Analysis of Factors Affecting the CSI300 Index Based on KPCA and Various Machine Learning Algorithms

Ziyue Wang[1,#], Hongyue Chen[2,#], Zile Xu[3,*,#]

[1] Birmingham Business School, University of Birmingham, Birmingham, United Kingdom, B15 2TT

[2] International Business School, Guangzhou City University of Technology, Guangzhou, China, 510850

[3] Faculty of Mathematics and Statistics, Chongqing University, Chongqing, China, 401331

* Corresponding Author Email: lio1249586941@163.com

[#]These authors contributed equally.

Abstract. This paper utilises Kernel Principal Component Analysis (KPCA) and various machine learning algorithms to analyze the importance of factors affecting the Chinese Securities Index 300 (CSI300). Based on previous research, this paper constructs an indicator system consisting of 4 secondary and 21 tertiary indicators affecting the CSI300. The data is then reduced through KPCA and processed by various machine learning algorithms, including LightGBM, XGBoost, SVM, and Random Forest, to compare their predictive ability and feature importance. The results indicate that: (1) Under appropriate model parameter settings, the LightGBM model performs the best, while the other algorithms, such as the XGBoost, SVM, and Random Forest models, perform worse and with greater variability than the former. (2) This paper identifies the most significant indicator factors that affect the CSI300 index, such as closing price, price-to-book ratio, and turnover rate. Conversely, some factors, such as the buy-to-sell ratio, exhibit lower importance. These research findings have certain reference and guiding significance for improving the accuracy and reliability of stock market forecasting and practical and theoretical research in financial markets.

Keywords: Machine Learning, CSI300, KCPA, LightGBM, Importance Analysis.

1. Introduction

Compared with the traditional quantification based on experience, machine learning algorithms such as LightGBM, XGBoost, SVM and Random Forest are more and more widely used in quantification. At the same time, in order to improve the efficiency and accuracy of the algorithm, dimensionality reduction techniques such as PCA and KPCA are also widely used in stock market forecasting and analysis. At present, there are many researches on the analysis and prediction of machine learning algorithms in the stock market.

In the study of the application of machine learning in the stock market, Nazareth et al.[1] studied 126 articles in 44 famous journals in recent years in the fields of finance and machine learning through literature analysis, systematically expounded the application and current situation of machine learning in six fields such as stock market in finance, and pointed out the possible research direction. Kumbure et al. [2] further systematically reviewed 138 related journals in the field of stock market, and pointed out that the most commonly used stock market prediction model based on machine learning is based on artificial neural network, SVM and fuzzy theory, and the use of deep learning gradually increased after 2020.

In the study of the application of specific methods of machine learning, researchers began to improve and integrate machine learning algorithms, preferring to use hybrid models rather than single models. Zhang et al. [3] used PCA to reduce the dimension of multiple factors, and reduced the dimension from 200 to 100. The stock after dimension reduction was trained with LightGBM. The results show that the accuracy after dimension reduction will not be reduced. Zhang et al. [4] used four machine learning algorithms, such as support vector machine, gradient boosting regression, random forest and linear regression, to predict the rise and fall of stocks with stock fundamentals as

input variables. It is concluded that the multi-factor quantitative stock selection strategy has a good effect, and with the increase of the number of factors, the fitting degree under various algorithms is inversely proportional to the yield. Basak et al. [5] used two algorithms, random forest and XGBoost, to create a classification problem framework to predict the direction of stock price changes by using decision tree sets to strengthen connections. Zhu et al. [6] used XGBoost model and SVM model to predict the rise, fall and fluctuation of CSI 300, SSE 50 and CSI 500 stock index futures, respectively. Chen et al. [7] used PSO-SVR-GRNN non-parametric hybrid model to predict this problem. The particle swarm optimization algorithm is used to optimize the parameters of the support vector machine model to improve the prediction ability of the support vector machine regression model for the CSI 300.

Existing research has made a lot of contributions to stock forecasting and quantitative trading, but has not proposed a more general model. This study aims to use KPCA and LightGBM, XGBoost, SVM, Random Forest machine learning algorithms to build models and analyze the importance of indicators affecting stock indexes. By constructing an index system including 4 second-level indicators and 21 third-level indicators, the influencing factors affecting the CSI 300 index are measured. The KCPA algorithm is used to reduce the dimension of the index system and extract the principal component factors. Through algorithm training and parameter adjustment, MSE and R2 are used as accuracy indicators to select the model with the highest prediction accuracy and analyze the importance.

The contributions of this paper are as follows: (1) It is to compare the optimal machine learning algorithm in the field of stock market forecasting and analysis, improve the accuracy and efficiency of investment decision-making, and improve the risk management ability of portfolio; (2) It enriches the research in related fields and provides a basis for further improvement of related nonlinear models and machine learning algorithms in the future.

2. Construction of index system affecting SCI300

2.1. Financial indicators

The financial indicators reflect the company's financial situation. A good financial situation will accelerate the rise of stock prices. This paper chooses Return on equity (X_1), price-earnings ratio (X_2), price-to-book ratio (X_3), fixed asset turnover (X_4), dividend rate (X_5), operating income (X6), earnings per share (X_7), financial leverage coefficient (X_8) as financial indicators. The return on net assets is positively correlated with the stock price. The greater the earnings per share, the greater the rising space of the stock market price. The price-earnings ratio reflects the profitability of the company, and the price-to-book ratio reflects the present value of the company's assets. Stocks with low price-earnings ratio and low price-to-book ratio have strong risk-resisting ability, low risk and high investment value. The turnover rate of fixed assets measures the solvency of the company. The high turnover rate of fixed assets indicates the high utilization efficiency of fixed assets and the good management level of the company. The dividend rate is related to the dividend and stock price of the stock. The dividend rate is low and the company's profitability is low.

2.2. Macroeconomic indicators

The stock market reflects the changes in the economic boom, so the macro economy is also one of the factors affecting the stock market. Therefore, inflation rate (measured by CPI proxy, X_9), money supply (measured by M2 proxy, X_{10}), interest rate (X_{11}), exchange rate (X_{12}), GDP (X_{13}), consumer confidence index (X_{14}), entrepreneur confidence index (X_{15}) is selected as macroeconomic indicators. On the one hand, the macro economy is good, the continuous growth of GDP, the rise of consumer confidence index and entrepreneurial confidence index are conducive to the continuous rise of the stock market, and moderate inflation is also conducive to the rise of stock market prices. On the other hand, the money supply will affect the interest rate. The larger money supply will make the interest

rate decline. It is easier for people to obtain funds to make more funds flow to the stock market. The interest rate is related to the exchange rate. The exchange rate declines and the stock market rises.

2.3. Technical indicators

Technical indicators focus on indicators that reflect market behavior in the market to infer the trend of stock prices. This paper selects momentum (X_{16}), turnover rate (X_{17}) and volatility (X_{18}) as technical indicators. Momentum is related to stock trading volume. It measures the speed of stock price rise and fall, and believes that the rise and fall of stocks in the short term has inertia. The turnover rate measures whether the stock 's trading is active or not. If the turnover rate is high, the stock 's trading is more active and liquid. When the stock price is low, the turnover rate rises from a lower point, and the stock price may rise. Stock price fluctuation refers to the change of stock price. Stock price fluctuates greatly and it is possible to obtain higher returns.

2.4. Other indicators

In addition to the above indicators, this paper selects the growth rate of cash flow per share (X_{19}), free cash flow ratio (X_{20}), and circulation market value (X_{21}) as indicators. Cash flow affects the company's subsequent investment, operation and development. Free cash flow ratio plays an important role in the valuation strategy. High-yield stock cash flow performs well, but too much free cash is not conducive to the company's high returns. The market value of circulation changes in the same direction as the stock price. The larger the market value of circulation is, the more conducive it is to the stability of the investment market.

In summary, the index system constructed in this paper is shown in Table 1.

Table 1. CSI300 impact index system selection

Second-level Indicators	Third-level Indicators	Index of the Unit	Index Sign	Reference
Financial indicators	Return on Equity	%	X_1	[8]
	Price-Earnings Ratio	%	X_2	
	Price-to-Book Ratio	%	X_3	
	Fixed Asset Turnover	%	X_4	
	Dividend Rate	%	X_5	
	Operating Income	RMB	X_6	
	Earnings Per Share	RMB	X_7	
	Financial Leverage Coefficient	/	X_8	
Macroeconomic indicators	CPI	/	X_9	[9]
	M2	RMB	X_{10}	
	Interest Rate	%	X_{11}	
	Exchange Rate	%	X_{12}	
	GDP	RMB	X_{13}	
Technical indicators	Consumer Confidence Index	/	X_{14}	[10]
	Entrepreneur Confidence Index	/	X_{15}	
	Momentum	/	X_{16}	
	Turnover Rate	%	X_{17}	[11]
	Volatility	%	X_{18}	
Other indicators	the Growth Rate of Cash Flow Per Share	%	X_{19}	[12]
	Free Cash Flow Ratio	%	X_{20}	
	Circulation Market Value	RMB	X_{21}	

3. Kernel Principal Component Analysis (KPCA) Method

3.1. Introduction to KPCA

It has been demonstrated in recent research [13] that incorporating noise-processed data can lead to improved prediction results. Principal component analysis (PCA) is a popular method for reducing the dimensionality of data, assuming that all dataset features are linearly independent. However, in practice, the reduction from high to low dimension is often nonlinear, and linear mapping may not yield satisfactory results. To address this issue, Kernel Principal Component Analysis (KPCA) was proposed by Bernhard Schölkopf and Alexander J. Smola in 1998 as a technique for dimensionality reduction of nonlinear data. KPCA maps the data to a high-dimensional space using a kernel function and then identifies principal components in this space, ultimately reducing the data to the principal component space. The kernel function plays a critical role in the effectiveness of KPCA, and leveraging kernel function ideas can lead to better feature extraction performance [14]. One of the advantages of KPCA is its ability to handle nonlinear data while preserving the main characteristics of the data. Therefore, KPCA is a valuable tool for reducing the dimensionality of data in machine-learning applications.

3.2. Steps for Implementing KPCA

STEP1: Choose a kernel function:

$$\mathbf{K}\left(\mathbf{x},\mathbf{x}'\right)=\phi(\mathbf{x})^{\mathrm{T}}\phi\left(\mathbf{x}'\right) \tag{1}$$

STEP2: Compute the kernel matrix:

$$\mathbf{K}=\left[\mathbf{K}\left(\mathbf{x}_{\mathbf{i}},\mathbf{x}_{\mathbf{j}}\right)\right]_{\mathbf{i},\mathbf{j}=\mathbf{1}}^{\mathbf{n}} \tag{2}$$

STEP3: Centre the kernel matrix:

$$\mathrm{K_c}=\mathrm{K}-\frac{1}{\mathrm{n}}\left(1^{\mathrm{T}}\mathrm{K}+\mathrm{K}1-\frac{1}{\mathrm{n}}1^{\mathrm{T}}\mathrm{K}1\right) \tag{3}$$

STEP4: Compute the eigenpairs:
Eigenvalue problem:

$$\mathrm{K_c}\alpha=\lambda \tag{4}$$

Eigenvectors: α
Eigenvalues: λ
STEP5: Select the top k principal components:
Select the top k eigenvectors corresponding to the k largest eigenvalues: $\alpha_1,\alpha_2,\ldots,\alpha_{\mathrm{k}}$
STEP6: Project the data:
Project the data onto the k principal components:

$$\mathrm{z}=[\mathrm{z}_1,\mathrm{z}_2,\ldots,\mathrm{z}_{\mathrm{k}}]^{\mathrm{T}} \tag{5}$$

Where $\mathrm{z_i}=\left[\alpha_{\mathrm{i}}^{\mathrm{T}}\phi(\mathrm{x}_1),\alpha_{\mathrm{i}}^{\mathrm{T}}\phi(\mathrm{x}_2),\ldots,\alpha_{\mathrm{i}}^{\mathrm{T}}\phi(\mathrm{x}_{\mathrm{n}})\right]^{\mathrm{T}}$

4. LightGBM

4.1. Introduction to LightGBM

LightGBM, developed by Microsoft, is an efficient gradient boosting decision tree (GBDT) framework. It addresses the challenges of large datasets and high-dimensional features by utilizing a histogram-based decision tree algorithm. LightGBM is primarily used for binary and multi-class classification problems and performs exceptionally well in large and high-dimensional datasets [15, 16]. In terms of algorithm integration, we can consider LightGBM = XGBoost + Histogram + GOSS

+ EFB, where GOSS and EFB are its most significant innovations. The growth strategy of LightGBM decision trees is leaf-wise rather than level-wise (Figure 1), meaning it splits the leaf with the highest gain instead of splitting all nodes, thereby improving the algorithm's accuracy and efficiency.

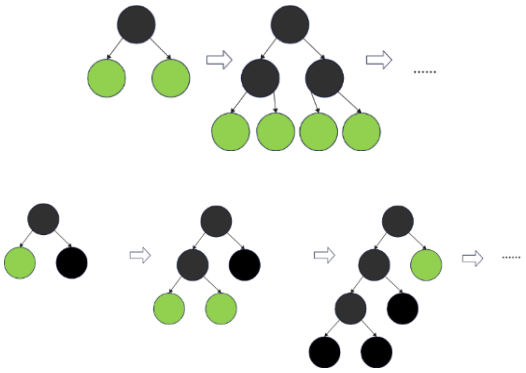

Figure 1. Schematic illustration of LightGBM leaf-wise node splitting

Firstly, building upon the foundation of XGBoost, LightGBM employs a histogram algorithm to construct decision trees. This algorithm divides the data into several small blocks and performs splits within each block, thereby reducing the number of searches and enhancing training efficiency, as illustrated in Figure 2.

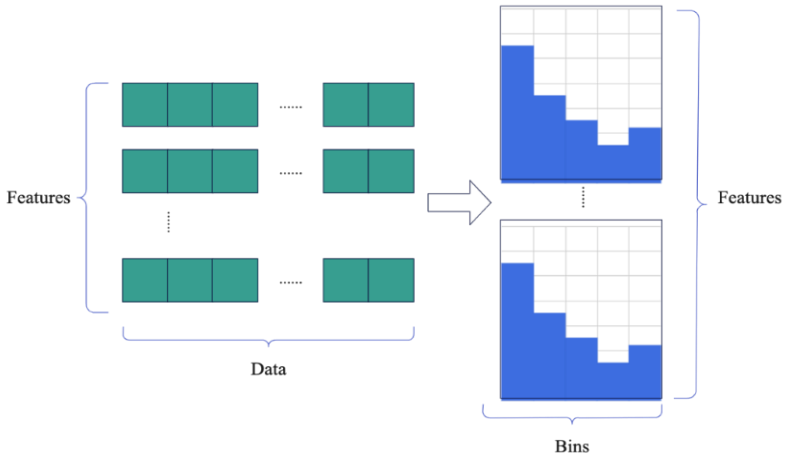

Figure 2. Schematic illustration of the histogram algorithm

In addition, the Gradient-based One-Side Sampling (GOSS) algorithm effectively retains samples with larger gradients while discarding those with smaller gradients during sampling, thus reducing the complexity of calculating objective function gains. The Exclusive Feature Bundling (EFB) algorithm reduces feature dimensions (the number of features when constructing histograms) by bundling mutually exclusive features together. Advantages of the LightGBM algorithm include fast speed (significantly faster training speed than XGBoost), low memory usage, excellent support for sparse features, and support for parallel learning [10].

4.2. Steps for implementing LightGBM

STEP1: Data Preparation: Prepare the training data as a matrix of features and a vector of corresponding labels.

STEP2: Gradient Calculation: The gradient is the negative derivative of the loss function with respect to the predicted values. For example, if we use MSE as the loss function, the gradient is given by:

$$\nabla_p \mathrm{MSE}(y, p) = -2(y_1 - p_i) \tag{6}$$

Where y_i is the real value and p_1 is the predicted value for year i.

STEP3: Hessian Calculation: The Hessian is the second derivative of the loss function with respect to the predicted values. In STEP2 case, the Hessian is given by:

$$\nabla_p^2 \mathrm{MSE}(y, p) = \frac{2}{n} I \tag{7}$$

Where I is the identity matrix. Note that if the gradient for the loss function, we choose is undifferentiable, alternative methods as gradient descent will be used to precede the algorithm.

STEP4: Leaf Value Calculation: The leaf value is the optimal output value for each leaf node in the tree. Denote gradient as g and Hessian as h, it is calculated as:

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{8}$$

Where $\mathrm{I_j}$ is the set of samples in the $\mathrm{j^{th}}$ leaf node, and λ is the $L2$ regularization term.

STEP5: Predicting: The predicted probability for a new sample is calculated by traversing the tree and summing the leaf values. The predicted probability is given by:

$$p = \frac{1}{1 + e^{-\sum_{j=1}^{T} w_j In\left(x \in R_j\right)}} \tag{9}$$

Where T is the number of leaf nodes in the tree, R_j is the region of the j^{th} leaf node, and $In(x \in R_j)$ is an indicator function that returns 1 if the sample x is in the region R_j and 0 otherwise.

STEP6: Model Training: LightGBM trains an ensemble of trees by iteratively adding new trees that fit the residual errors of the previous trees. The final prediction is the weighted sum of the predictions of all the trees. The training objective is typically defined as the sum of the loss function and a regularization term, which can be written as:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} L\left(y_i, \hat{y}i\right) + \sum_{j=1}^{T} T\Omega\left(w_j\right) \tag{10}$$

Where θ is the model parameters, L is the loss function, $\widehat{y_i}$ is the predicted value for the $\mathrm{i^{th}}$ sample, Ω is the regularization function, and $\mathrm{w_j}$ is the leaf value of the $\mathrm{j^{th}}$ leaf node in the $\mathrm{T^{th}}$ tree.

STEP7: Hyperparameter Tuning: LightGBM has many hyperparameters that can be tuned to improve the model performance, such as the number of trees, the learning rate, the maximum depth of the tree, and the regularization terms.

5. Test index setting

Different algorithms can be used to solve a problem, and the process results obtained by different algorithms are not the same. Therefore, it is necessary to have a certain standard to measure its accuracy and reliability. Regression analysis accounts for a large proportion of machine learning, so in a large number of machine learning studies, regression test indicators are usually used to evaluate performance. Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2) and Mean Absolute Percentage Error (MAPE) are usually used as test indicators. In this paper, MSE and R^2 are selected as the algorithm test indexes [17].

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t - \bar{y}_t\right)^2 \tag{11}$$

$$R^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \quad (12)$$

If there are outliers to be detected, MSE can be used, and the square part of the function will amplify the error. Therefore, MSE measures the error of the model and directly reflects the gap between the predicted value and the true value. The smaller the predicted error, the better. Its value range is [0, + ∞). When MSE is 0, it is completely similar, and the value greater than 1 indicates that the similarity is small. R^2 reflects the fitting accuracy of the model by the ratio of the predicted value to the real value. The range is [0, 1]. The closer to 1, the better the model fitting effect. When R^2 is negative, the model performs poorly.

6. Empirical analysis

6.1. Data collection and display

In order to improve the accuracy of stock forecasting, this paper selects the relevant data indicators of CSI300, including the return on net assets (X_1), price-earnings ratio (X_2), price-to-book ratio (X_3), fixed asset turnover (X_4), dividend rate (X_5), operating income (X_6), earnings per share (X_7), financial leverage coefficient (X_8) ; macroeconomic indicators CPI (X_9), M2 (X_{10}), interest rate (X_{11}), exchange rate (X_{12}), GDP (X_{13}), consumer confidence index (X_{14}), entrepreneur confidence index (X_{15}) ; as well as the corresponding technical indicators of the Shanghai and Shenzhen 300 momentum (X_{16}), turnover rate (X_{17}), volatility (X_{18}), cash flow per share growth rate (X_{19}), free cash flow ratio (X_{20}), circulation market value (X_{21}) and other 21 indicators as characteristic indicators to study the closing price of the CSI300. The specific indicators are shown in Table 1.

This paper uses the Wind database to collect data on various indicators for 15 years from 2007 to 2022, with the CSI300 closing price as the explained variable and the remaining indicators as the explanatory variables. The data were standardized descriptive statistics, the results are shown in table 2.

Table 2. Data descriptive statistics

Variable Name / Unit	Sample Size	Maximum Value	Minimum Value	Mean Value	Standard Deviation	Median	Variance
X_1/%	15	5406.81	1463.68	3253.815	1058.261	3066.05	1119916.376
X_2	15	33.56	9.9	17.287	6.434	15.18	41.398
X_3	15	5.13	1.78	2.543	0.782	2.32	0.611
X_4	15	0.64	0.57	0.613	0.019	0.61	0
X_5/%	15	3.17	1.6	2.513	0.471	2.54	0.222
X_6/ trillion	15	142198.72	30919.19	76493.236	36294.203	72735.45	1317269178.472
X_7/yuan	15	2.66	0.94	1.828	0.59	1.8	0.348
X_8	15	2.8	2.03	2.348	0.226	2.37	0.051
X_9	15	105	99.1	101.68	1.326	101.9	1.759
X_{10}/%	15	15	1	8	4.472	8	20
X_{11}/%	15	0.07	0.043	0.057	0.009	0.058	0
X_{12}/%	15	7.295	6.155	6.64	0.341	6.617	0.116
X_{13}/ trillion	15	14	1	7.867	4.274	8	18.267
X_{14}/%	15	125.7	94.1	112.84	9.438	116.4	89.077
X_{15}	15	107.9	89.5	100.26	6.483	99.6	42.025
X_{16}/ trillion	15	92.51	-67.71	14.431	40.89	11.47	1672.012
X_{17}/%	15	61	31.96	45.53	8.924	45.09	79.646
X_{18}	15	59.96	12.79	27.762	12.746	25.21	162.467
X_{19}/ trillion	15	0.486	0.163	0.332	0.098	0.336	0.01
X_{20}	15	0.254	-0.039	0.091	0.063	0.08	0.004
X_{21}	15	15	1	8	4.472	8	20

6.2. KPCA dimension reduction processing

After data pre-processing, this paper chooses to use SPSS for KPCA feature extraction. Here, several key parameters are involved, and the setting of these parameters will affect the accuracy of the feature extraction process.

(1) The number of principal components extracted. After comparing the results of extracting different numbers of principal components, six principal components were selected to be extracted in this paper. The algorithm automatically selects two eigenvalues based on the eigenvalue from largest to smallest, calculates the eigenvector and returns the six principal components directly. (b) Selection of kernel function

(2) The kernel parameters determine the choice of the kernel function in kernel principal component analysis. As Gaussian kernel has good high-dimensional generalization ability and has been recognized by many scholars, after a comprehensive comparison of linear kernel, polynomial kernel and Gaussian kernel, Gaussian kernel function $k(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2})$ is selected for kernel principal component analysis in this paper.

In this paper, a KPCA with a Gaussian kernel as the kernel function was used to simplify the 21 indicators previously collected into six dimensions, and the resulting factor correlation plot was obtained as shown in Figure 3.

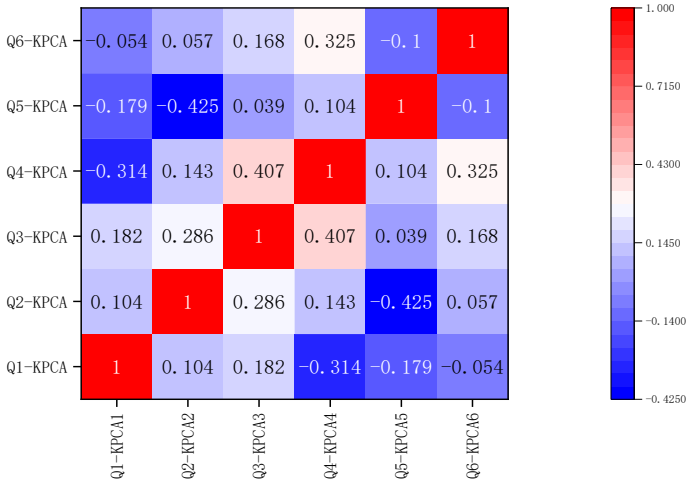

Figure 3. Factor correlation graph

Figure 3 depicts the correlations between the factors. From the results, it can be found that the correlation between Q2-KPCA and Q5-KPCA is slightly stronger than that between the other factors, but the absolute value of the correlation coefficient between these two factors is still less than 0.5, indicating that the correlation between them is still very weak, so the correlation between these six factors is very small and the factors are independent of each other, and the selection of the factors The effect is very good.

6.3. LightGBM Algorithm Learning

Based on the relevant programming software, the LightGBM algorithm is called, and the six factors are used as independent variables. The closing price of the Shanghai and Shenzhen 300 is used as the dependent variable, and the number of trees is adjusted with a step size of 100. From 100 to 1000, 70 % of the samples were selected as the training set, and 30 % of the samples were selected as the test set, and L2 was used for regularization. The prediction accuracy of each time is shown in Table 3 and Figure 4.

Table 3. Model accuracy demonstration

Number	R^2	MAE
100	0.985	4.242
200	0.925	0.025
300	0.855	0.001
400	0.885	0.001
500	0.99	0.001
600	0.92	0.001
700	0.98	0.001
800	0.97	0.001
900	0.935	0.001
1000	0.945	0.001

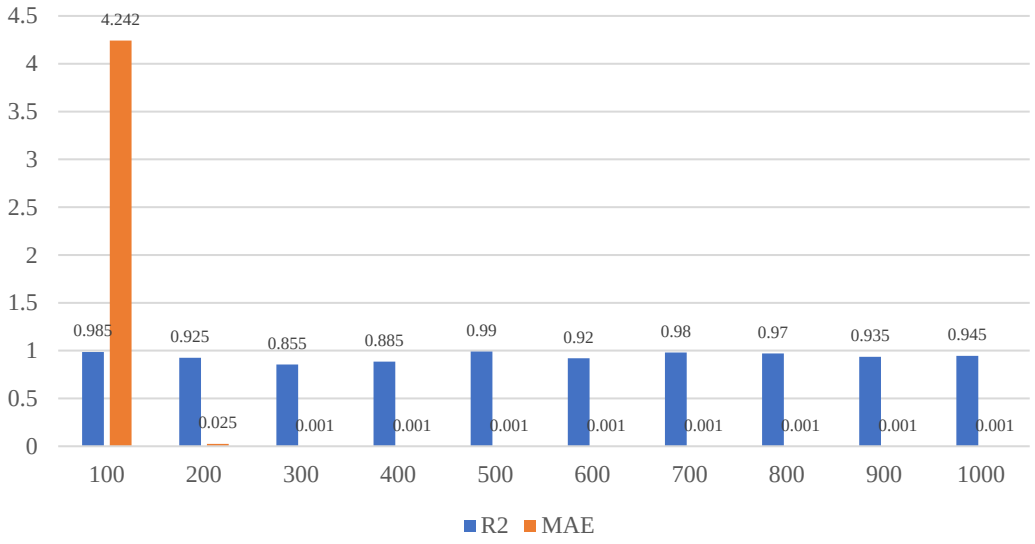

Figure 4. Model accuracy demonstration

The model accuracy reaches optimal convergence when the number of trees reaches 500. At this point the parameter values are as follows: R^2 = 0.99 and MAE = 0.001.

Based on the LightGBM algorithm, this paper also studied and analyzed the prediction accuracy of other algorithms, including four algorithms, namely, Random Forest Regression, Adaboost, XGBoost and ExtraTrees, still from 100-1000, adjusting the number of trees in steps of 100, selecting 70% of the samples as the training set and 30% as the test set, and regularizing them with L2. The MAE and R^2 for each prediction are shown in Tables 4 and 5.

Table 4. MAE metrics of other algorithms

Number of trees	Random Forest	Adaboost	XGBoost	Extra Trees
100	980.098	901.982	776.496	981.385
200	933.366	691.562	663.177	826.399
300	665.853	572.454	647.83	873.765
400	1337.495	492.444	469.078	1158.698
500	980.687	1216.18	1605.801	880.831
600	608.811	956.286	1158.406	862.206
700	1387.669	1766.678	1056.631	941.536
800	1102.176	992.912	1422.711	692.881
900	1115.154	860.92	731.508	1499.92
1000	894.125	1388.318	806.859	1018.899

Table 5. R^2 metrics for other algorithms

Number of trees	Random Forest	Adaboost	XGBoost	Extra Trees
100	-0.351	-1.953	-0.035	-1.887
200	0.01	0.215	0.102	-0.219
300	0.484	0.633	0.566	-1.023
400	-2.885	0.611	0.53	-1.719
500	-1.161	-1.019	-5.248	0.024
600	0.352	0.071	-0.589	-1.421
700	-3.901	-6.224	-0.82	-0.371
800	-1.422	0.26	-4.475	0.116
900	-0.967	-1.432	0.053	-0.573
1000	-1.523	-5.306	0.17	-0.425

Based on the fitting effect above, this paper concludes that the KPCA-LightGBM algorithm has the highest fitting accuracy for the prediction of CSI 300 and can basically meet the prediction requirements. The fitting effect and forecasting effect of other algorithms including Random Forest Regression, Adaboost, XGBoost and ExtraTrees are obviously not as good as the LightGBM algorithm, and there is a large gap.

6.4. Importance analysis

According to the above, by comparing the MAE and R^2 indicators of the LightGBM algorithm with the other four algorithms, the analysis shows that LightGBM has good accuracy in fitting and predicting the closing price of the CSI 300, so the LightGBM model can be used to predict the future trend of the closing price of the CSI 300. At this point, based on the results of 6.3, we analyse the characteristic importance of the group with the highest accuracy and obtain the importance of each variable to the CSI 300 as shown in Figure 5.

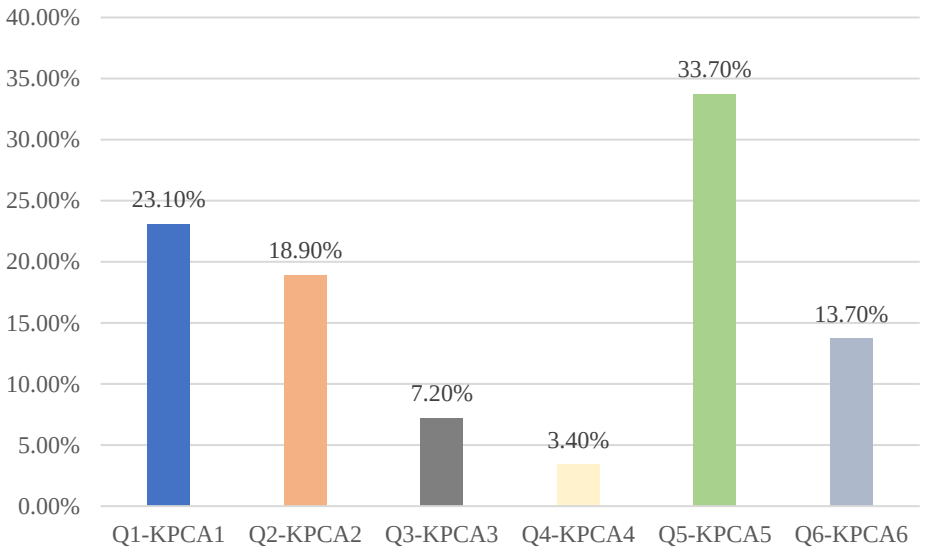

Figure 5. Importance of the factors

The analysis of the model results shows that the six factors are ranked in descending order of importance: Q5-KPCA, Q1-KPCA, Q2-KPCA, Q6-KPCA, Q3-KPCA and Q4-KPCA. The importance of Q4-KPCA4 is the weakest factor, with only 3.40%.

Meanwhile, Q1-KPCA1, Q2-KPCA2 and Q6-KPCA6 also have good explanatory power with importance ratios of over 10%, 23.10%, 18.90% and 13.70% respectively. Q3-KPCA3, on the other hand, has an importance share of only 7.20%, not much different from the variable with the weakest

explanatory power, Q4-KPCA4, and the sum of the importance of the two factors is only 10.60%, less than a third of Q5-KPCA.

7. Conclusion

This study analyzed the importance of indicators affecting stock market indices by applying KPCA and various machine learning algorithms. Based on existing literature and research, we constructed an indicator system consisting of 4 secondary and 21 tertiary indicators to measure the factors influencing the CSI300 index.

In this paper, we used KPCA algorithm to reduce the dimensionality of the indicator data and extract principal component factors. After training and parameter tuning using algorithms such as LightGBM, XGBoost, SVM, and Random Forest, we selected the model with the highest prediction accuracy to conduct feature importance analysis using R^2 and MAE as accuracy standards.

The results showed that the KPCA-LightGBM algorithm can efficiently reduce dimensionality, predict stock market data, and accurately analyze the main factors affecting the CSI300 index. Among them, we found that some important factors have a significant impact on the index, such as Q5-KPCA5, which is the most explanatory factor among the six factors, with an important ratio of 33.70%; Q4-KPCA4 is the weakest explanatory factor, with importance ratio of only 3.40%.

Furthermore, we also found that macroeconomic conditions, policy regulation, and other industry factors have a significant impact on the CSI300 index. These research findings have certain reference and guiding significance for improving the accuracy and reliability of stock market forecasting, as well as for practical and theoretical research in financial markets. Additionally, our research methods provide a reference for data analysis and prediction in other fields.

References

[1] Nazareth N, Reddy Y Y R. Financial applications of machine learning: a literature review[J]. Expert Systems with Applications, 2023: 119640.

[2] Kumbure M M, Lohrmann C, Luukka P, et al. Machine learning techniques and data for stock market forecasting: A literature review [J]. Expert Systems with Applications, 2022: 116659.

[3] Zhang N, Gao C, Xiao M. LightGBM stock forecasting model based on PCA[C]//2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). IEEE, 2021: 396-399.

[4] Zhang C, Tang H. Empirical Research on Multifactor Quantitative Stock Selection Strategy Based on Machine Learning[C]//2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML). IEEE, 2022: 380-383

[5] Basak S, Kar S, Saha S, et al. Predicting the direction of stock market prices using tree-based classifiers[J]. The North American Journal of Economics and Finance, 2019, 47: 552-567.

[6] Zhu H, Zhu A. Application Research of the XGBoost-SVM Combination Model in Quantitative Investment Strategy[C]//2022 8th International Conference on Systems and Informatics (ICSAI). IEEE, 2022: 1-7.

[7] Chen J, Yang H. A CSI 300 Index Prediction Model Based on PSO-SVR-GRNN Hybrid Method [J]. Mobile Information Systems, 2022, 2022.

[8] Li Z, Xu W, Li A. Research on multi factor stock selection model based on LightGBM and Bayesian Optimization [J]. Procedia Computer Science, 2022, 214: 1234-1240.

[9] Liu Yuxuan, Jin Weize, Yuan Liang. Multi-factor quantitative stock selection model optimization and empirical research-analysis of introducing financial cycle indicators [J]. Price theory and practice, 2022, No.454 (04): 141-145. DOI: 10.19851/j.cnki.cn11-1010/f.2022.04.172.

[10] Wang, Difei & Huang, Yi & Ren, Chengxin. (2022). Prediction and contribution analysis of CSI 300 Stock Index Futures based on KPCA-LightGBM. BCP Business & Management. 28. 1-11. 10.54691/ bcpbm.v28i.2135.

[11] Zhou Liang. Research on stock multi-factor investment based on random forest model [J]. Financial Theory and Practice, 2021, No.504 (07) : 97-103.

[12] Yan W L. Stock index futures price prediction using feature selection and deep learning [J]. The North American Journal of Economics and Finance, 2023, 64: 101867.

[13] Meng Xiang Jun. (2019). Dupont analysis on the impact of financial indicators on stock prices of CSI 300 companies (Master's thesis, University of International Business and Economics).

[14] (2018). Machine Learning; New Findings from Beijing Institute of Technology Update Understanding of Machine Learning (Fault detection and fault-tolerant control over signal-to-noise ratio constrained channels). Journal of Robotics & Machine Learning.

[15] Sun, Xiaolei, Mingxi Liu, and Zeqian Sima. A Novel Cryptocurrency Price Trend Forecasting Model Based on LightGBM. Finance Research Letters, no.32 (January 2020): 101084. https://doi.org/10.1016/j.frl.2018.12.032.

[16] Tian, Liwei, Li Feng, Lei Yang, and Yuankai Guo. Stock Price Prediction Based on LSTM and LightGBM Hybrid Model. The Journal of Supercomputing 78, no.9 (1 June 2022): 11768–93. https://doi.org/10.1007/s11227-022-04326-5.

[17] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation [J]. PeerJ Computer Science, 2021, 7: e623.