

Analysis of quantitative trading investment influence factors based on KPCA-LightGBM algorithm

Chuwei Feng^{1, #}, Jiali Li^{2, #}, Mengliang Liu^{3, #, *}

¹ School of economics and finance, south China University of Technology, 510641, Guangzhou, China

² The College of Humanities, Arts, and Social Sciences, University of California, Riverside, 92521, Riverside USA

³ School of Management Science and Engineering, Central University of Finance and Economics, 102206 Beijing, China

* Corresponding author: liumengliang0477@163.com

#These authors contributed equally.

Abstract. This study proposes a stock price forecasting model based on Kernel Principal Component Analysis (KPCA) and Light Gradient Boosting Machine (LightGBM) to examine the connection between the impact factors and the CSI 300 index. As a case, this study focuses on the CSI 300 Index. Data gathered from January 2005 to July 2022, including 16 clear indicators such as CPI, IPI, and Dow Jones Industrial Index, were simplified into five dimensions by the KPCA. Subsequently, the LightGBM algorithm was applied through 5-fold cross-validation to adjust the parameters and achieve optimal model performance. The results demonstrate that the combined KPCA-LightGBM model has strong generalizability and stability, with an RMSE and R2 of 0.001 and 0.837, respectively. Furthermore, reducing factors to five dimensions by KPCA can also provide improved decision-making metrics for investors.

Keywords: Impact Factors, Quantitative Investment, KPCA, LightGBM, Stock Price Forecast.

1. Introduction

February 1, the full implementation of the stock issuance registration system reform officially launched. After the registration system reform, investors will face a more harsh market. After the implementation of the comprehensive registration system, the way for companies to list in various sectors is more smooth. Because there will be no too much human intervention and administrative intervention, the conditions for companies to issue and list are more inclusive, and the number of listed stocks has skyrocketed. In particular, innovative and growth-oriented companies will accelerate the listing. These companies naturally have high-risk characteristics. Therefore, for ordinary investors, the investment risks are getting bigger and bigger, and stock selection becomes more and more difficult. As one of the forms of financial technology innovation, intelligent quantitative investment, based on artificial intelligence technology to carry out asset management business, can greatly improve the efficiency of asset management, and will play an incomparable role in many stocks in the future. In this context, how to improve the accuracy of intelligent quantitative investment is particularly important.

In recent years, with the rapid development of artificial intelligence technology, machine learning and deep learning have become intelligent quantitative investment methods that have attracted much attention in recent years. Machine learning learns a reproducible model from a large amount of data and uses the acquired model to predict. Because the machine learning algorithm has high prediction accuracy and good generalization ability, the excellent prediction ability of the model can be applied to out-of-sample data. This technology is widely used in quantitative investment [1]. Pan and Xu (2011) tested the relationship between six factors and stock returns [2], and constructed a factor model suitable for the A-share market. Kozak [3] and Kelly [4] used PCA method and IPCA method (Instrumented PCA) to extract common factors in factors respectively, and found that models based

on a few principal components can independently predict cross-sectional returns. Li (2019), Shao (2019) and Li (2019) used 12 kinds of machine learning algorithms such as recurrent neural network and long-term and short-term memory network [5] to construct stock return prediction models and investment portfolios, which have good benefits. In machine learning, GBDT is an enduring model, which has the advantages of good training effect and not suitable for overfitting, so it is widely used in machine learning.

However, in the face of high data dimensions, the traditional GBDT will increase the computational complexity of the algorithm and is difficult to apply to the massive data of the stock market. Therefore, this paper applies the LightGBM algorithm to quantitative investment, giving full play to its excellent reliability and flexibility. At the same time, because there are nonlinear patterns between the anomaly factors that are difficult to identify by traditional linear regression, we use the nonlinear extension algorithm of PCA-KPCA to extract the principal components. By processing the original data, the indicators that consumers need to consider are reduced, and the signal-to-noise ratio of data information is improved. The LightGBM method is used in parallel to contribute to the factors. The KPCA-LightGBM coupled quantitative trading model proposed in this paper has the advantages of efficiently processing a large amount of data in the stock market, accurately identifying the nonlinear pattern of anomaly factors, supporting efficient parallel training, faster training speed, and lower memory consumption. It can provide investors with a reasonable decision support system.

2. Construction of CSI 300 index system

2.1. Introduction of CSI 300

The CSI-300 index is composed of 300 securities with large scale and good liquidity in the CSI markets. The CSI 300 Index was born on April 8, 2005, which means that since the establishment of China's securities market on December 19, 1990, for the first time, there has been an index that can reflect the entire A-share market. Compared with the Shanghai Stock Exchange 50 Index and the Shenzhen Stock Exchange 100 Index, the CSI 300 Index can basically represent the overall trend of stocks in China [7].

Therefore, this paper chooses CSI 300 as the research object, which can have certain representativeness and practical significance, and can further study the impact of volatility in China's domestic securities market and quantify investment decision-making recommendations.

2.2. The construction of index system affecting CSI 300

2.2.1. Macroeconomics

Macroeconomic fundamentals and their changes are the main sources of systemic risk in the stock market. Their impact on stock market volatility is characterized by wide coverage, large volatility and long duration. The impact of macroeconomic fundamentals is to affect the performance of listed companies through macroeconomic prosperity factors acting on macroeconomic industrial factors or microeconomic enterprise factors, thus indirectly acting on investors' investment psychology to make the stock market volume price relationship change significantly, which leads to the overall fluctuation of the market.

Although Harris's research shows that the relationship between the stock market and the macro economy in underdeveloped countries is very weak, the impact of the macro economy on the stock market cannot be ignored considering China's economic physique, large volume and rapid development of the financial sector. Therefore, this paper selects 10 indicators such as GDP and CPI from the macro economy to study the impact.

2.2.2. Investor sentiment

Traditional economics is used to study from the perspective of 'rational man', but behavioral finance believes that investment behavior can be regarded as a psychological process. It is believed that in addition to the objective factors of stocks, the psychological factors and behavioral

characteristics of investors will also have an important impact on the price of stocks. For example, the psychology of ' overreaction ' will lead investors to follow the information related to their own interests with a keen attitude, so that the yield in the market will continue to reverse ; the psychology of ' underreaction ' will lead investors not to respond to the information in the market in time. According to Lu and Li [2], investor sentiment has a significant impact on the stock market index.

Since the turnover rate can reflect investor sentiment, the higher the turnover rate, the higher the investor sentiment. Therefore, this paper selects the turnover rate as an indicator to reflect investor sentiment [6].

2.2.3. Financial indicators

Financial status is an important factor affecting the stock price changes of listed companies. Investors can make correct investment decisions by studying the financial indicators of listed companies. At the same time, listed companies can attract more investors and raise more funds for the company by enhancing their profitability. For example, the higher the price-earnings ratio and return rate, the stronger the company 's profitability, which is more favored by investors. Therefore, among the many financial indicators that reflect the financial status of the company, it is particularly important to obtain important financial indicators with investment identification [7].

Therefore, this paper selects two indicators of price-earnings ratio and return rate to study the relationship with stock prices [8].

2.2.4. Stock index futures market

The research on the contribution of the futures market and the spot market to the price discovery process shows that [3] the impact of changes in the futures market on the spot market is significant, and the stock index futures market plays a dominant role in the price discovery process, especially the main market for setting prices. Although the characteristics of China 's stock market are different from those of developed countries, the research on the market of developed countries has a certain reference for the launch of China 's stock index futures.

Therefore, this paper selects the price discovery process of five stock index spot markets and futures markets in the United States, Britain, Japan and China 's Hong Kong market. The five stock indexes are the Standard & Poor 's 500 Index (S & P500), the Dow Jones Industrial Average (DJIA), the Hang Seng Index (HIS), the Nikkei 225 Index (NIKKEI225) and the Financial Times 100 Index (FTSE-100). To sum up, the index system of CSI 300 constructed in this paper is shown in Table 1 :

Table 1. Indicators affecting the CSI 300 Index

Indicator number	Name of indicator	Reason for selection	unit
X1	CPI	CPI is an important index to measure inflation.	%
X2	Interbank 30-day lending rate	The level of interbank interest rate sensitively reflects the supply and demand of money market funds.	%
X3	IPI	IPI is an industrial production index, which is used to calculate and reflect the speed of industrial development. It is also the preferred index for business analysis.	-
X4	sum of investments in fixed assets used	The amount of investment in fixed assets is a comprehensive index reflecting the scale, speed, proportion and direction of investment in fixed assets.	100 million yuan
X5	M2	M2 is the broad money supply. The higher the index is, the higher the inflation risk is.	100 million yuan
X6	u.s. dollar against the rmb exchange rate	As the world 's common currency, the exchange rate of the RMB exchange rate directly affects the domestic stock market.	yuan
X7	consumer confidence index	Consumer confidence index is an indicator of consumer confidence.	

X8	GDP	GDP is the core indicator of national economic accounting and an important indicator to measure the economic status and development level of a country or region.	100 million yuan
X9	Dow jones industrial average index	The Dow Jones Industrial Index is an index measuring the development of industrial composition in the US stock market.	%
X10	s&p500	The S & P 500 Index is a stock index that records 500 listed companies in the United States.	%
X11	Hong Kong Hang Seng	Hang Seng Index is the most influential stock price index reflecting the floating trend of Hong Kong stock market.	%
X12	Nikkei 225	The Nikkei index is a barometer of whether the Japanese stock market or capital market is good or bad.	-
X13	FTSE 100	The FTSE 100 Index is an important measure to reflect the changes in the stock market in London.	-
X14	rate of return	Return rate is an indicator used to measure how much net profit is created per unit of assets.	-
X15	price-earning ratio	The price-earnings ratio is an indicator used to measure the relationship between stock prices and stock returns.	-
X16	buying-selling rate	Turnover rate is one of the indicators reflecting the strength of stock liquidity.	-

3. KPCA-LightBGM Method

3.1. KPCA Method

Kernel Principal Component Analysis (KPCA) is a non-linear extension of Principal Component Analysis (PCA), first proposed by the Schölkopf team in 1988. Unlike PCA, which is a linear dimension reduction method, KPCA is based on the principle of non-linear mapping. The KPCA algorithm transforms the original data into a high-dimensional feature space through a non-linear mapping function, making it linearly separable, then performs PCA dimension reduction in the high-dimensional space. Compared to PCA, KPCA has a stronger non-linear processing ability. The application principle of this algorithm is as follows [9]:

Under the condition that R^p represents the input space and p represents the dimension of the data, with i being a positive integer ($1, 2, \dots, n$), this study considers a data set $X = \{x_1, x_2, \dots, x_n\}, x_i \in R^p$. By applying a specific mapping function ϕ to the input space, the mapped data $R_p \rightarrow F, x \rightarrow X$ is obtained. Based on this, if $\sum_{i=1}^m \phi(x_i) = 0$ also meets the required conditions, the covariance matrix of the data can be derived:

$$C = \frac{1}{m} \sum_{i=1}^m \phi(x_i) \phi(x_i)^T \quad (1)$$

The eigenvalues λ and their corresponding eigenvectors ν can be solved respectively. Assuming the obtained eigenvalues satisfy $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ and each eigenvalue has a unique corresponding eigenvector $\nu_1, \nu_2, \dots, \nu_n$, as follows:

$$\nu_r = \sum_{i=1}^m \alpha_r^i \phi(x_i) \quad (2)$$

By assuming $K = (k_{ij})_{n \times n} = \phi(x_i) \cdot \phi(x_j) (i, j = 1, 2, \dots, n)$ and substituting both equations (1) and (2) into $\phi(x_i) \cdot C\nu = \lambda(\phi(x_i) \cdot \nu)$, $K\alpha = n\lambda\alpha$ can be obtained. In this equation, the kernel matrix and eigenvalue are K and $n\lambda_i$, respectively, and the eigenvector is $\alpha_1, \alpha_2, \dots, \alpha_n$. The projection of $\phi(x_i)$ onto ν_r can be seen in equation (3):

$$g_r(x_j) = \sum_{i=1}^m \alpha_r^i [\phi(x_i) \cdot \phi(x_j)] \quad (3)$$

In the majority of cases, $\sum_{i=1}^m \phi(x_i) = 0$ requires a certain transformation. However, due to its difficulty in satisfying the relevant conditions, it is assumed that:

$$\tilde{\phi}(x_i) = \phi(x_i) - \frac{1}{m} \sum_{i=1}^m \phi(x_i), i = 1, 2, \dots, m \quad (4)$$

Based on equation (4), it can be derived that:

$$\tilde{K}_{ij} = (\tilde{\phi}(x_i) \cdot \tilde{\phi}(x_j)) = K_{ij} - \frac{1}{m} \sum_{i=1}^m K_{ip} - \frac{1}{m} \sum_{i=1}^m K_{qi} + \frac{1}{m^2} \sum_{p,q=1}^m K_{pq} \quad (5)$$

Based on equation (3), in accordance with the diverse research scenarios, the significant principal components can be selected from the original data to replace the others for further analysis.

3.2. LightGBM Method

Light Gradient-Boosting Machine (LightGBM) is a modified and optimized version of the gradient-boosting decision tree algorithm, proposed by the Microsoft Asia Research Institute [6]. During its machine learning model training process, LightGBM creates multiple decision trees using a certain amount of sample data, resulting in outputs consistent with the combined decision tree outputs. The solutions offered by LightGBM typically include the boosting tree method and the random forest method which combines multiple decision trees. With the traditional gradient boosting decision tree (GBDT) algorithm, multiple traversals of the entire training data are required for each iteration, leading to memory limitations due to a large amount of training data. LightGBM reduces memory consumption and computational requirements through its histogram-based partition point search algorithm, effectively solving the problem of GBDT's high memory consumption when dealing with large data sets. The fundamental principle underlying LightGBM is congruent to that of XGBoost, however, LightGBM boasts a more expeditious training speed in comparison to XGBoost. Following the principle of the XGBoost model, $\hat{Y}^t$ is the result of gradient boosting model prediction with t decision trees integrated on the basis of labeling Y as the target result. With f_i serving as each newly grown decision tree model, the objective function of the gradient boosting algorithm is O [10]:

$$O = L(Y, \hat{Y}^t) + \varphi(F_t) = \sum_{i=1}^n l(y_i, \hat{y}_i^t) + \sum_{i=1}^t \varphi(f_i) \quad (6)$$

In equation (6), L, n represent the loss function and the number of test samples, respectively. The regularization term is: φ . The purpose of the regularization term here is to effectively reduce the risk of overfitting in this model.

The growth process of decision trees in this gradient-boosting algorithm model is as follows:

$$\hat{Y}^t = \sum_{k=1}^t F_k(X) = \hat{Y}^{t-1} + F_t \quad (7)$$

The loss function of the objective function can be obtained by Taylor expansion of $\hat{y}_i^{t-1}$, where C is a constant, which is:

$$O^t = \sum_{i=1}^m \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_i(x_i) + \frac{h_i f_i^2(x_i)}{2} \right] + \varphi(F_t) + C \quad (8)$$

In equation (8), g_i is defined as:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}} \quad (9)$$

In equation (8), h_i is defined as:

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{t-1})}{\partial (\hat{y}_i^{t-1})^2} \quad (10)$$

To simplify the objective equation, it is necessary to remove all constants in the objective function, yielding:

$$O_s^t = \sum_{i=1}^n \left[g_i f_i(x_i) + \frac{h_i f_i^2(x_i)}{2} \right] + \varphi(F_t) \quad (11)$$

Each growing decision tree benefits from the continuous optimization of the target equation within the gradient boosting framework in equation (6). These growing decision trees contribute to the classification of the entire model, rendering the prediction results increasingly precise and gradually approximating the classification labels. LightGBM boasts a competitive edge over XGBoost and GBDT in several ways: through its uni-side gradient sampling algorithm, it prioritizes high gradient samples when processing a substantial amount of information, thereby reducing the intensity of computation to a considerable extent; due to its leaf-wise growing strategy with a depth limit, it prevents unnecessary splits and searches, conserving computation time; it employs an efficient parallel mode, thereby enhancing hit rate. In summary, LightGBM holds immense potential for application in the fields of regression and prediction, owing to its efficiency, high accuracy, and various advantages.

4. Setting of evaluation metrics

The coefficient of determination (R^2) in machine learning is the ratio of the sum of squares regression to the sum of squares total. It indicates the consistency between the actual and model-predicted results and has a range of $[0, 1]$. The higher the value, the higher the degree of conformity between the model predicted and actual results, indicating a better fit and greater explanatory power of the dependent variable by the independent variable. Additionally, this also means that the distribution of the prediction points near the regression line is denser. As shown in formula (12):

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (12)$$

Root mean squared error (RMSE) is the square root of the ratio of the sum of the squared. Its result can be used to evaluate the degree of change of the data, with the range of $[0, +\infty)$. RMSE, as a method of measuring average error [11], the smaller the result, the higher the accuracy of the corresponding prediction model. The calculation process of RMSE is shown in equation (13):

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (13)$$

5. Empirical analysis

5.1. Data collection

This paper collects the data of relevant indicators from January 2005 to July 2022 from the CSMAR database, Choice Financial database and Wind database. Macro indicators include CPI, interbank 30-day lending rate, IPI, fixed asset investment completion, M2, US dollar against RMB exchange rate, and consumer confidence index; the stock market indicators include the Dow Jones Industrial Index, S & P 500, Hong Kong Hang Seng, Nikkei 225 and FTSE 100; financial indicators include the return rate and price-earnings ratio of the CSI 300 Index; the investor sentiment index is the turnover rate of the CSI 300 index. Among them, the completion of fixed asset investment, price-earnings ratio, turnover rate for individual missing value processing and quarterly GDP data for monthly interpolation. To check the obtained data, this paper calculates the descriptive statistics of each index, and the specific results are shown in Table 2.

Table 2. Descriptive statistics of indicators

Indicator/unit	maximum	Minimum	Mean	Median
X1/%	8.70	-1.80	2.49	2.10
X2/%	7.18	1.10	3.39	3.17
X3	114.05	49.85	101.48	103.75
X4/million yuan	635636.00	4221.78	188318.74	127614.14
X5 /million yuan	2581451.20	257708.47	1170964.29	1077379.16
X6/yuan	8.28	6.10	6.81	6.71
X7	127.00	86.70	106.17	104.20
X8/%	5.34	0.19	0.91	0.53
X9 /%	48.30	8.11	16.10	13.65
X10/yuan	381223.20	13376.97	122198.45	96898.27
X11/%	0.28	-0.26	0.01	0.01
X12	35848.57	7235.47	17726.61	15390.21
X13	4674.77	757.13	2032.47	1720.03
X14	31831.89	12794.95	22529.69	22663.95
X15	29893.57	7694.78	16615.60	16235.39
X16	7695.65	3760.23	6264.22	6299.43
Y	5632.61	853.14	3172.24	3247.23

5.2. KPCA dimensionality reduction

Based on SPSS software, KPCA is used to reduce the dimension of the independent variable and dependent variable data. The target dimension is set to 5 dimensions, and the RBF function is used as the kernel function to calculate the results. The specific results are shown in Table 3.

Table 3. KPCA partial dimension reduction results

KPCA1	KPCA2	KPCA3	KPCA4	KPCA5
-0.0545	-0.2234	-0.0500	-0.0801	0.1340
0.1132	0.0108	-0.0669	0.0164	-0.0292
-0.0992	0.0083	0.0924	-0.1473	-0.0458
-0.0787	-0.0007	-0.1766	-0.1314	0.3815
-0.0152	-0.2235	-0.0233	-0.0553	-0.0696
0.1676	-0.0673	-0.0235	0.0012	0.0208
-0.0043	0.0876	-0.0330	0.0613	-0.0092
0.2812	-0.1367	0.1693	0.1167	0.1978
0.0056	0.0078	-0.0202	-0.0090	-0.0023
-0.1090	-0.0671	-0.0388	-0.1034	-0.1412
-0.0603	-0.0029	-0.0258	0.1208	0.0246
0.4339	-0.1759	0.0074	0.0048	0.1917
-0.0008	0.0372	0.0364	-0.0347	-0.0562
0.0039	-0.1547	-0.0021	0.0660	-0.0043

To test the reliability of the data after dimensionality reduction, the Kendall consistency test was performed on KPCA1, KPCA2, KPCA3, KPCA4 and KPCA5. The test results are shown in Table 4.

Table 4. Analysis results of Kendall's W

Name	Average rank	Median	Kendall's W coefficient	X ²	P
kpca1	2.981	0	0.002	1.892	0.756
kpca2	3.071	0			
kpca3	2.943	0			
kpca4	2.919	0			
kpca5	3.085	0			

The Kendall consistency test was performed on the five indicators. The results showed that the significance p value of the overall data was 0.756, which was not significant at the level, and the null hypothesis could not be rejected. Therefore, the data could not be consistent. At the same time, the Kendall coordination coefficient w value was 0.002, which was extremely low. Consistency, so KPCA dimensionality reduction effect is better.

5.3. Light GBM model solution

Using SPSSPRO software, the data after 5.2 KPCA dimension reduction is brought in and solved by the Light GBM model. The GBDT is used as the base learner, and the ratio of the training set to the test set is maintained at 7 to 3. The 5-fold cross-validation is used, and the default parameters of the model are set: the learning rate is 0.1, the minimum number of leaves is 10, the sample rate and the tree feature sampling rate are 1, the single leaf minimum weight and the node splitting threshold are 0, the L1 regularization term is 0, and the L2 regularization term is 1. Adjust the number of model core parameter learners and the maximum tree depth. Among them, the number of learners is 50 steps and limited to 100-500. The evaluation results of multiple sets of Light GBM models with the different number of base learners and the maximum depth of the tree are shown in Table 5.

Table 5. Multiple Light GBM model evaluation results

Number of base learners	Maximum depth of the tree	RMSE	R2
100	10	0.001	0.822
150	10	0.001	0.707
200	10	0.001	0.777
250	10	0.001	0.837
300	10	0.001	0.833
350	10	0.001	0.754
400	10	0.001	0.823
450	10	0.001	0.78
500	10	0.001	0.775
100	9	0.001	0.753
100	8	0.001	0.773
100	7	0.001	0.819
100	6	0.001	0.775

5.4. Importance analysis

From the results of 5.3, the RMSE of the Light GBM model with 250 learners and the maximum tree depth of 10 is the same as that of other Light GBM models with different parameters, but R2 is slightly better than other models. Therefore, this set of parameters is selected as the optimal Light GBM model in this paper, and the contribution of the independent variables KPCA1, KPCA2, KPCA3, KPCA4 and KPCA5 to the dependent variable HS300 index is analyzed. By summing the contribution of each indicator on a single decision tree, the importance of the five indicators is 23.0 %, 21.2 %, 20.1 %, 19.9 % and 15.8 %, respectively. The histogram of importance is shown in Figure 1, and the comparison between the actual value and the simulated value is shown in Figure 2.

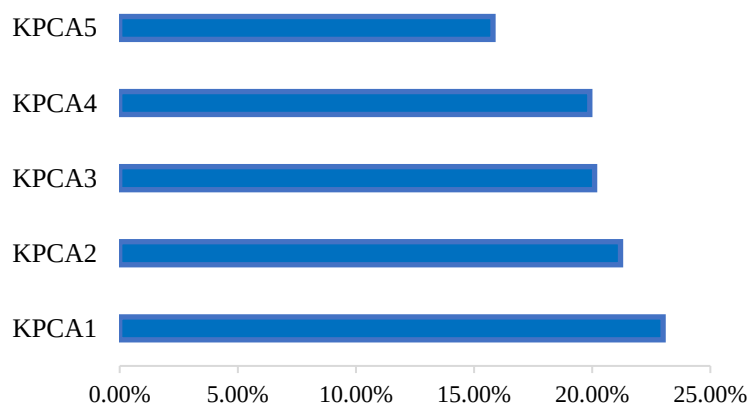


Figure 1. The histogram results of the importance analysis

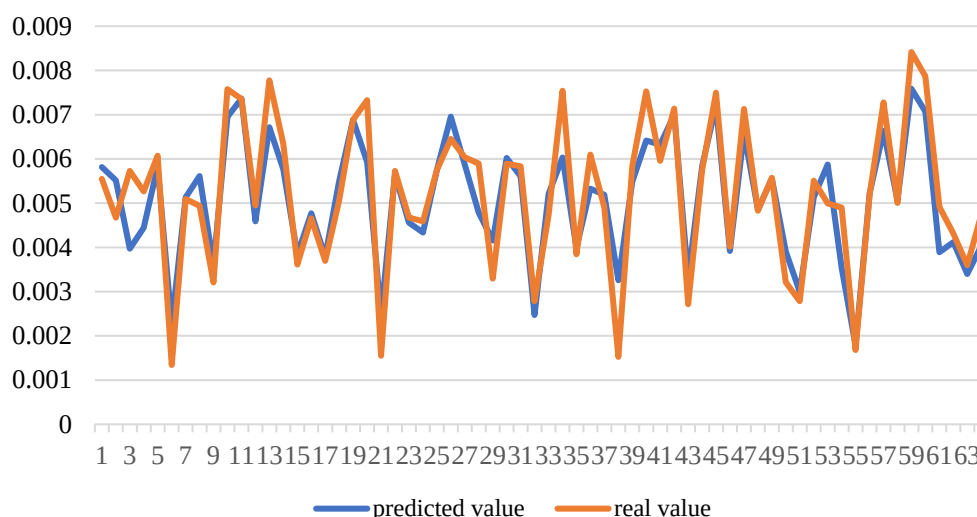


Figure 2. The line charts of simulated value and actual value

From the comparison diagram, it can be intuitively concluded that the KPCA-Light GBM model under this set of parameter settings has high accuracy. Most of the real value curve and the simulated value curve almost coincide, so the predicted value error is small, that is, CSI300 can be effectively predicted. On the other hand, the R2 of the KPCA-Light GBM model on the training set and the test set is close to 1, indicating that its generalization ability is strong.

6. Conclusion

In this study, data from 16 relevant indicators based on the CSI 300 stock market from January 2005 to July 2022 is collected, covering macroeconomics, investor sentiment, financial indicators, stock index futures market, etc. The KPCA-LightGBM algorithm is employed to lower the dimensionality of the impacting elements in the established index system, resolve the novel dataset, and conduct an importance analysis of the outcomes. The conclusion is as follows:

(1) Following the dimensionality reduction of the independent variables into five standalone indicators through KPCA, an analysis of the contribution of each of these variables to the dependent variable CSI 300 index was conducted. It was discovered that the importance of KPCA1 was the most pronounced; KPCA2, KPCA3, and KPCA4 displayed levels of importance that were commensurate and slightly lower than that of KPCA1; the importance of KPCA5 was comparatively low in comparison to the other four indicators. As a result, when making investment decisions, investors only need to consider the first four indicators that exhibit higher levels of significance.

(2) The prediction outcome indicated that when the parameters of the LG algorithm were configured to 250 learners and a maximum tree depth of 10. The algorithm demonstrated its highest accuracy, and the prediction values were remarkably aligned with the actual values, yielding an RMSE and R2 of 0.001 and 0.837, respectively, thereby facilitating the forecasting of the stock price trend of CSI 300.

References

- [1] Li Bin, Lin Yan, Tang Wenxuan. ML-TEA: A set of quantitative investment algorithms based on machine learning and technical analysis [J]. Systems engineering theory and practice, 2017, 37 (05): 1089 - 1100.
- [2] Pan Li, Xu Jianguo. Risk and characteristic factors of A-share market [J]. Financial Research, 2011, No.376 (10): 140 - 154.
- [3] Li Bin, Shao Xinyue, Li Yueyang. Quantitative investment research on fundamentals driven by machine learning [J]. China Industrial Economy, 2019, No.377 (08): 61 - 79.

- [4] Kozak, S., S. Nagel, and S. Santosh. Shrinking the Cross Section [EB / OL]. Journal of Financial Economics,
- [5] Kelly, B.T., S. Pruitt, and Y. Su. Characteristics Are Covariances: A Unified Model of Risk and Return [EB / OL]. Journal of Financial Economics,
- [6] Ma Xiaojun, Sha Jinglan, Niu Xueqi. Design and Application of P2P Project Credit Rating Model Based on LightGBM Algorithm. Quantitative Economics and Techno-Economic Research, 2018, 35 (05): 144 - 160.
- [7] Gao Ni, He Yiyue, Ma Xincheng. Index timing strategy based on low-frequency component EEMD-SVR prediction modeling [J]. Statistics and Decision, 2022, 38 (02): 140 - 145.
- [8] Xu Xuechen, Tian Kan. A new method of stock index prediction based on emotional analysis of financial text [J]. Quantitative Economic and Technological Economic Research, 2021, 38 (12): 124 - 145.
- [9] Huang Qiuli, Huang Zhuxing, Yang Yan. Empirical study on stock prediction based on recursive feature elimination and stacking integrated learning [J]. Journal of Nanning Normal University (Natural Science Edition), 2021, 38 (03): 37 - 43.
- [10] Ye Kang, Deng Xiaowei, Lu Xuebin. Research on stock market index prediction based on CEEMDAN-GRU model [J]. Business Research, 2021, 28 (03): 64 - 72+132
- [11] Xie Chengyan, Fang Hongbin, Guo Mengjie, Yang Mengzhuo. Research on the prediction of the CSI 300 index based on the inversion error combination optimization algorithm [J]. Journal of Fujian University of Business, 2020 (06): 14 - 22.