

Gold and Bitcoin Price Prediction based on KNN, XGBoost and LightGBM Model

Ziyang Yuan

Hohai University, Nanjing 210098, China

1908050130@hhu.edu.cn

Abstract. In the past few decades, there has been an increasing demand for assets trading with help of machine learning. Contemporarily, the cryptocurrency and gold market has become prosperous with extremely dramatical fluctuations. This paper aims to study the trading price laws based on machine learning scenarios of Bitcoin and Gold to predict the price of the two currencies. To be specific, this study gives an inside view of the application of a method combined three algorithms (i.e., KNN, XGBoost and LightGBM) to predict the future Gold and Bitcoin price browser based on past data from 2017 to 2022. According to the analysis, the study shows the difference of three models, the accuracy of the combined algorithms and proves the related metrics to predict the price of the Gold and Bitcoin. Overall, these results give a guideline for the investor to make sensible decisions about Bitcoin and Gold price and shed light on guiding further exploration of price forecasting in terms of machine learning approaches.

Keywords: Machine Learning; Price Prediction; KNN Regression; XGBoost; LightGBM.

1. Introduction

Bitcoin is the first proposed Crypto-Currencies which was introduced in 2009 based on blockchain techniques [1]. Instead of controlled by government or any organization, Bitcoin is produced by complex algorithms [2]. As a digital currency, it addresses two critical issues, i.e., creation and avoiding duplication (risk rules) [3]. Nowadays, it has the potential to be more popular since it facilitates the use if a virtual currency, including electronic payments. Gold is a chemistry element in the nature as well as a currency in the global market, i.e., global gold market [4]. There is a high correlation between gold and oils. The gold price in the long-term is nonstationary. Gold price forecasting is significant in the past several years since people buy the gold jewelries and coins [5], investing in gold Exchange trade fund, which reduces the minimize the cost in the annually inflation.

Guha and Bandyopadhyay applied ARIMA (Autoregressive integrated moving average) approach to make predictions of Gold [5]. They collected data about 10 years to mitigate the risk in purchases of gold and found out that the technique was only used for short run, to detect the small variation in the data. Andi used LSTM in machine learning to predict the bitcoin price [6]. With a large dataset of LSTM, the researcher had a successfully results based on the normalization method. Li and Dai in their paper used the CNN-LSTM hybrid model [7], which can significantly boost the accuracy as well as performance of the model. Faghih and Heidari found that the GM (1, 1) type of grey model can also enhance the forecasting accuracy of BTC [8]. Indera and Rizman used the MLP-based NARX, which achieve excellent result for Bitcoin prediction [9]. Aggarwal and Goel looked for various effects affecting the price of BTC including gold price bitcoin price prediction in terms of the metric RMSE based on CNN, LSTM and GRU [10].

The study tries to forecast the prize of Bitcoin and Gold using machine learning model. In Sec. 2, the paper discusses the data collected and visualize these data. In this part, the paper introduces three models and set the parameters. Sec. 3 discusses the result about the three models and analysis the results about the parameters used in this study. In Sec. 4, the limitations and the future outlook of the research are demonstrated. Eventually, a brief summary will be given in Sec. 5.

2. Data & Method

2.1 Data

The study is based on the daily data (closing prices) for CME and XAU ranging from October 2017 to October 2022 collected from the investing.com, which is a stock website provided with all the information globally. The data is shown in Fig. 1. The study chooses the whole data collected from the website as the training data. 1584 valid data of Bitcoin and 1302 valid data of Gold are in the study. The algorithms are adopted in order to optimize the learning rate and regularization parameters, and the upper limit of learning rate was 0.3. The lower limit is 0.1. The proportion of prediction data and training set data is controlled between 10% and 30%. The study sets the minimum sample number of leaf node and the max depth of the tree all equals to 10.



Fig 1. The price evolution of gold and BTC during last 5 years in the unit of USD

Because gold and Bitcoin belong to two different investment assets with different intrinsic values, which leads to different dollar prices in the investment market. In fact, it is observed that the dimensional units of data gold and Bitcoin are different in the time interval between the 5 years, and direct analysis of them will reduce the characteristics. It seriously affects the performance of the model, thus affecting the given investment strategy. Therefore, in order to solve the comparability between the data and make the processed data within the specified interval, the study uses z-score standardization to scale the data, i.e., standard deviation standardization. This study calculates the mean and standard deviation of the original price data of gold and Bitcoin, and standardizes them. In the study, the data after standardization and the difference between two days are set as the independent variables.

2.2 Models & Metrics

K-Nearest Neighbor (KNN) regression algorithm is a model. KNN regression takes advantages in two aspects. First, the model is easy to understand and can be performed well without adjusting too much parameters. Second, the speed of constructing the model is relatively fast. However, the model has a deficiency in the large quantities of data. Additionally, the efficiency can be bad and the speed will be low in the situation that a lot of eigenvalues equal to zero. The study processes the data in preparation to clean the data. The study deletes the index which the price is below 2500 dollars, because they are little change in these intervals.

In essence, XGBoost regression is a Gradient Boosting Descent Tree (GBDT). XGboost regression prediction has a good prediction effect on the huge and nonlinear data sets, and the processing speed is relatively fast. The study uses the XGboost regression prediction to avoid the underfitting or overfitting estimation of the prediction model to the data to train the team and the results. The model uses the cross-validation method to evaluate the results and return the best parameter values. It can find the split-point accurately. Nevertheless, XGBoost requires a large space of memory. It does a

calculation each time when gets into a split-point, which costs a lot of time. In addition, if the time step of grid search is settled too small, more parameter values can be found, but it will consume more computing space. So, the study chooses the previous time step index to get the local optimal solution, and then filter the local optimal solution to adapt to most situations.

Light Gradient Boosting Machine (LightGBM) regression is a framework to implement the GBDT algorithm. LightGBM is a high-effectiveness implement of XGBoost. XGBoost performs well but has disadvantages above. Therefore, the research adopts the advanced algorithm to construct the model in order to better predict the price. It differs from several aspects. First, LightGBM regression uses the Gradient-based One-Side Sampling (GOSS) method to reduce a large quantity of data with little gradient instance, thus saving a lot of time and space when traversing the eigenvalue. Second, it uses the Exclusive Feature Bundling (EFB) method to reduce dimensions by binding some different characteristics. Third, it uses the leaf-wise algorithm because there is always with little benefit to search and divide the leaves. However, LightGBM regression has disadvantages in generating the overfitting model as well as being sensitive the noise. The study uses some symbols to better evaluate the accuracy of the price prediction, i.e., MSE, RMSE, MAE, MPAE and R^2 .

2.3 Model Selection

The three different prediction models have different advantages and disadvantages. XGBoost regression model runs longer than the other two models, has slightly better prediction accuracy and better fitting effect of various nonlinear functions. Whereas under the requirements of the topic, there is only one data per day and other factors are not considered. XGBoost model also has higher prediction accuracy, and the prediction effect decreases with the increase of prediction days in the KNN regression. It should be noted that there are many price influencing factors in the financial investment market in real life, e.g., policy, public opinion, natural disasters. The amount of data is huge and nonlinear, and the prediction accuracy of the single-factor XGBoost regression model is higher than the other two models but takes a long time. For the purpose of predicting the medium and long-term price change trend and making efficient use of time, the study selects and combine the prediction data of the three regression prediction models above to carry out the following research.

3. Results & Discussion

3.1 Correlation Analysis

The Fig. 2 lists the heatmap of the correlation analysis, which helps the study choose different valuable independent variables. Seen from the results, the factors are somehow relevant to the prediction target.

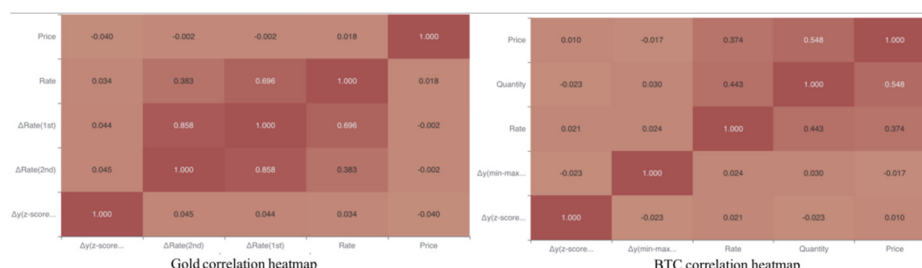


Fig 2. Correlation heatmap of Gold and BTC.

3.2 Model Performance

The Table. 1 summarizes the performance of different approaches on the two underlying assets. According to the analysis, the results of the three statistical methods perform well overall. The prediction results of the three models are presented in Figs. 3, 4 and 5.

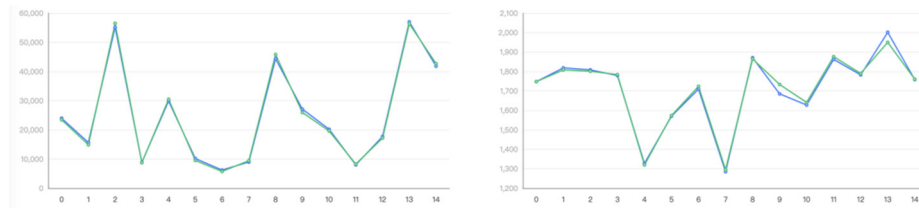


Fig 3. XGBoost regression prediction.

Table 1. Performances of different models for BTC and Gold

Parameters	MSE	RMSE	MAE	MAPE	R ²
KNN(Training_Bitcoin)	812712.888	901.506	574.582	2.06	0.997
KNN(Tesing_Bitcoin)	1114513.201	1055.705	678.045	2.477	0.997
XGBoost(Training_Bitcoin)	89426.766	299.043	172.572	0.601	1
XGBoost(Testing_Bitcoin)	1755292.576	1324.875	806.185	3.174	0.995
LightGBM(Training_Bitcoin)	10444.549	102.199	74.729	7.421	0.874
LightGBM(Testing_Bitcoin)	14043.277	118.504	85.092	7.887	0.834
KNN(Training_Gold)	129.259	11.369	8.435	0.488	0.996
KNN(Testing_Gold)	205.093	14.321	11.104	0.642	0.994
XGBoost(Training_Gold)	1.01	1.005	0.682	0.041	1
XGBoost(Testing_Gold)	13.036	3.611	2.55	0.151	0.996
LightGBM(Training_Gold)	5.782	2.405	0.878	0.052	1
LightGBM(Testing_Gold)	11.041	3.323	2.014	0.119	0.997

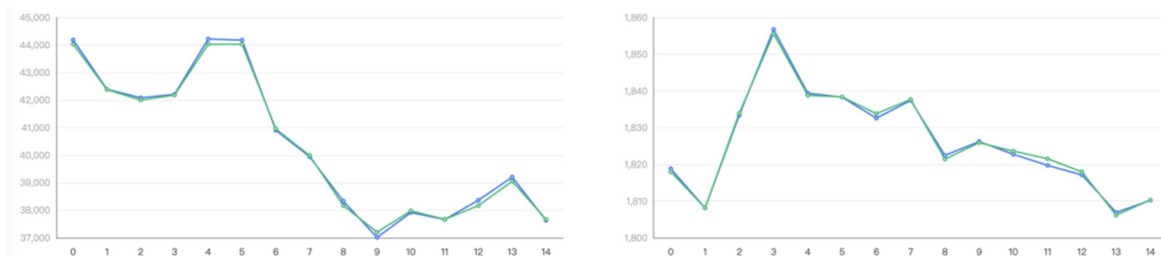


Fig 4. KNN regression prediction

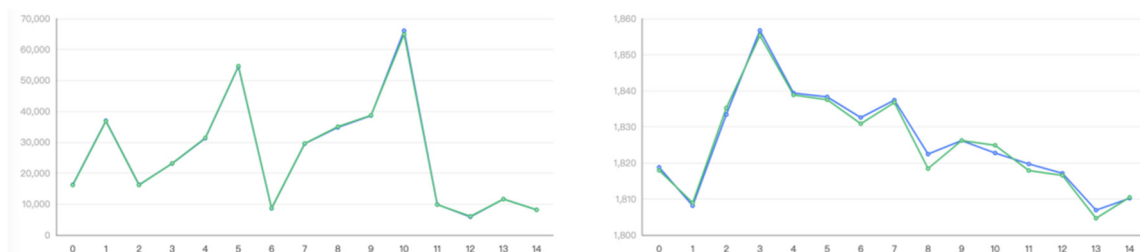


Fig 5. LightGbm regression prediction.

3.3 Analysis

This project uses Python packages including TensorFlow, Seaborn, Pandas, and Numpy to process and evaluate the data. To be specific, the Tensorflow is implemented for the construction of various deep learning scenarios. The Seaborn is used to visualize the date in the research. While the Pandas and Numpy are used to process the data.

Based on the evaluations, the larger the training set, the more susceptible to the overfitting issues. As a result, the use of models has helped researchers to have a better chance to see the regression architectural design from the beginning. Especially, the LightGBM regression operates quickly and effectively with good performance. The study sets the date in the five years, the changing rate of the daily price, the transaction quantities, and the difference between two days as the independent variable. In fact, without the intervene of the outside factors, date is the only variable involved in the study. The date gives an idea to make observation about the price trend focusing only on the date.

4. Limitations & Outlooks

Nevertheless, the results presented in this study have some shortcomings and defects. To be specific, the model only analyzes the price of two currencies in the past half decades and does not consider the factors outside the free trading market. It may substitute a small part of data that does not conform to the market rules into the calculation, which ultimately leads to the inaccuracy of the model. In addition, the relevant algorithms used in the above model need huge data to support, and the more data, the more accurate the fitting results will be. The prediction results may be biased to a certain extent when the data provided are not sufficient. From the discussion and analysis before, one can see that a prediction model usually performs best in a specific aspect, but often has some defects in other aspects. In fact, price trends often change dynamically and frequently, hence a single prediction model cannot fully predict accurate data. Thus, one needs to avoid the limitations of a single model, determine the weight proportion of each prediction model in the portfolio model, and obtain the optimal prediction data under the same conditions.

In the study, the research also takes advantages in these aspects:

- Universality. The model in the study is highly universal. In fact, these models can be applied to any time unit of gold, Bitcoin trading market, or even any freely traded commodity market, just by adjusting the parameters in the training data according to the model.
- Accuracy. In theory, as long as the training data base in machine learning is large enough, these models can get the transaction amount at the target time point relatively accurately.
- Clarity. The models in the study includes specialized concepts such as machine learning, but the basic concepts and properties of the model can be understood in a popular way, and everyone can understand how the model works.
- Longevity. The models in the study are based on previous data and machine learning. In this way, one can get more and more accurate calculation results with the length of time by providing continuous and corresponding time data without modifying the model too much.

Hence, one can apply the advanced and combined model to different currencies and traded assets thanks to these advantages. Further study can forecast the Bitcoin and Gold price more accurately and help investors more possible to make sensible decisions.

5. Conclusion

In summary, this paper investigates Gold and Bitcoin price based on KNN Regression, XGBoost Regression and LightGbm Regression. Specifically, KNN is the easiest to operate and easy to understand. To compare, XGBoost requires a large space of memory. In addition, LightGbm is evolved from the XGBoost. According to the analysis, LightGbm is more effective and space-saving. Contemporarily, more and more market traders frequently buy and sell volatile assets to maximize their total returns. Therefore, it is of great significance and influence to accurately predict the rise and fall of the free trading market based on past trading data and to formulate appropriate strategies for effective trading. In this study, the data set is examined and clarified, performing missing value processing and data normalization to retain reasonable time and price information. Then, this study uses KNN regression, XGboost regression and LightGBM to establish the model of different weight ratio. Next, according to different evaluation indicators, the performances of the three models are

evaluated. Finally, the study makes a conclusion about the accurate prediction. Based on the analysis, using these models will make easier for the prediction of Gold and Bitcoin price, even other properties in the global market. Overall, these results offer a guideline for price prediction based on machine learning scenarios.

References

- [1] Rahouti, M., Xiong, K., Ghani, N.: Bitcoin concepts, threats, and machine-learning security solutions. *IEEE Access*, 6, 67189-67205 (2018).
- [2] Velde, F.: Bitcoin: A primer. Blog, Philippe-Poisie, Eu, 1 (2013).
- [3] Böhme, R., Christin, N., Edelman, B., et al.: Bitcoin: Economics, technology, and governance. *Journal of economic Perspectives*, 29(2), 213-38 (2015).
- [4] Shafiee, S., Topal, E.: An overview of global gold market and gold price forecasting. *Resources policy*, 35(3), 178-189 (2010).
- [5] Guha, B., Bandyopadhyay, G.: Gold price forecasting using ARIMA model. *Journal of Advanced Management Science*, 4(2), (2016).
- [6] Andi, H. K.: An accurate bitcoin price prediction using logistic regression with LSTM machine learning model. *Journal of Soft Computing Paradigm*, 3(3), 205-217, (2021).
- [7] Li, Y., Dai, W.: Bitcoin price forecasting method based on cnn-lstm hybrid neural network model. *The journal of engineering*, 2020(13), 344-347 (2020).
- [8] Faghih, M. J. M., Heidari, H. Predicting changes in Bitcoin price using grey system theory. *Financial Innovation*, 6(1), 1-12 (2020).
- [9] Indera, N. I., Yassin, I. M., Zabidi, A., et al.: Non-linear autoregressive with exogeneous input (NARX) Bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators. *Journal of fundamental and applied sciences*, 9(3S), 791-808 (2017).
- [10] Aggarwal, A., Gupta, I., Garg, N., et al.: Deep learning approach to determine the impact of socio-economic factors on bitcoin price prediction. In 2019 twelfth international conference on contemporary computing (IC3) (pp. 1-5). IEEE (2019, August).