Gold futures price forecast research ----A combinatorial prediction method based on CEEMDAN-GARCH-SVR

Simin Chen*

Institute of economics, Jinan university, Guangzhou, Guangdong, 510632

*Corresponding author:19867814050@163.com

Abstract. In recent years, the international form has become more and more complex, and various emergencies have constantly impacted the world economy. In this context, countries in the world need to maintain the stability of the domestic economic market and reasonably avoid financial risks. Gold plays a major role for both countries and individuals, and predicting the gold price is a prerequisite for making important decisions.In this paper, with the closing price of gold futures AU9999 from January 2,2008 to February 13,2023 as the research object. This paper mainly uses the traditional time series model for modeling, combined with the machine learning method, and introduces the decomposition reconstruction algorithm, hoping to achieve a better prediction effect. The traditional time-series model-error t-based ARIMA (2,1,2) -EGARCH (1,1) was established, and then the machine learning method SVR was used to model the data. In order to improve the performance of the model, this paper uses the CEEMDAN method to decompose the price sequence, and then decompose the IMF according to the high frequency sequence judgment condition summary into high frequency, low frequency and remaining term three sequence, make the original complicated sequence relatively simple, then with the traditional time series model for high frequency and low frequency modeling, SVR for the remaining term. Finally, we found that the decomposition modeling algorithm proposed here has the best prediction effect and made reasonable suggestions according to the prediction results.

Keywords: Predicting the Gold Price, Traditional Time Series Model, Decomposition Reconstruction Algorithm, CEEMDAN.

1. Introduction

Gold is a financial investment variety with a long history, and its price is affected by the global economic conditions, political situation, financial markets and other factors. For consumers, the jewelry made of gold has unique charm, so the rise and fall of gold price also means the rise and fall of gold jewelry; for investors, the fluctuation of gold price will increase or decrease of their assets; for countries, gold as an important financial product can effectively maintain financial stability. Therefore, predict the price of gold can help consumers more affordable buy gold jewelry, can help investors make better decisions when investment gold, and get the best return on investment, can also provide predictions for the government, so that they can make effective decisions according to the future gold price trend, maintain financial stability. Accurate gold price prediction is the basis of planning by using the prediction results. In order to predict the gold price more accurately, this paper studies the various properties of the gold price data, and compares a variety of gold price prediction methods, and in-depth study of the gold price prediction.

The innovation of this paper is: first, using decomposition algorithm CEEMDAN to the original sequence decomposition, signal signal decomposition algorithm fusion of traditional time series and machine learning algorithm applied in the field of economy, especially in the field of gold futures, the "decomposition-reconstruction-integration-prediction" method is few research in the field of gold futures. Second, in the reconstruction process, the high frequency sequence judgment condition uses the t-test to test whether the mean of the sequence is zero, but also adds a variable percentage condition, which can make the fluctuation of the high frequency sequence more obvious, which is more conducive to the fitting of the model.

2. Literature review

2.1. Traditional time-series model

In essence, gold futures price data is a kind of financial time series data. For financial time series data, many scholars at home and abroad have invested a lot of time and energy to study, which makes the research in this field has a long history and deep precipitation. Ding lei [1] using on January 2,2018 to 2018 gold from December 28 (T + D) closing price data, using different error distribution of ARIMA-GARCH family modeling to predict the gold price, found that ARIMA-GARCH model prediction effect is significantly better than ARIMA model, and the error distribution of GED distribution can depict data error, finally found that the gold futures price fluctuation has obvious TGARCH effect. Sun Hao [2] uses fTGARCH model to analyze and predict the volatility of financial time series, so as to provide investors with more comprehensive information on the degree of asymmetry of data and model, as well as the degree of influence of current disturbance on future volatility.

2.2. machine learning model

Machine learning methods have been widely used in various fields, including finance. The commonly used machine learning models are support vector machine, random forest, naive Bayes model, and XGBoost[3]. Xie Qi [4] proposed a feature selection algorithm based on the improved forest optimization algorithm, and the experiments showed that the new feature selection algorithm is better than the other algorithms. The regression prediction model based on improved SVM [5] is also a good prediction method. Using the search optimization process, the grid search algorithm is relatively simple, with high search accuracy, which has certain advantages in parameter optimization.

2.3. Neural network model

In the development of deep learning, people have found the RNN model with mining time data dependence. Literature [6] uses the LSTM model to predict the gold price, and obtains good prediction results. Zhang et al. used deep belief network (DBN) [7] to predict gold prices from 1984 to 2019, and compared the results with the traditional BP neural network model, genetic algorithm optimized BP neural network model (GA-BP), and autoregressive integrated moving average mathematical linear model (ARIMA). The empirical results show that the proposed DBN model has good predictive performance and direction, with the lowest root mean square error (RMSE), average absolute percentage error (MAPE) and mean absolute error (MAE), and the highest directional statistic (Dstat). Jiang Chaoyu [8] input 20 feature variables into single Transformer and gold futures price prediction model based on CNN-Transformer combination model, and finally verified the superiority of the proposed CNN-Transformer model in the experiment.

2.4. Combined prediction model

The combined prediction model, that is, is modelusing multiple methods, such as Liang [9] predicts the gold price using ICEEMDAN-LSTM-CNN-CBAM combined with empirical decomposition method and neural network, which proves the superiority of the method through experiments. He Linyun [10] uses the ICEEMDAN algorithm to decompose the closing price of gold futures from January 4,2010 to December 13,2021, Then we calculate the sample entropy of the IMFS, Reconstituted and high, Medium, and the low-frequency sequence data, The three sequences were then modeling and predicted separately using the SSA-ELM method, Finally, the nonlinear integration of the prediction results of the three sequences, To obtain the final predicted value, The proposed ICEEMDAN-SE-ELM-ELM model under the four evaluation indexes than ELM, SSA-ELM, The ICEEMDAN-SSA-ELM model achieved better results. Yuan Dongfang [11] used CEEMDAN method to decompose the gold futures price, and then used the method of principal component analysis to reconstruct the resulting IMFs and residual sequence. Finally, the LSTM

model was established for the reconstructed features, and the superiority of the model was verified in experiments.

2.5. Research Trends

In recent years, at home and abroad about economic classes such as gold time series data prediction research method deepening and complicated, most of the researchers use deep learning for time series data prediction, but the established model mostly lack of interpretability, the reasons behind the mining data changes, and more and more complex model can achieve the prediction effect of promotion is smaller, even than the traditional model. The literature: Do We Really Need Deep Learning Models for Time Series Forecasting? The [12] points out that the reasonable data processing and simple model can also achieve good effect, even can be more than complex deep learning model, not to pursue the model of large and complicated to the time series data prediction, so the regression simple model method, through reasonable data processing and analysis to the gold futures price forecast.

3. Establishment and analysis of individual models

3.1. Data description

This paper uses the closing price of AU9999 from January 2,2008 to February 2023,2023 as the data for this study. With a total of 3657 data, the last 100 data were taken as test samples and the first 3557 samples as training samples. The price trend of the entire data is shown in the Figure 1. From 2009 to 2011, it can be seen that the price of AU9999 was on an upward trend, while from 2011 to 2019, it fluctuated significantly and the overall trend was downward. After 2019, the amplitude and fluctuation increased, and the overall trend was upward.Table 1 is the full sample data, the descriptive statistics of the sample data of the training set and the test set, and it can be seen from the table that the data do not obey a normal distribution.The whole sample data and the training sample data showed a significant right bias, while the test sample showed a slight right bias, and the kurtosis of all data samples was negative, indicating that the sample distribution was flatter than the normal distribution.

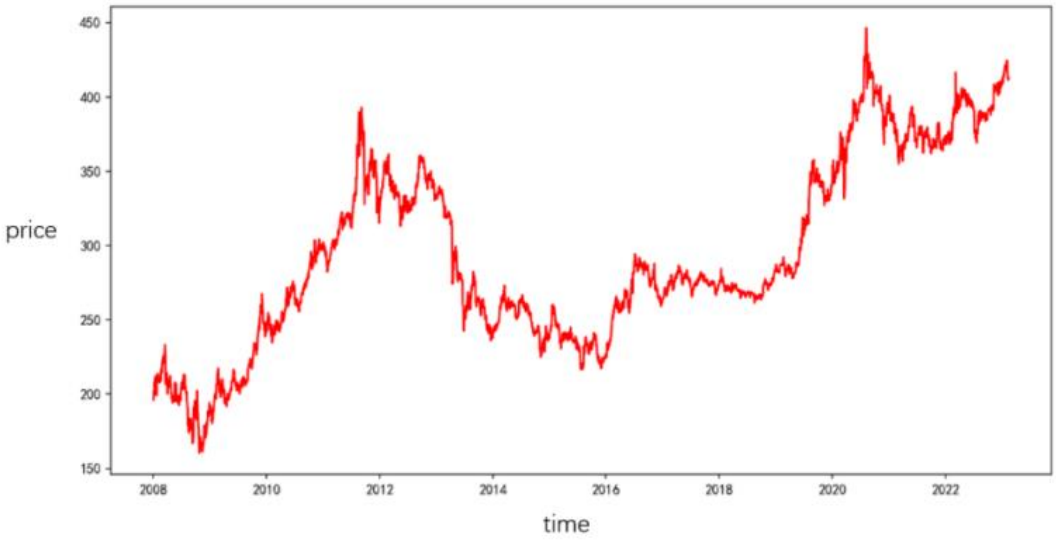

Figure 1. Fluctuation curve of raw data

Table 1. Descriptive statistics of raw data

	number	mean	variance	range	skewness	kurtosis
full sample	3567	293.08	3834.88	285.98	0.26	-0.86
training sample	3557	290.03	3597.91	285.98	0.28	-0.79
test sample	100	401.73	126.19	40.90	0.012	-1.21

In order to avoid the impact of large numerical differences on the fitting of the model, this paper adopts the log transformation method commonly used in most economic data to reduce the extreme value of the data, that is, the log transformation of each price Pt.

3.2. Traditional time-series modeling

3.2.1ARIMA Model building

The data were tested for stationarity using the unit root test. The sequences are found to be non-stationary, requiring differential sequences and stationary after first order difference, so they are all modeled later in this section. For the ARMA model, there exist two parameters p and q, and the AIC criterion is used to determine the order of the traditional time series models. By calculation, the ARMA (2,2) model has the minimum AIC value, so the ARMA (2,2) model is selected for modeling, and the coefficient is estimated using the OLS method.The fit of the prediction result is shown in Figure 2.

Figure 2. ARIMA (2, 1, 2) real value and predicted value fluctuation curve

The model passes a white noise test and considers no other useful information in the residuals. The LM test was used to verify the sequence heteroquasticity, and the results showed that the sequence has autocorrelation conditional heteroququasticity, so the GARCH model is needed to differential fit the data information. Considering whether the model has a leverage effect, compare the model (GARCH, EGARCH). Also make assumptions and compare the errors (normal distribution, T distribution and GED distribution). While the GARCH and EGARCH models, the most commonly used GARCH (1,1) and EGARCH (1,1).

The results show that the gold price has a significant leverage effect, and the bad news has a greater impact on the volatility of gold futures prices than the good news. When there is bad news in the market, the market and earnings are negative, and the volatility rises sharply. Since the EGARCH model AIC based on the t distribution assumption is used, the ARMA (2,2) -EGARCH (1,1) model with t distribution is used for analysis and prediction in the future. The parameters of the model were mostly significant and the residual passed the white noise test as well as no ARCH effect.

3.2.2 Prediction results

We established the ARMA (2,2) -EGARCH (1,1) model for the difference sequence {yt} of the logarithmic form of the original data, and estimated the parameters of the training set as follows:

$$y_t = -1.13092 y_{t-1} - 0.247612 y_{t-2} + 1.084411 \varepsilon_{t-1} + 0.201981 \varepsilon_{t-2} + \varepsilon_t \quad (1)$$

$$\varepsilon_t = \sigma_t * e_t \quad (2)$$

$$\sigma_t^2 = 0.014318\varepsilon_t^2 + 0.980092\sigma_{t-1}^2 \tag{3}$$

In the formula,e_t obey the t-distribution.

Using the one-step forecast to predict the gold futures price in the next 100 days, the results are shown in Figure 3, which can be seen that the model can make a better one-step prediction. The predicted results with the true value of MAE, MSE is shown in the Table 2.

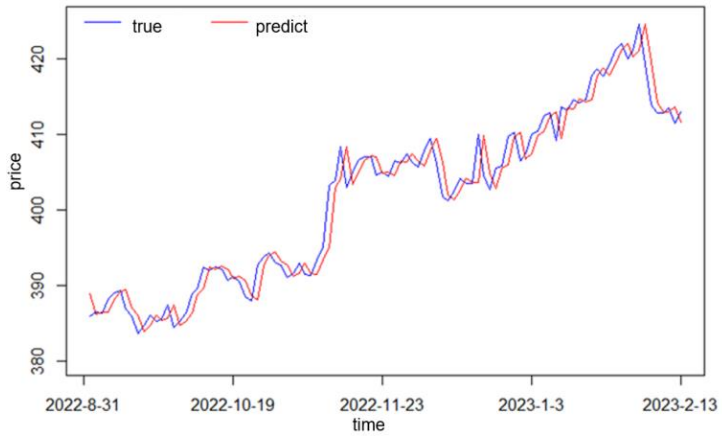

Figure 3. ARMA (2,2) - EGARCH (1,1) - t model real value and predicted value fluctuation curve

Table 2. ARMA (2,2) - EGARCH (1,1) - t model evaluation indicators

Index	MAE	MSE
Value	1.7246	5.4733

3.3. Establishment of the SVR model

3.3.1 SVR modeling

The SVR model in this paper uses the RBF kernel function, and for modeling the SVR using the RBF kernel function needs to determine 3 hyperparameters, one is the size of the time window,the penalty factor C and the kernel parametersγ in the SVRSVR.In this paper, the top 80% of the data in the training set are taken as the training set, and the bottom 20% of the data are taken as the verification set. We find the hyperparameter combination with the highest accuracy in the validation set. Through the experiment, the hyperparameter is determined to be, and the time window size is 30, C=100,γ=5e-5.

3.3.2 SVR prediction results

The one-step prediction results and effects for the test set are shown as follows in Figure 4 and Table 3 respectively.

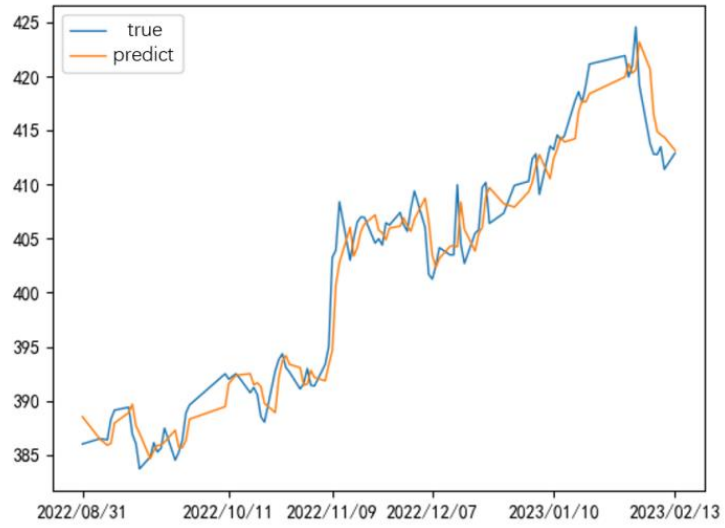

Figure 4. Fluctuation Curve of SVR Model Real Value and Predicted Value Test Set

Table 3. SVR model scoring indicators

Index	MAE	MSE
Value	1.8533	5.7519

4. Establishment and analysis of the model after data decomposition

The original sequence is more difficult to directly model the original sequence due to its non-stationarity, nonlinearity and the presence of noise, which adds difficulty to the learning of the model. To get better results using relatively simple models requires processing the data and simplifying the complex data. This chapter uses the CEEMDAN decomposition method to decompose the original sequence data, which decomposes the complex data into multiple frequency-more concentrated and relatively simple sequences, and improves the accuracy of the final prediction results by modeling the multiple sequences obtained by the decomposition separately.

4.1. Data decomposition and reconstruction

The CEEMDAN module using the pymed package of python was set to 100 added noise, noise standard deviation of 0.1, and noise SNR of 0.5. The decomposition sequence obtained from the decomposition of the training set is shown in Figure 5. It can be seen that the original sequence is divided into 9 IMF and 1 remaining term. The frequency of 9 IMF is gradually reduced from high to low, and the change form is more and more simple, while the remaining term presents a monotonic rise of S-type curve.

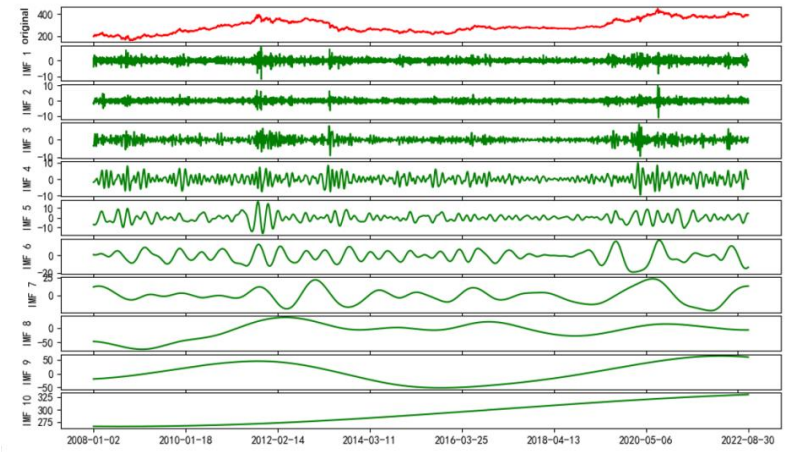

Figure 5. Perform CEEMDAN module decomposition on the training set

If modeling each decomposed sequence separately will lead to increased computation and because white noise is added to the CEEMDAN decomposition process, modeling each sequence separately may lead to lower prediction accuracy due to the presence of these white noise. Therefore, it is necessary to summarize IMF into high frequency sequences (imfH), low frequency sequences (imfL), and remaining terms (imfR). The high frequency sequence reflects the short-term market fluctuations of the original sequence, and the low frequency sequence reflects the periodic influence of the original sequence, while the remaining term reflects the long-term slow change trend of the original sequence. Model the three sequences separately to mine information on different frequencies in the original sequences.

Since the high-frequency data are fluctuations caused by random factors and fluctuate around zero, the IMF is a sequence of high-frequency by judging whether the sequence mean is 0 by using t-test for IMF. In order to make the frequency of high frequency data more concentrated, there are enough high frequency fluctuations, can more reflect the market short-term random fluctuations, the judgment of the IMF for the high frequency sequence conditions add a change percentage condition, namely if the time point and the next time point of data negative different change times plus one, finally with the change number divided by all the data get change percentage, only when the change percentage is greater than K that the IMF has enough high frequency fluctuations, for high frequency sequence. A t-test p-value greater than 0.05 and percent change greater than 5% were considered as high frequency.

By comprehensive analysis, IMF1-IMF4 is high frequency component, and IMF5-IMF9 is low frequency component. IMF1-IMF4 is summarized into high-frequency sequence imfH, and IMF5-IMF9 is summarized into low-frequency sequence imfL. The results are shown in the following figure 6:

Figure 6. Fluctuation curves of raw, high-frequency, low-frequency, and residual sequences

As can be seen from Figure 6, the high frequency sequence is a random fluctuation with high frequency and low range up and down around zero, reflecting the random volatility of the short-term market, with large fluctuations after 2008,2012,2014 and 2020. While the low frequency sequence is similar in shape to the original sequence, indicating that the low frequency sequence reflects the periodic change of the original sequence. The remaining items show an S-shaped curve rise, reflecting the long-term trend change of gold futures.

4.2. Establishment of the CEEMDAN-ARIMA-SVR model

4.2.1 High frequency and low frequency components modeling

The traditional time-series modeling method in Chapter 3 is used to model the high-frequency component imfH and low-frequency component imfL. Since the high-frequency sequence distribution changes greatly, the original sequence may not be the best, so the traditional time series model suitable for the sequence before modeling the high-frequency sequence and the low-frequency sequence. Through experimental analysis, the high-frequency sequence is suitable for the GARCH-get model, and the low-frequency sequence is suitable for the EGARCH-t model. The modeling of high-frequency and low-frequency sequences was carried out according to the following process. The iterative decomposition of data obtained high-frequency data, and then the one-step prediction results were obtained by modeling, and the one-step prediction results of 100 days were obtained after 100 iterations.

4.2.2The remainder of the modelling

The rest was modeled using the SVR with the RBF kernel function, C=10, γ=0.001.

4.3. Experimental results and their analysis

The resulting 100-day one-step prediction results of high-frequency, low-frequency and remaining sequences were summed to obtain the final 100-day one-step prediction in Figure 7 and Table 4.

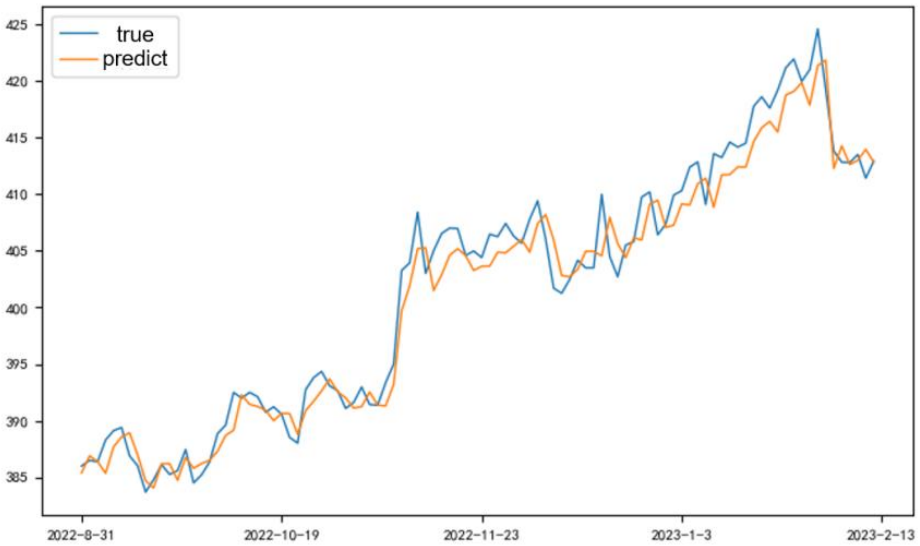

Figure 7. Final EDDMDAN-GARCH-SVR prediction results

Table 4. Final EDDMDAN-GARCH-SVR prediction evaluation indicators

Index	MAE	MSE
Value	1.6984	4.3050

According to Table 5,comparing the three models of MAE and MSE, it can be seen that no CEEMDAN decomposition modeling, EGARCH model MAE and MSE are less than SVR, illustrates the EGARCH model has a better prediction effect, the reason is that although SVR model has strong nonlinear fitting ability, but two super parameter selection difficulty, difficult to select the optimal super parameter combination, lead to the performance of the SVR model is not very good play. It also shows that the use of the traditional time series model is more convenient and can get a better prediction effect. While the CEEMDAN-SVR-EGARCH model using CEEMDAN decomposition with minimal MAE and MSE, 8.54% and 25.16% compared than SVR model indicators, By 1.52% and 21.35% compared to the EGAECH model, Show that the proposed method has obvious advantages in prediction, By splitting the complex data into frequency sets, Several relatively simple

sequences are summarized into high frequency, low frequency and residual sequences can effectively change the complex signal into simplification, Momodel the three sequences.

Table 5. Predictive ability of three sequence models

	MAE	MSE
EGARCH	1.7246	5.4733
SVR	1.8533	5.7519
CEEMDAN-SVR-GARCH	1.6984	4.3050

5. Conclusion

This paper draws the following conclusions: (1) through the analysis of AU9999 data, the conclusion that the sequence distribution is right-skewed, the peak ratio is flat normal distribution, and it has ARCH effect and leverage effect, which is suitable for the establishment of EGARCH-t model, and the model α+β is close to 1, indicating that the frequency of fluctuation is very high. However, the distribution of high frequency sequences after "decomposition-reconstruction" shows sharp and thick tail, while the distribution of low frequency and remaining sequences is similar to the original sequence. Through analysis, high frequency sequences are suitable for GARCH-get model, and low frequency sequences are suitable for EGARCH-t model.(2) The one-step prediction accuracy of the CEMDAN-SVR-GARCH model based on the decomposition algorithm is higher than the accuracy of the GARCH model and SVR model without decomposition, decreasing by 8.54% and 25.16%, and by 1.52% and 21.35% compared with the EGAECH model, which shows the obvious superiority of the proposed algorithm.

Based on the above conclusions, suggestions are put forward: (1) from the perspective of the sequence after data decomposition and reconstruction, gold futures have a long-term upward trend. If individuals need long-term investment and cannot find better investment targets, they can choose long-term investment in gold, and they can add gold to their investment portfolio to hedge risks. From the perspective of low frequency sequence, in the medium and long term, the cyclical fluctuation of gold has a great impact on the gold price. If you need to invest in gold, you need to pay more attention to various factors affecting the cyclical changes of gold and grasp the price fluctuation of gold.(2) For the government, on the one hand, it is necessary to consider sufficient gold reserves to resist risks and maintain the stability of the financial market. On the other hand, as gold fluctuates greatly in the medium and long term, it is necessary to pay more attention to the market risks and liquidity risks of gold, fully consider various factors, and plan gold reserves.

Reference

[1] Ding Lei, Guo Wanshan. Study on Gold Price Prediction based on ARIMA-GARCH family hybrid model [J]. Journal of Xuchang College, 2019,38 (06): 124-129.

[2] Hao Sun. The fTGARCH model and the GARCH-SVR model for predicting financial time series volatility [D]. Dalian University of Technology, 2020.

[3] Huang Ying, Yang Huijie. Financial time series prediction based on the XGBoost and LSTM models [J]. Technology and Industry, 2021,21 (08): 158-162.

[4] Qi Xie, a machine learning-based financial time series analysis method and its application [D]. Wuhan University of Science and Technology, 2020.

[5] Chen Lingling. Application of machine learning in financial time series prediction [D]. Hangzhou Dianzi University, 2020.

[6] Jianwei E , Ye J , Jin H . A novel hybrid model on the prediction of time series and its application for the gold price analysis and forecasting[J]. Physica A: Statistical Mechanics and its Applications, 2019, 527:121454.

[7] Zhang P , Ci B . Deep belief network for gold price forecasting[J]. Resources Policy, 2020, 69(10):101806.

[8] Jiang Chaoyu. Gold futures price forecast based on CNN-Transformer [D]. Shanghai University of Finance and Economics, 2021.

[9] Liang Y, Lin Y, Lu Q. Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM[J]. Expert Systems with Applications, 2022, 206: 117847.

[10] He Linyun. Gold futures price prediction based on the ICEEMDAN-SE-SSA-ELM algorithm [J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2023,37 (01): 35-39.

[11] Yuan Dongfang. Gold futures price forecast based on the CEEMDAN-PCA-LSTM model [D]. Shandong University, 2021.

[12] Elsayed S, Thyssens D, Rashed A, et al. Do we really need deep learning models for time series forecasting?[J]. arXiv preprint arXiv:2101.02118, 2021.