Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/47/20230371

Impact of Models and Eigenvalues on Gold Price Forecasting

Jin Zhang[1, a, *]

[1]*Mechanical and Electrical Engineering, Guangdong University of Technology, No. 100, Zhonghuan West Road , Panyu District, Guangzhou , China*
a. 3120000212@mail2.gdut.edu.cn
**corresponding author*

Abstract: The burgeoning synergy between computer science and finance has fostered an increasing integration of these domains. Machine learning has become a prevalent tool in aiding financial analysis and forecasting. Compared to traditional forecasting techniques, machine learning-based models exhibit enhanced accuracy and broader applicability. This study introduces three models, namely linear regression, random forest, and support vector machine, to analyze and predict gold prices. The influence of Eigenvalues on model performance is also examined. In the end, the support vector machine model constructed by using two kinds of US dollar exchange rates, US Treasury bond interest rates, and the 10-day moving average of gold prices and passed cross-validation obtained the best model performance evaluation index, and its R2 index reached nearly 0.99. It can be concluded from this study that the performance of the model is poor when only one eigenvalue is used to build the model, while for the case of building a model with multiple eigenvalues, the contribution of the U.S. Treasury bond rate to the improvement of the performance of the prediction model is the smallest. Therefore, appropriately increasing the number of eigenvalues is conducive to improving the performance of the model, and selecting the types of eigenvalues reasonably is also conducive to improving the accuracy of the model.

Keywords: machine learning, gold price forecast, eigenvalue analysis

1. Introduction

Gold is widely used for storage or investment around the world, due to its particularity as a bulk commodity and hard currency. Given the current turbulent world situation, the price of gold has surpassed $2,000 an ounce. Due to the properties of gold itself, necessitating a thorough investigation and projection of its price trends. The inherent characteristics of gold further accentuate the importance of comprehensively studying and forecasting its price trajectory. In this regard, machine learning, an automated data-driven methodology enabling predictive insights, has been employed in many domains of finance. The application of machine learning algorithms to predict gold prices stands as a notable exemplification of this approach within the financial landscape.

In the field of financial forecasting, there are many traditional forecasting methods, such as the forecasting of securities prices, the traditional time series forecasting technology, which uses the relationship between time and securities prices to analyze and predict the price of securities, but Since the stock price will be affected by the volatile stock market, it is a very complex nonlinear system, so the accuracy of using the traditional time series technology to predict the stock price is not high[1].

DOI: 10.54254/2754-1169/47/20230371

This phenomenon will also appear in the prediction of the price of other items whose price changes are intricate and affected by many factors, such as the price of crude oil and the price of gold [2].

Since the 1950s, with the continuous development of computer technology, data science and machine learning have been greatly developed. These technologies use the analysis and learning of a large amount of data to extract the relationship contained in the data. And then make better use of data. So far, artificial intelligence technology has been constantly changing our lives. As an important issue of artificial intelligence, machine learning is also worthy of in-depth research as an important difficulty. In the financial field, machine learning is combined with various traditional models to obtain a variety of better financial data analysis and processing methods, which can help financial practitioners perform better related operations in their respective fields [3].

There are also many related applications in many other financial fields. For example, in the prediction of China's stock market, You et al. proposed quantitative research on investment portfolios based on machine learning [4]; Manjula et al. employed the ensemble-based machine learning approach to forecast gold prices and understand the relationship between the gold price and factors [5]. In the prediction of Bitcoin prices, there are also attempts to use deep machine learning such as Recurrent Neural Networks (RNN), Long Short Memory (LSTM), Gate Recurrent Unit (GRU) and other methods for predicting [6], since these models' satisfactory performance in other domains [7-9]. In terms of gold forecasts, there are also many cases of research using different machine learning methods such as various neural network learning methods for gold price forecasting.

This article aims to employ linear regression analysis, random forest, and support vector machine these three machine learning models to predict gold prices. To use machine learning to predict gold prices, the author imports relevant databases perform merge preprocessing, and conducts research and analysis on the processed data. Subsequently, the acquired gold price forecast outcomes are scrutinized to discern the impact of various factors, such as the U.S. treasury bond interest rate, gold 10-day moving average, and US dollar foreign exchange rate, either individually or in combination, on the accuracy and efficacy of the gold price predictions. Through this analysis, the article seeks to enhance our understanding of the intricate relationship between these factors and their influence on the forecasting of gold prices.

2. Method

2.1. Data Preparation

The data used in the study is mainly obtained from Kaggle [10]. Three small databases were employed, namely the gold price database, the US dollar exchange rate database, and the US treasury bond interest database to build the machine learning models.

Before training the machine learning models, relevant data preprocessing was performed on the collected database. Given the disparity in the time durations and occurrence of missing data across the three collected databases, a decision has been made to standardize the time intervals for machine learning and analysis purposes. Specifically, the time range selected for this analysis spans from March 1, 2017, to December 31, 2019. This time period was chosen based on the availability of complete data within the databases, and it ensures the fulfillment of machine learning requirements while adhering to the specified size limitations of the required database. By employing this standardized time range, the analysis can be conducted effectively while maintaining the necessary level of accuracy. For the correction of missing values and wrong values, the data value of the previous day of the data is used to replace and correct the missing or wrong values. Since the selected time span is relatively long, the error caused by this replacement method is acceptable.

Given the abundance of eigenvalues present in the merged database, it is crucial to address the potential issue of overfitting when constructing a machine learning model. Utilizing all eigenvalues

DOI: 10.54254/2754-1169/47/20230371

indiscriminately may lead to a decrease in model accuracy. Therefore, after the revised total database is generated, the correlation analysis between each eigenvalue and the gold price is carried out, and several eigenvalues with positive correlation and negative correlation with the gold price are selected to build a model, so as to prevent over-correlation. Fitting occurs, improving the accuracy of predictions.

Since the price of gold will undergo serious jumps due to violent fluctuations in the world environment in some periods, the changes in the price of gold in the selected range are used to draw the relevant data of the ten-day moving average, and the data of the ten-day moving average is used to build the machine Learning model with gold price forecasts.

2.2. Machine Learning Models

2.2.1. Linear Regression

Linear regression analysis is a statistical analysis method for predicting variables based on large amounts of data. It is mostly used to construct a linear regression model for data, perform related machine learning and analysis, and predict the required results. This is the step of using the linear regression model to make predictions.

Since multiple eigenvalues are selected to predict the price of gold, the method of multiple linear regression is used to predict the price of gold. In order to ensure the accuracy of the model and avoid collinearity among the features, the eigenvalues are required to be independent of each other [11]. The formula of the linear regression model can be found as follows:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon, \quad (1)$$

2.2.2. Random Forest

The random forest model is a machine learning prediction method that builds multiple decision trees and fuses them to obtain a more accurate model. Several decision trees constructed by random forest, when a sample needs to be predicted and analyzed, each tree in the random forest corresponds to a possible prediction result, and the random forest model is obtained by voting on its prediction results.

When considering the eigenvalues of the random forest, it is essential to be mindful of their correlation as it directly impacts the prediction error rate. An increase in correlation among the eigenvalues leads to a higher error rate within the random forest. Consequently, careful attention must be given to the correlation of the input eigenvalues in the data. Moreover, the classification capability of each individual tree directly influences the overall classification ability of the random forest model. Therefore, in order to obtain more accurate prediction results, attention should be paid to the selection of each feature value of the model [12].

2.2.3. Support Vector Machines

The support vector machine model is a model that uses a partition method to process data and make related predictions. Its basic model is a nonlinear classifier defined on the feature space, and its machine learning strategy is how to maximize the segmentation interval. The basic idea of performing support vector machine is to solve the hyperplane that can correctly divide the training data set and have the largest geometric interval. Finally, using the hyperplane divided according to the learned eigenvalue data to predict the result.

DOI: 10.54254/2754-1169/47/20230371

2.2.4. Implementation Details

In this paper, sklearn is used to implement these models. Before building the model, the data is first divided into the train data set and a test set. The training set is utilized to train the models, and subsequently, the trained models are applied to the test set for evaluation. To enhance the accuracy of each model's predictions, a cross-validation approach is employed to obtain optimized hyperparameters for different models. The performance of the models is assessed by evaluating the mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) metrics on the test set, providing insights into the model's effectiveness.

3. Results and Discussion

Through iterative iterations of model establishment and evaluation, the predictive accuracy of the established model was assessed using linear regression, specifically examining the influence of the number and type of feature parameters. The evaluation outcomes are presented in Table 1. When only the U.S. treasury bond interest rate is selected as the characteristic value affecting the gold price, it can be obtained according to three evaluation indicators, and the measurement results of the model are very unsatisfactory. When the characteristic value of the ten-day moving average is added, the result has been greatly improved, and the MAE evaluation index has tended to be stable. When the characteristic value of the exchange rate of the US dollar against other countries is added, the results are further improved, reaching 0.97 on the R2 indicator.

Table 1: The influence of selecting different quantity of eigenvalues on linear regression models.

	MSE	MAE	R2
Select 1*	3738.08	51.52	0.02
Select 2**	241.52	15.54	0.56
Select 3***	199.49	14.12	0.97
Select 4****	192.92	13.88	0.97

*Select 1 is to select only the exchange rate of US treasury bond bonds as the characteristic value.
**Select 2 is to select the ten-day moving average of US treasury bond exchange rate and gold price as the characteristic value.
***Select 3 is to select the exchange rate of US treasury bond bonds, the ten-day average line of gold price, and a US dollar exchange rate as the characteristic value.
****Select 4 is to select the exchange rate of US treasury bond bonds, the ten-day average of gold price, and the two US dollar exchange rates as the characteristic values.

Table 2: Performance of different models using the same four eigenvalues.

	MSE	MAE	R2
LinearRegression	192.92	11.39	0.97
random forest	122.22	8.15	0.97
Support vector machines	89.49	7.06	0.98

In the process of using different models for machine learning, by using the cross-validation method to optimize the hyperparameters of each model, the final evaluation indicators of the three models are relatively close shown in Table 2. Among them, cross-validation has the weakest optimization effect on the linear regression model, and the strongest optimization effect on the SVM model. The best predictive model was obtained using the SVM model.

The aforementioned findings highlight the strong association between the prediction accuracy of each model and the number of predictor variables utilized. Notably, when the number of predictor variables is small, the accuracy of the predictions progressively improves with an increase in the

DOI: 10.54254/2754-1169/47/20230371

number of variables. However, as the number of input predictor variables continues to rise, the predictive results tend to stabilize. Consequently, the inclusion of additional predictor variables may potentially have a detrimental effect on prediction accuracy, leading to overfitting or a decline in performance.

From the results in Table 1, it can be observed that the 10-day moving average index has the most obvious effect on the optimization of the linear regression model. Compared with the extremely low prediction results when using the U.S. Treasury bond interest rate alone, the results are greatly optimized after adding eigenvalues such as the ten-day moving average or the U.S. dollar exchange rate. It is speculated that for the use of linear regression models to predict gold prices, it is a good choice to choose its ten-day moving average or the US dollar exchange rate as its characteristic value. For the random forest model, by calling the feature_importances function to analyze the importance of each feature value to the prediction result, the US dollar exchange rate is slightly more important to the prediction accuracy of the model than the 10-day moving average, while the U.S. Treasury bond interest rate is also important to Forecast accuracy is less important. The same result demonstrated that in the forecast of gold price, more reference to its own 10-day moving average or the dollar exchange rate with strong correlation with gold will help improve the accuracy of the forecast.

4. Conclusion

In this study, the database reconstructed from multiple small databases is used to analyze the relationship between gold price forecasts, various eigenvalues, and various models, to find ways to optimize gold price forecasts. Three models of linear regression, random forest, and support vector machine are used to predict prices and make relevant comparisons. The results demonstrated that using a single eigenvalue to predict the price of gold is not ideal, but a reasonable increase in the number of eigenvalues is conducive to improving the performance of the established model. Moreover, the data analysis reveals that the US dollar exchange rate holds the greatest influence on the model's accuracy. Future research endeavors aim to incorporate additional predictor variables for analysis. This will facilitate the identification of predictor variables that yield greater benefits in optimizing the gold price prediction model. Furthermore, there is a plan to explore the fusion of multiple models to construct a more robust and high-performing model for gold price prediction.

References

[1] *Li, S. (2008) Several forecasting methods and empirical research on securities prices (in Chinese). (Doctoral dissertation, Jiangsu University*

[2] *Yang, X. (2006) Research on International Crude Oil Futures Price Forecast Based on Combination Forecast (in Chinese), Beijing Institute of Technology.*

[3] *Lou, Z. Q. (2019) The theoretical development and application status of machine learning (in Chinese). China New Communications (1), 3.*

[4] *Fang, Y., et al. (2022) Artificial Intelligence and China's Stock Market——A Quantitative Research on Investment Portfolio Based on Machine Learning Forecast. Industrial Technology Economics, 41(8), 9.*

[5] *Manjula, K. A., et al. (2019) Gold Price Prediction using Ensemble based Machine Learning Techniques. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).*

[6] *Ke, K. F. (2020) Research on Bitcoin Price Prediction Based on Deep Learning (Master's Thesis, Harbin Institute of Technology)*

[7] *Fu, R., Zhang, Z., Li, L. (2016) Using LSTM and GRU neural network methods for traffic flow prediction, 2016 31st Youth academic annual conference of Chinese association of automation (YAC). IEEE, 324-328.*

[8] *Yu, Q, Wang, J, Jin, Z, et al. (2022) Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323.*

[9] *Zha, W., Liu, Y., Wan, Y., et al. (2022) Forecasting monthly gas field production based on the CNN-LSTM model. Energy,124889.*

DOI: 10.54254/2754-1169/47/20230371

[10] Kaggle. (2021) Gold prices. https://www.kaggle.com/datasets/kamyababedi/gold-prices

[11] Kaggle (2022) Market yield us treasury securities percent.https://www.kaggle.com/datasets/taranmarley/market-yield-us-treasury-securities-percent

[12] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.