Research on gold price forecasting based on lstm and linear regression

Weichen Gong*

Ocean University of China, Faculty of Information Science and Engineering, Qingdao, 266100, China

Abstract. Gold price forecasting is critical in financial decision-making, providing valuable insights for in-vestors and stakeholders in the gold market. Deep learning methods have witnessed remarkable progress in various domains, including image recognition and sentiment analysis. This paper integrates LSTM (Long Short-Term Memory) and Linear Regression models to forecast the rise and fall of gold prices. The analysis of the prediction accuracy regarding the rise and fall of the daily gold price reveals that the LSTM model achieved an accuracy rate of 50.67%, while the Linear Regression model achieved a slightly higher accuracy rate of 53.02%. By combining the strengths of these models, this research provides valuable insights to investors in the gold markets.

1 Introduction

Finance is an integral part of our lives, and accurate financial forecasts are essential for economic development. In the realm of finance, gold holds significant importance as a precious metal and a store of value. Individuals, businesses, and major institutions all have a vested interest in understanding and predicting the future trends of gold prices. Therefore, employing effective methods for gold price forecasting holds great value for these three stakes- holders.

Traditionally, economists and researchers have utilized mathematical methods, such as the Naive Bayes algorithm, as statistical tools to predict future trends in financial prod- ucts. However, the field of finance presents unique chal- lenges, including the computational demands and the in- tricate nature of financial data. The effectiveness of the Naive Bayes classifier is limited by its ability to handle large-scale data and continuous features through binning procedures.

Moreover, traditional quantitative factors models face challenges in capturing the intricate and nonlinear relationships that exist in the financial domain. Various approaches have been explored, including dynamic models, ad-hoc factor-timing techniques, and dynamic learning capabilities in linear regression models. However, criticisms arise due to the complexities involved, as numerous factors influence the estimation of the relationship between potential predictors and expected returns.

Recent advances in deep learning, which have demonstrated remarkable performance in fields such as computer vision and natural language processing, have sparked interest in applying these techniques to financial data analytic.

While the application of deep learning in financial forecasting is still a topic of debate, the fundamental principle remains to map input data to corresponding returns. This characteristic theoretically enables deep learning to discover relationships associated with gold price movements, disregarding the irregular and nonlinear nature of the input data. More specifically, several relevant studies have already been carried out by some researchers.

Shafiee and Topal have examined the advancements and developments of gold price forecasting models, including reveal the connection between oil price and gold price [1]. Building upon their work, no- table studies that employ machine learning techniques for gold price prediction can be briefly highlighted.

With the abundant availability of figures in financial domain and improved calculating effectiveness in recent decades, researchers have developed various machine learning-based gold price forecasting models. Tripathy compares the performance of two models (ARIMA and MLR) in monthly gold price prediction [2]. The results indicate that the ARIMA model outperforms the MLR model, and the study identifies the significant impact of the latest month gold price on the current price. Bandyopadhyay and Guha further highlight the limitations of time-series techniques, particularly the shortcomings of the ARIMA model, in forecasting gold prices [3]. As the gold price exhibits nonlinearity, the ARIMA model's effectiveness is limited to short-term predictions based on minor variations observed in the dataset. Makridou et al. Propose the Adaptive Neural-Fuzzy Inference System (ANFIS) model as a superior alternative to auto-regression (AR), autoregression moving average (ARMA), and ARIMA for gold price forecasting [4]. Their study demonstrates the strong predictive performance of the ANFIS model and suggests the potential application of neuro-fuzzy-based models in forecasting the prices of other commodities.

Hadavandi et al. introduced a time series model for precise gold price forecasting, employing a particle swarm optimization (PSO) approach. Their research showcased

*corresponding author: 20020036017@stu.ouc.edu.cn

the model's effectiveness in managing fluctuations and achieving favorable prediction accuracy, thus positioning it as a suitable tool for addressing financial forecasting challenges [5]. Liu and Li conducted research on gold price fluctuation trend prediction and proposed the use of the random forest method for predicting gold price fluctuations [6]. They acknowledged the importance of accurately predicting gold price trends and identified various factors considered in related literature.

However, gathering data for multiple factors can be challenging. Through extensive experiments with real-world data, the authors found that the random forest method was powerful in predicting gold price trends. Their findings indicated that only two factors, DJIA and SP500, were crucial for achieving accurate predictions using the random forest algorithm [6]. Madziwa et al. utilized the ARDL model to predict yearly gold prices and found it to outperform stochastic mean reverting and ARIMA approaches. Their findings emphasized the significant impact of gold demand on prices, while indicating that treasury bill rates had no notable influence [7]. Li proposed a gold price forecasting approach that combines a WNN with a novel ABC algorithm [8]. The improved algorithm introduces a replacement for the conventional roulette selection strategy, incorporating feedback messages from previous iterations' convergence statuses. This modification enhances the search intensity in subsequent cycles. Experimental results have shown that this novel algorithm exhibits faster convergence compared to the conventional Artificial Bee Colony method when applied to benchmark functions. Additionally, it effectively enhances the modeling capacity of the Wavelet Neural Network (WNN) for forecasting gold prices. Chen and Huang developed algorithms using various input features, including gold prices and financial indicators, to accurately predict stock price trends [9]. Their proposed approach achieved a 67% accuracy rate and demonstrated higher ROI compared to other models, highlighting the importance of considering gold and crude oil factors for specific industries. Verma et al. enhanced gold price forecasting in the Indian market through the application of modified Artificial Neural Network (ANN) techniques, resulting in enhanced efficiency [10]. Khani et al. developed an LSTM- based model for stock market prediction during the COVID-19 pandemic [11]. Their study utilized features such as COVID-19 cases and market tickers, with the vector sequence output LSTM achieving the best performance among other methods.

These studies collectively contribute to the exploration of machine learning techniques for gold price prediction, providing different aspects of the advantages and drawbacks of various models.

This article aims to examine the effectiveness and rationality of employing Linear Regression and LSTM models for gold price prediction. By supplementing and expanding upon existing research on gold price forecasting, this study offers novel prediction approaches and strategies to investors and relevant institutions. Additionally, by conducting a comprehensive analysis of the LR and LSTM models, this research not only advances knowledge in the field of gold price prediction but also offers valuable insights for decision-making in the financial market.

2 Dataset

2.1 Data analyzing

The dataset used in this study was collected from Kaggle. The dataset consists of seven columns, including Date, Open, High, Low, Close, Volume, and Currency. These columns offer valuable insights into the daily trading activities associated with a particular asset.

The dataset covers a considerable time span, ranging from January 4, 2000, to September 2, 2020, allowing for a comprehensive analysis of long-term trends and patterns. The inclusion of data from over two decades enables researchers to explore the dynamics of the asset's price and volume over different market conditions.

In terms of its size, the dataset consists of 5703 samples, representing individual data points or observations. Each line corresponds to a specific date and includes information such as the opening and closing prices, the highest and lowest prices recorded during the trading day, the trading volume, and the currency in which the asset is denominated.

This dataset serves as a valuable resource for conducting empirical analyses and developing predictive models in the financial domain. The extensive coverage and richness of information within the dataset enable researchers to gain insights into the asset's historical performance and make informed decisions based on the observed patterns and trends.

Data analysis involves gathering and assessing basic in- formation about the dataset, such as its size, and variables. Descriptive statistics are calculated to understand the data distribution.

Gathering basic information about the dataset is crucial. This involves obtaining details regarding the source of the data, such as the organization or entity that collected or provided the dataset. Understanding the dataset's size refers to the number of observations or records contained within it, which helps in assessing its comprehensiveness and potential limitations.

The correlation between numeric columns in a data frame represents the strength and direction of the linear relationship between those columns. It provides insights into how closely related two variables are and how they change in relation to each other. The result is shown in the Figure 1.

	Open	High	Low	Close	Volume
Open	1.000000	0.999879	0.999825	0.999740	0.692123
High	0.999879	1.000000	0.999778	0.999861	0.693861
Low	0.999825	0.999778	1.000000	0.999893	0.688983
Close	0.999740	0.999861	0.999893	1.000000	0.690534
Volume	0.692123	0.693861	0.688983	0.690534	1.000000

Fig. 1. Pairwise correlation of the dataset

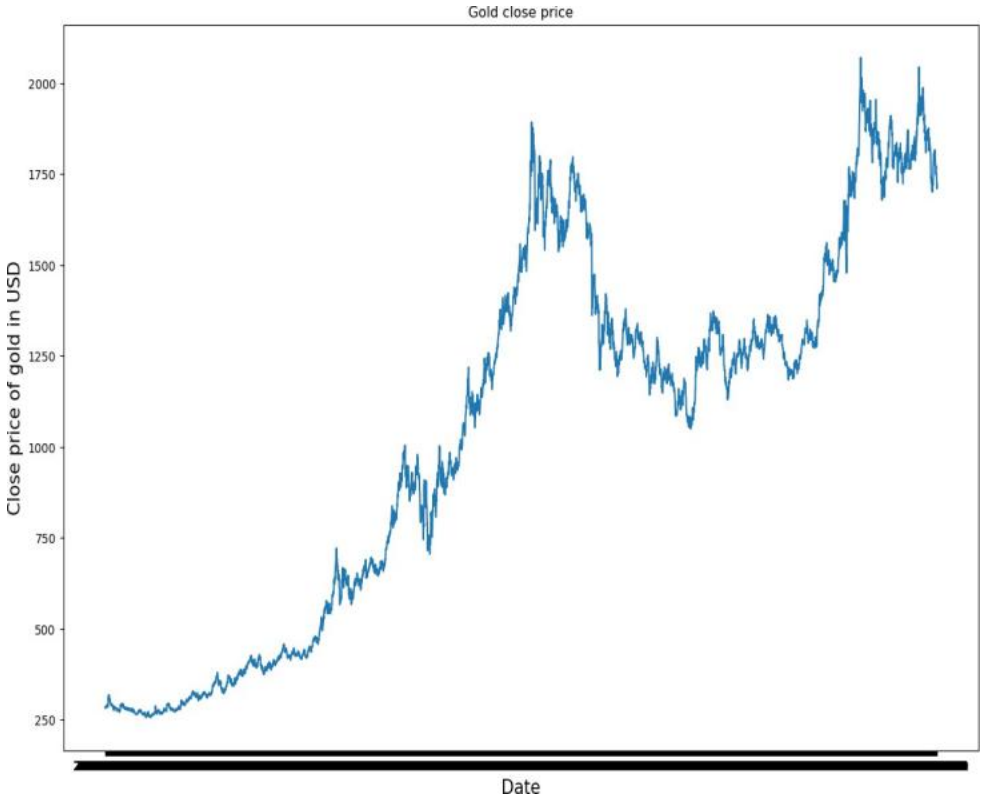

Fig. 2. Lineplot

Based on the result shown in the Figure 1, it appears that the Open, High, Low, and Close columns have a very high positive correlation with each other, as indicated by correlation coefficients close to 1. This suggests a strong linear relationship between these variables, which is ex- pected since these columns are likely related to stock mar- ket data such as the opening, highest, lowest, and closing prices of a stock.

The Volume column also shows a positive correlation with the other columns, but the correlation coefficients are lower compared to the price-related columns. This sug- gests a weaker linear relationship between the volume and price variables.

It's worth noting that correlation coefficients above 0.9 in- dicate a very high correlation, and the differences between 0.999879, 0.999825, and 0.999740 are minimal. These values suggest an almost perfect positive correlation be- tween the price-related columns.

2.2 Data Visualization

Data visualization is an essential step in data analysis that involves representing data visually to gain insights, identify patterns, and communicate findings effectively. It allows for a more intuitive understanding of complex datasets and facilitates the exploration of relationships and trends.

A line plot is a common data visualization technique used to display the trend or pattern of a variable over time. It involves plotting data points connected by straight lines. The line plot shown in Figure 2 illustrates the historical trend of the close gold price over a specific time period. The x-axis represents the time range from January 4, 2000, to September 2, 2020, while the y-axis represents the corre- sponding close gold prices.

Initially, the gold price started at around 250. From Jan- uary 4, 2000, to approximately October 30, 2006, the gold price gradually climbed to approximately 450. During the period from October 31, 2006, to December 29, 2016, the gold price experienced a rapid ascent from 450 to around 1750, accompanied by significant fluctuations.

In the final phase, spanning from December 30, 2016, to September 2, 2020, the gold price exhibited a U-shaped pattern. It first declined from 1750 to 1250 and remained at this level for a while. Subsequently, it rose again to ap- proximately 2000. Throughout this period, the gold price

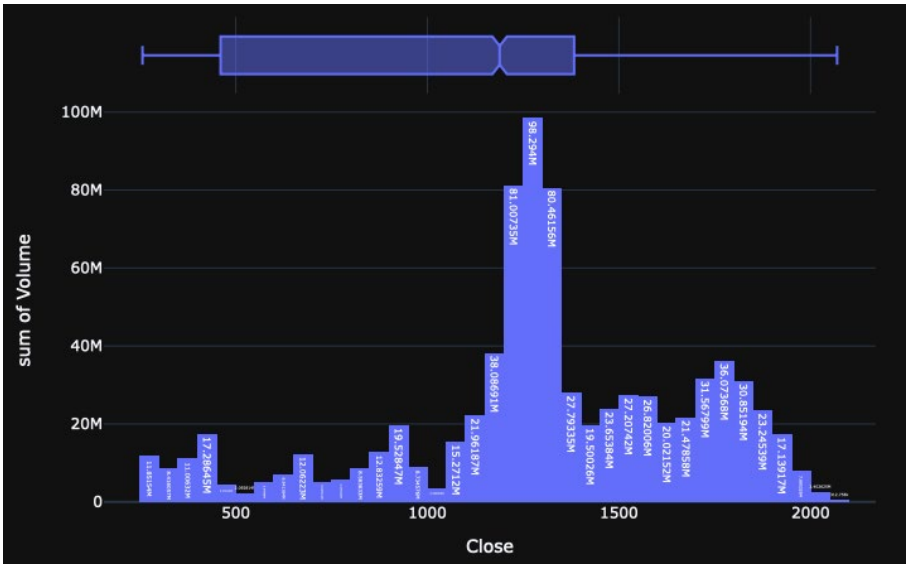

Fig. 3. Lineplot

experienced notable fluctuations.

The histogram shown in the Figure 3 provides an overview of the distribution of gold prices within a specified range. The x-axis represents the range of gold prices, ranging from 256 to 2100 with intervals of 50. The y-axis represents the frequency or count of prices occurring within each respective range.

The most frequent range of gold prices observed is between 1250 and 1300, occurring a total of 98.284 million times. Conversely, the least frequent price ranges are the intervals between 500-549, 1000-1049, and 2000-2099. The frequency of gold prices below 1100 is generally low, with each range having occurrences of no more than 20 million. The range of 1150 to 1350 demonstrates the high- est frequency, averaging around 70 million occurrences. The frequency remains relatively stable between 20 mil- lion and 40 million occurrences within the range of 1350 to 1750. Subsequently, as the price range increases, the frequency gradually decreases until it reaches zero.

3 Methodology

3.1 Data Processing

Data processing involves various steps to transform, manipulate, and prepare the dataset for further analysis or modeling. This includes creating new data structures, normalizing data values, and reshaping the data as necessary. Create a new DataFrame that focuses specifically on the ”close” variable, extracting it from the original dataset. This allows for a more targeted analysis about the trends of gold prices and modeling based on the specific variable. Fits the scaler to the data, learning the minimum and maximum values of the dataset, and then applies the transformation to scale the data accordingly. The scaled data is assigned to the variable scaled data. Creating a training dataset for LSTM model by selecting a subset of scaled data. Prepares training datasets for LSTM and linear regression models. It selects a subset of scaled data and constructs input features and target values. For the LSTM model, the input features are sequences of 68 previous scaled ’close’ prices, while the target values are the next three scaled ’close’ prices. For linear regression, lagged features are created by shifting the ’10’ column. Rows with missing values are dropped. Finally, the index is reset, resulting in a new DataFrame with a consecutive index.

3.2 Train and Evaluation

Regarding gold price prediction models, LSTM models and linear regression models have been built and used to forecast gold prices. In evaluating the performance of gold price prediction models MSE and MAPE are commonly used metrics.

The LSTM model is used for time series prediction and consists of two LSTM layers with 128 units each, followed by two layers which are fully connected. The model is compiled using the Adam optimizer and mean squared error loss function. Once trained, the model can be used for predictions and calculating the average predicted value. The Linear Regression train dataset is divided using a temporal order to create a training set and a test set. The training set is composed of 70% of the complete dataset, while the test set consists of the remaining 30%. Slice operations are performed to select the input features and target variables. Subsequently, the linear regression model is fitted to the training data, and the model is used to make predictions on both the training set and the test set. Finally, the actual target values from the test set are extracted to evaluate the performance of the model.

The evaluation process encompasses two aspects: The performance of predictions is evaluated using the MSE and MAPE metrics. Additionally, the model's capability to accurately predict the directional changes (increase or decrease) in the target variable is assessed.

Using Mean Squared Error and Mean Absolute Percentage Error to evaluate the performance of every model. The ac- curacy of predicting the directional change of gold prices is calculated. This is determined by comparing the differences between consecutive values to identify the trend of increase or decrease. The accuracy is computed by comparing the number of correct predictions with the total number of predictions made.

Table 1. Evaluation Result

LSTM		Linear Regression
Accuracy	50.67%	53.02%
MSE 376.603		381.653
MAPE	20.190	20.353

4 Result

4.1 Predict Result by LSTM Model

As the result shown in Table 1, the analysis of the prediction accuracy regarding the rise and fall of the daily gold price reveals that the LSTM model achieved an accuracy rate of 50.67%. This indicates that the LSTM model had a moderate success rate in correctly predicting the directional changes in the gold price.

The LSTM model has a MSE of 376.603 and the value of MAPE is 20.190. These metrics represent the average squared difference and average percentage difference between the predicted values and the actual values, respectively. The LSTM model achieved a lower MSE and a slightly lower MAPE compared to the Linear Regression model.

4.2 Predict Result by Linear Regression Model

As the result shown in Table 1, the analysis of the prediction accuracy regarding the rise and fall of the daily gold price reveals that the Linear Regression model achieved a slightly higher accuracy rate of 53.02%. Upon comparing the prediction accuracy of the LSTM and Linear Regression models regarding the rise and fall of the daily gold price, it is observed that the accuracy of the Linear Regres- sion model achieved slightly higher. This indicates that the Linear Regression model had a slightly higher success rate in correctly predicting the directional changes in the gold price compared to the LSTM model.

The Linear Regression model has a slightly higher MSE of 381.653 and a slightly higher MAPE of 20.353 compared to the LSTM model. These metrics indicate the average squared difference The Linear Regression model demonstrated marginally lower accuracy and precision in comparison to the LSTM model, as indicated by the mean percentage deviation between the predicted and actual values.

5 Conclusion

In conclusion, the research conducted in this study compared the predictive performance of the Linear Regression and LSTM models for gold price forecasting.

The implications of this research are significant in several aspects. When the primary goal is to accurately predict the long-term trends of gold prices, it is recommended to utilize the LSTM model. Its ability to capture temporal dependencies and intricate patterns can be advantageous for such forecasting purposes. On the other hand, when the focus is on assessing the finer details of gold price fluctuations, such as predicting short-term changes or analyzing specific price movements, the Linear Regres- sion model may be more suitable. Its simplicity and interpretability allow for a clearer understanding of the impact of individual factors on gold prices.

Generally speaking, this research contributes valuable in- sights into the strengths and limitations of the LSTM and Linear Regression models for gold price forecasting. It highlights the need for further advancements in modeling techniques to overcome the identified limitations and emphasizes the importance of incorporating a comprehensive set of influential factors for accurate and holistic gold price predictions.

The results revealed certain limitations associated with both models. One notable limitation is their constraint to predicting prices within a 3-day timeframe, which restricts their usefulness for longer-term forecasting scenarios. This temporal limitation hampers the models' ability to capture and project complex price trends and fluctuations that occur over extended periods.

To overcome these imperfections, future research should prioritize the development of models that address the identified shortcomings that can accommodate longer-term forecasting horizons and integrate a more comprehensive set of influential factors. This could involve enhancing the architecture and training process of the LSTM model to improve its performance in capturing intricate temporal patterns within gold price data. By addressing these shortcomings, researchers can develop more robust and accurate models that support effective gold price prediction and analysis.

References

1. S. Shahriar, and E. Topal. Resources policy, **35**, 3 (2010)
2. T. Naliniprava. International Journal of Economics and Financial Is- sues, **7**, 4 (2017)
3. B. Guha, and G. Bandyopadhyay. Journal of Advanced Management Science, **4**, 2 (2016)
4. M. Georgia, et al. International Journal of Financial Engineering and Risk Management, **2**, 1 (2013)
5. E. Hadavandi, A. Ghanbari, S. Abbasian-Naghneh . *Developing a time series model based on particle swarm optimization for gold price forecasting,*Third International Conference on Business Intelli- gence and Financial Engineering. IEEE, (2010)
6. D. Liu, Z. Li. *Gold price forecasting and related influence factors analysis based on random forest,* Proceedings of the Tenth International Conference on Management Science and Engineering Management. Springer Singapore, (2017)
7. L. Madziwa, et al. Resources Policy, **76**, (2022)
8. B. Li. Computational intelligence and neuroscience, **2014**, (2014)
9. YC. Chen, WC. Huang. Applied Soft Computing, **112**, (2021)
10. S. Verma, et al. IAES International Journal of Artificial Intelligence, **9**, 1 (2020)
11. M. Mohtasham Khani, et al. SN Computer Science, **2**, 4 (2021)