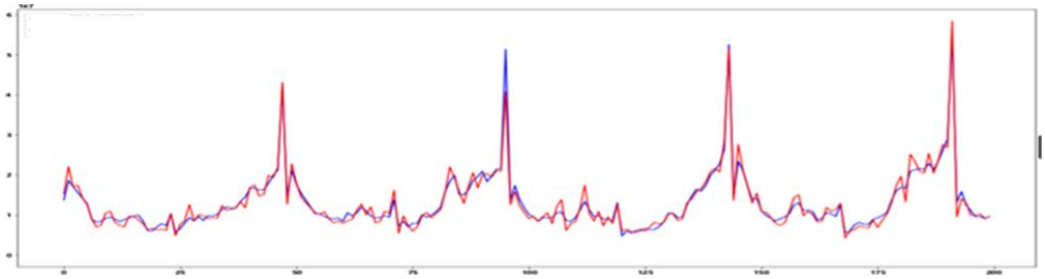

Figure 5: Comparison of trading volume prediction effect of LighTGBM model

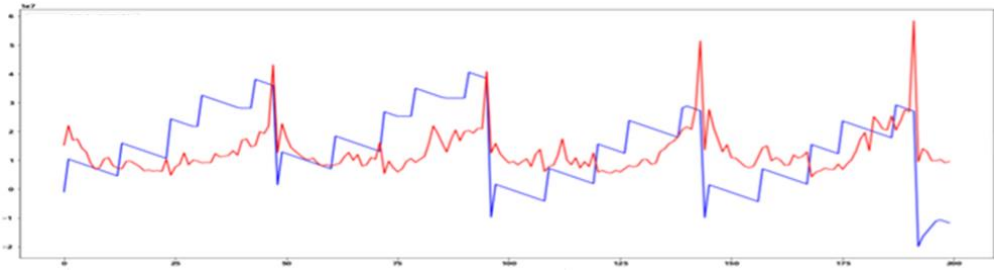

Figure 6: Comparison of trading volume prediction effect of multiple linear regression model

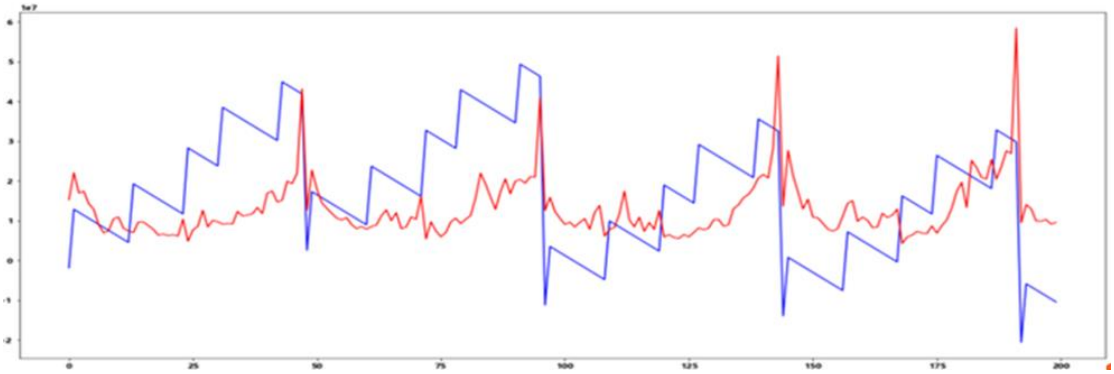

Figure 7: Comparison of trading volume prediction effect of BP network model

4. Evaluation of the model

This paper adopts a variety of models for prediction analysis, and comparative analysis of the results, it is concluded that LightGBM model is more suitable for quantitative investment trading, unique thinking, novel method, faster training speed and higher efficiency of the model, can reduce the memory utilization rate.From the perspective of accuracy, LightGBM model has a higher accuracy. The model has a small amount of calculation and fast convergence speed, which can quickly obtain data results.

Acknowledgements

Fund Project: Teaching reform project of College university mathematics teaching research and development center (Project No.: CMC20210119)

References

[1] Chen Yantai, Chen Guohong, Li Meijuan. Classification and research progress of comprehensive evaluation methods [J]. Journal of management science, 2004,7(2).
[2] Chen Minqiong, PENG Donghai. A note on the calculation formula of partial correlation coefficient [J]. Journal of chuzhou university, 2014,16(2).
[3] Guo Xiaojing, HE Qian, Zhang Dongmei, et al. Science and technology management research, 2012,32(20).
[4] Wu Jingyi, Shi Benshan. Analysis of the influence of evaluation mode on evaluation reliability [J].

ISSN 2616-5775 Vol. 5, Issue 7: 40-46, DOI: 10.25236/AJCIS.2022.050707

Systems engineering-theory & practice, 1993,5(3).
[5] Yang Y Y. Research on theory and Application of real estate Appraisal -- Comment on Real Estate Appraisal [J]. Contemporary Education Science, 2015(24):72.]
[6] Wang Yixiang, Chen Jiying, Wang Shengquan, Li Ang. Real estate price prediction based on improved rf-bp neural network [J]. Industrial control computer, 2019,32(10):122-124.
[7] Li Yingbing, Chen Yujin, Ouyang Qian. Research on wuhan Second-hand House Valuation Model Based on BP Neural Network [J]. Digital Manufacturing Section Science, 2017, (Z1) : 66-70.
[8] Ma X , Sha J , Niu X . An Empirical Study on the Credit Rating of P2P Projects based on LightGBM Algorithm[J]. The Journal of Quantitative & Technical Economics, 2018.
[9] Zhou d s. Research on data preprocessing method under the background of big data [J]. Shandong chemical industry, 2020,49(01):110-111+122. (in Chinese)