Research on Gold Price Prediction Model Based on CEEMDAN-XGBoost

Xin Xie

Central South University, Changsha, Hunan, China

Xiexin666xin@163.com

Abstract. This study proposes a hybrid model that combines Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Extreme Gradient Boosting (XGBoost) for predicting the time series of gold prices. Addressing the performance bottlenecks of traditional XGBoost when dealing with nonlinear and non-stationary signals, this research first applies CEEMDAN to decompose historical gold price data into multiscale signals, extracting Intrinsic Mode Functions (IMFs) of different frequency components. These IMFs are then used as input features for training and prediction with the XGBoost model. Experimental results show that the CEEMDAN-XGBoost model achieves high prediction accuracy on both the training and testing sets, outperforming the standalone XGBoost model in terms of generalization ability and prediction performance. This study not only provides an effective modeling approach for gold price prediction but also offers new insights for handling nonlinear and non-stationary features in other financial time series data.

Keywords: Gold Price Prediction; CEEMDAN; XGBoost; Time Series Analysis; Machine Learning.

1. Introduction

Gold, as one of the most important global financial assets, has long been considered a safe-haven tool against inflation and economic uncertainty. Its price fluctuations not only reflect the macroeconomic conditions but also have significant impacts on investor behavior, monetary policy formulation, and financial market stability[1]. Gold prices are influenced by multiple factors, including the US dollar exchange rate, oil prices, interest rates, stock market volatility, and geopolitical risks, among other complex variables, making gold price forecasting a core issue in financial research and investment decision-making[2]. Traditional statistical models, such as Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH), face limitations in handling nonlinear and non-stationary time series, making it difficult to capture the complex dynamic relationships among economic variables, leading to insufficient prediction accuracy. With the rapid development of artificial intelligence and big data technologies, machine learning (ML) and deep learning (DL) methods have been widely applied to financial market prediction tasks, offering new possibilities for modeling complex time series data and extracting nonlinear features. These methods can automatically learn latent patterns from historical data, thereby effectively improving the accuracy and stability of gold price predictions[3].

In related studies, traditional machine learning methods such as Linear Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Trees (GBT) have been successfully used for gold price prediction[4]. For instance, gradient boosting regression models constructed using macroeconomic variables such as the US dollar index, crude oil prices, and the S&P 500 index have achieved high prediction accuracy[5]. Random Forest-based analysis of 22 market variables' impact on gold prices further validated the applicability of machine learning models in complex market environments[6]. In contrast, deep learning models, with their superior feature representation and temporal dependency capture capabilities, have seen increased applications in gold price prediction. A comparison between Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in the Indonesian gold market found that the GRU model significantly outperforms traditional regression methods in terms of prediction accuracy[7]. A two-stage deep fusion framework that combines feature fusion and residual correction effectively improves the

stability and accuracy of multi-market gold price predictions[8]. Moreover, ensemble learning and hybrid modeling studies show that heterogeneous ensemble learning frameworks that combine multiple algorithms can further improve prediction performance and model robustness. Research indicates that multi-layer ensemble models combining Random Forest, SVM, Gradient Boosting, and LSTM achieve superior prediction results across multiple metrics compared to single models[9].

Despite significant progress in these studies, there remain several shortcomings in the current literature. First, most studies focus only on single model structures or limited feature sets, failing to adequately capture the nonlinear interactions and temporal dependencies among multi-source economic indicators[10]. Secondly, although deep learning models have powerful fitting capabilities, they face risks of overfitting and lack of interpretability, limiting their application in high-risk financial decision-making[11]. Additionally, the price linkage effects between different gold markets (e.g., London, New York, Shanghai) and cross-market dynamic features have not been sufficiently considered, which undermines the generalization performance of the models. Therefore, how to improve the model's interpretability and stability while ensuring prediction accuracy remains a critical issue in the field of gold price forecasting.

2. Model Methodology

2.1 CEEMDAN

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an advanced algorithm that further develops Empirical Mode Decomposition (EMD) and its improved version, Ensemble Empirical Mode Decomposition (EEMD). Traditional EMD decomposes non-stationary signals into several Intrinsic Mode Functions (IMFs) and a residue term based on local feature scales. However, EMD is prone to mode mixing when processing signals with strong noise, leading to the blending of features from different time scales into the same IMF, which affects the accuracy of signal feature extraction. EEMD introduces white noise and averages multiple decompositions to alleviate the mode mixing problem, but it suffers from large reconstruction errors and high computational complexity.

CEEMDAN improves upon EEMD by introducing an adaptive noise strategy, ensuring that the IMF components obtained in each decomposition step more accurately represent the intrinsic features of the signal at corresponding time scales. The basic idea is to add Gaussian white noise of varying amplitude to the original signal and decompose it multiple times to obtain the first-order IMF by averaging each decomposition. The IMF is then removed from the original signal, and adaptive noise is added to the residual signal for further decomposition. This process is repeated until all IMFs are obtained. Compared to EEMD, CEEMDAN offers higher signal reconstruction accuracy and better stability, effectively avoiding the generation of pseudo-IMFs while preserving the nonlinear and non-stationary characteristics of the signal.

CEEMDAN's good adaptability and high-precision signal reconstruction capabilities have made it widely used in nonlinear, non-stationary signal analysis fields, such as fault diagnosis, seismic wave analysis, financial time series modeling, and energy consumption forecasting. In this study, CEEMDAN is used to decompose the original time series signal, extracting IMF components at different frequency scales, providing more physically meaningful input features for the subsequent XGBoost-based prediction model and improving overall prediction performance and model generalization ability.

2.2 XGBoost

Extreme Gradient Boosting (XGBoost) is an efficient ensemble learning algorithm based on the gradient boosting framework. It makes several improvements over the traditional Gradient Boosting Decision Tree (GBDT), offering higher computational efficiency, stronger generalization ability, and excellent model stability. XGBoost builds multiple weak learners (usually regression trees) and

iteratively optimizes the loss function in an additive model form to achieve high-precision fitting of nonlinear relationships.

Compared to traditional GBDT, the main improvements of XGBoost include: first, the introduction of second-order derivative information to accelerate convergence and make the approximation of the loss function more accurate; second, the addition of a regularization term to the objective function, effectively controlling model complexity and preventing overfitting; furthermore, XGBoost employs feature selection strategies based on approximation algorithms when splitting nodes and supports sparse matrix processing and parallel computing, greatly enhancing training efficiency and model scalability. The objective function of XGBoost can be expressed as:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \tag{1}$$

In time series modeling and prediction tasks, XGBoost demonstrates excellent prediction performance with its powerful nonlinear modeling capabilities and adaptability to high-dimensional features. Unlike traditional statistical models (e.g., ARIMA, SVR), XGBoost does not rely on data stationarity assumptions, making it capable of flexibly capturing complex time series features and multi-scale dynamic changes. By weighting and combining input features, XGBoost effectively explores potential variable interaction relationships, thus enabling precise prediction of the target variable.

In this study, XGBoost is used to model and predict the IMFs and residual signals obtained from the CEEMDAN decomposition. Each IMF component corresponds to features at different frequency scales, and XGBoost can independently train sub-models for the temporal characteristics of each component, achieving fusion of multi-scale information. Finally, the results of each sub-model are weighted and reconstructed to effectively improve overall prediction accuracy and robustness. This approach fully combines CEEMDAN's signal decomposition ability with XGBoost's nonlinear modeling advantages, providing an efficient and reliable hybrid modeling framework for predicting complex non-stationary signals.

3. Dataset Overview

3.1 Data Partitioning and Visualization

Figure 1. Time Series Trend of Gold Prices

This study uses historical gold price data as the experimental dataset, sourced from publicly available financial market data. The time span covers from October 31, 2012, to October 28, 2022, and includes daily opening prices of gold during this period. To construct effective training and testing sets, the dataset is partitioned in an 8:1:1 ratio, with 80% of the data used for model training, 10% for model validation, and the remaining 10% for final testing. This partitioning ratio ensures that the model can fully learn the trends and patterns in the data while effectively evaluating the model's generalization ability.

To better understand the distribution characteristics of the data, this study first conducts a visualization analysis. By plotting a time series graph, **Figure 1** shows the fluctuation trend and periodic variations of gold prices over the entire time period. From the visual results, it is evident that gold prices exhibit significant nonlinear and non-stationary characteristics at different time periods, which provides the theoretical basis for applying the CEEMDAN algorithm for signal decomposition in subsequent steps.

3.2 Experimental Framework

The experimental framework of this study consists of two main steps: First, CEEMDAN is used to preprocess the original gold price data to extract Intrinsic Mode Functions (IMFs) at different frequency scales. Second, these IMF components are used as input features for prediction with the XGBoost model. The framework aims to validate the effectiveness of CEEMDAN in extracting signal features and assess the enhancement effect of combining it with XGBoost for gold price prediction. Specifically, the CEEMDAN algorithm first decomposes the gold price time series data, obtaining multiple IMF components at different frequency scales and preserving the primary features of the signal by removing the residual part. These IMFs are used as independent input variables and fed into the XGBoost model for training and prediction. The XGBoost model then uses gradient boosting methods for nonlinear modeling, continuously optimizing the prediction results through tree-based models.

To further evaluate the effectiveness of the CEEMDAN and XGBoost combined model, comparative experiments are also conducted. The comparison group uses traditional single time series forecasting methods, where the original gold price data is directly used to train the XGBoost model without the CEEMDAN preprocessing. By comparing the prediction performance of these two models, the impact of CEEMDAN on the accuracy and stability of gold price predictions can be analyzed. In terms of evaluation metrics, this study employs several common regression model evaluation methods, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), to comprehensively assess the model's prediction accuracy and generalization ability.

4. Experimental Comparative Analysis

4.1 Data Decomposition Based on CEEMDAN

Figure 2 shows the Intrinsic Mode Functions (IMFs) obtained after applying the CEEMDAN algorithm to decompose the gold price time series data. The different IMF components in the figure clearly reveal the changing characteristics of the signal at various frequency scales. In this experiment, the decomposition process uncovers the complex dynamic fluctuations of gold prices, which correspond to multiple time series components at various frequencies.

From **Figure 2**, it can be observed that IMF1 and IMF2 correspond to high-frequency components, primarily reflecting the details of short-term fluctuations. IMF3 represents mid-frequency components, capturing relatively stable trend changes, while IMF4 presents low-frequency long-term trend components. Each IMF layer effectively reveals different feature levels in the original signal, fully demonstrating the advantages of CEEMDAN in extracting nonlinear and non-stationary data features.

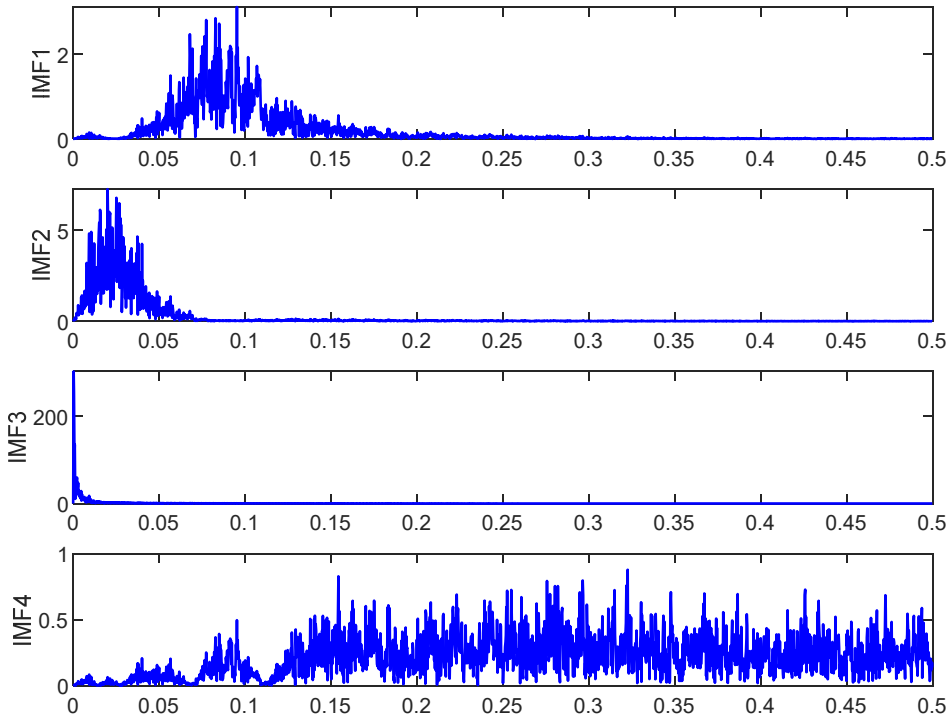

Figure 2. Intrinsic Mode Functions (IMFs) After CEEMDAN Decomposition

Through this decomposition, noise and the true trend in the original data are separated, providing clearer and more meaningful input features for the subsequent prediction model. These decomposed signal components will be used as input data for the XGBoost model, enabling the model to more accurately capture the changing trends of each frequency component when predicting gold prices.

4.2 XGBoost Prediction

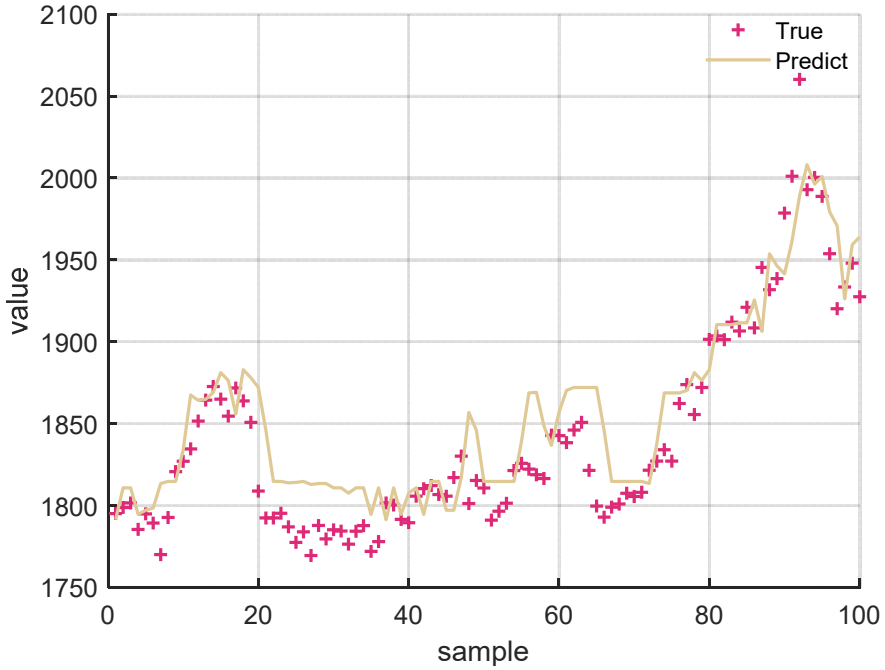

Figure 3. Comparison of Predicted and Actual Values for the XGBoost Model on the Test Set

Figure 3 illustrates the results of gold price prediction using the XGBoost model, where the actual values are represented as scatter points and the predicted values are shown as a continuous line. From the figure, it is evident that the prediction curve closely follows the real price changes, effectively capturing the main features of gold price fluctuations in both short-term volatility and medium-to-long-term trends. Although there are some deviations in regions of high volatility, the overall trend remains consistent, demonstrating that the XGBoost model has strong fitting capability and predictive reliability when handling nonlinear financial time series data.

To further quantify the model's performance, this study calculates the prediction metrics for the training set, validation set, and test set separately. Table 1 presents the evaluation results of the XGBoost model on each dataset. During the training phase, the model demonstrates high precision, with a MAE of 7.1615, MAPE of 0.0053829, RMSE of 9.6414, and an R^2 of 0.99745, indicating that the model has achieved a high degree of accuracy in fitting the training data. The validation set results show a MAE of 20.0696, RMSE of 26.0623, and R^2 of 0.79376. While the error is slightly higher compared to the training set, the model still maintains a relatively high explanatory power, indicating strong generalization ability on unseen data. The test set results, with a MAE of 21.0753, MAPE of 0.011639, RMSE of 26.4743, and an R^2 of 0.90556, further confirm the model's robustness in real-world prediction scenarios. Overall, XGBoost shows strong fitting performance during the training phase and maintains good prediction accuracy in both the validation and test phases, suggesting its ability to effectively model the nonlinear dynamic structure of gold prices, making it valuable for practical applications.

Table 1. Evaluation Metrics of the XGBoost Model on Different Datasets

Dataset	MAE	MAPE	MSE	RMSE	R^2
Training Set	7.1615	0.00538	92.9563	9.6414	0.99745
Validation Set	20.0696	0.01114	679.2455	26.0623	0.79376
Test Set	21.0753	0.01164	700.8899	26.4743	0.90556

4.3 Ablation Study

In this section, we conducted an ablation experiment to compare the performance of two models: one that directly uses XGBoost for modeling and another that incorporates CEEMDAN for signal decomposition before applying XGBoost. This comparison helps assess the improvement in XGBoost's performance when CEEMDAN is introduced.

First, when training the XGBoost model directly, the training set shows excellent prediction accuracy with a MAE of 3.6907, MAPE of 0.0027677, RMSE of 4.7473, and R^2 of 0.99938, indicating that the model fits the training data very well, as shown in **Table 2**. However, the performance on the validation and testing sets drops compared to the training set. The MAE for the validation set is 41.8155, RMSE is 46.9056, and R^2 is 0.33198, while the test set shows an MAE of 31.7936, RMSE of 38.8063, and R^2 of 0.79709. This suggests that although the model fits the training data well, it struggles with generalization, especially with significant prediction errors on the validation set.

In contrast, the XGBoost model with CEEMDAN shows improvements in the training, validation, and test sets. The MAE for the training set is 7.1615, MAPE is 0.0053829, RMSE is 9.6414, and R^2 is 0.99745. Although these values are slightly lower than those of the pure XGBoost model, the prediction accuracy remains high. On the validation set, CEEMDAN-XGBoost achieves a MAE of 20.0696, RMSE of 26.0623, and R^2 of 0.79376, demonstrating more stable performance. The test set shows a MAE of 21.0753, RMSE of 26.4743, and R^2 of 0.90556, further confirming that this model significantly outperforms the pure XGBoost model, especially on the test set.

By comparing the results of the two models, it is clear that introducing CEEMDAN for data preprocessing effectively improves XGBoost's performance on both the validation and test sets, particularly in terms of prediction accuracy and generalization ability. CEEMDAN extracts different frequency components from the signal, providing more precise and meaningful features for XGBoost, which in turn enhances the overall performance of the model.

Table 2. Ablation Experiment Evaluation Analysis

Model	Dataset	MAE	MAPE	MSE	RMSE	R^2
XGBoost	Training	3.6907	0.00277	22.5366	4.7473	0.99938
XGBoost	Validation	41.8155	0.02327	2200.1346	46.9056	0.33198
XGBoost	Test	31.7936	0.01769	1505.9255	38.8063	0.79709
CEEMDAN- XGBoost	Training	7.1615	0.00538	92.9563	9.6414	0.99745
CEEMDAN- XGBoost	Validation	20.0696	0.01114	679.2455	26.0623	0.79376
CEEMDAN- XGBoost	Test	21.0753	0.01164	700.8899	26.4743	0.90556

5. Conclusion

In this study, we proposed a hybrid modeling method combining CEEMDAN and XGBoost to address the issues of nonlinearity and non-stationarity in gold price prediction. We first used CEEMDAN to decompose the time series data of gold prices, extracting multiple Intrinsic Mode Functions (IMFs), and then input these decomposed components as features into the XGBoost model for training and prediction. By comparing this model with the traditional XGBoost, we demonstrated that the introduction of CEEMDAN significantly improved the model's prediction accuracy and generalization ability.

The experimental results show that when XGBoost is used alone, the model performs excellently on the training set but exhibits considerable prediction errors on the validation and test sets. After introducing CEEMDAN for data preprocessing, XGBoost's performance on all datasets improved notably, especially in terms of prediction accuracy and model generalization. Specifically, the CEEMDAN + XGBoost model not only captures the dynamic trends of gold prices more effectively but also enhances the model's stability and predictive ability on unseen data.

This study verifies the effectiveness of CEEMDAN in time series data processing and demonstrates that combining it with XGBoost significantly boosts the model's prediction performance. Future research could explore the effects of different signal decomposition methods in combination with machine learning models, as well as extend this approach to the prediction of other non-stationary time series data, enhancing its application potential in fields such as finance, energy, and meteorology.

References

[1] S. Chandar, S. Mahendran, and S. Natarajan, “Forecasting Gold Prices Based on Extreme Learning Machine,” Int. J. Comput. Commun. Control, vol. 11, pp. 372–380, 2016.

[2] R. Kumar, J. Moolchandani, A. Shukla, S. Sahu, V. Thada, and V. Chole, “Machine Learning-Based Prediction of Gold Prices Using Economic Indicators,” in Proc. 13th Int. Conf. System Modeling & Advancement in Research Trends (SMART), 2024, pp. 520–524.

[3] C. Qiu et al., “A Two-Stage Deep Fusion Integration Framework Based on Feature Fusion and Residual Correction for Gold Price Forecasting,” IEEE Access, vol. 12, pp. 85565–85579, 2024.

[4] A. Gadhave, “Gold Price Prediction using Machine Learning,” Int. J. Sci. Res. Eng. Manag., 2022.

[5] S. Liu Sentiko, A. Y. Zakiyyah, and Meiliana, “Gold Price Prediction Using Machine Learning and Deep Learning,” in Proc. 6th Int. Conf. Cybernetics and Intelligent System (ICORIS), 2024.

[6] Y. Wang and T. Lin, “A Novel Deterministic Probabilistic Forecasting Framework for Gold Price with a New Pandemic Index,” Mathematics, 2023.

[7] N. M. Trieu and N. T. Thinh, “A Robust Prediction for Gold Price using Heterogeneous Ensemble Learning,” in Proc. 13th Int. Conf. Control, Automation and Information Sciences (ICCAIS), 2024, pp. 1–5.

[8] W. Gong, “Research on Gold Price Forecasting Based on LSTM and Linear Regression,” SHS Web Conf., 2024.

[9] D. M. T. Nguyen, N. C. Debnath, L.-D. Quach, and V. D. Nguyen, “Machine Learning Algorithms for Gold Price Prediction,” in Proc. Int. Conf. Advances in Information Technology and Education, 2023, pp. 212–220.

[10] H. Zangana and S. R. Obeyd, “Deep Learning-based Gold Price Prediction: A Novel Approach using Time Series Analysis,” SISTEMASI, 2024.

[11] S. Duman, S. Turnacıgil, E. Arık, and M. A. Aktaş, “The Role of International Variables in Predicting Gold Prices: Analysis with Machine Learning Algorithms,” Sosyoekonomi, 2024.