PAPER • OPEN ACCESS

Prediction of gold price with ARIMA and SVM

View the article online for updates and enhancements.

Prediction of gold price with ARIMA and SVM

D Makala and Z Li

Department of Computer Science, China University of Petroleum (East China), Qingdao 266580, china.

Email: lizongmin@upc.edu.cn

Abstract. Gold has become more popular as well as very useful commodity in terms of investment. Gold has been used as a national reserve for many years, and that makes it very crucial in the economics of any country. Most of the investors running to gold as a safe area from uncertainty and political chaos. Determining of the price movement of gold helps the investors in focus in their investments, government to make correct decision about economy since Gold price is a key element is world economy. For the purpose of predicting the price of gold, this article research uses ARIMA and SVM model in prediction. The study uses the daily data from world Gold Council from 1979 to 2019 in analysis. The data up to 2014 are used for the training of the models and the rest are used validation. The study results show that the SVM is better one compares to ARIMA using the performance measurement tools of RMSE and MAPE by having RMSE of 0.028 and MAPE of 2.5 for the SVM and 36.18 and 2897 for ARIMA respectively. The results suggest SVM to be used in prediction of any commodity price due to his high accuracy.

1. Introduction

One of the most important minerals in the world is gold. Despite making valuable commodities, gold acts as a reserve in any country. A gold reserve is an amount of gold held by the central bank of any country for the purpose of the guarantee to be used to pay or trading in the world market and hence increase the country economically. Amongst all mineral in the world, gold is the most popular selection for the investment [1].

The price of gold is affected by different factors, thus making the movement of price to be unstable. These factors include inflation rate, demand and supply, and political issue among others. Inflation as one of the signs of economic growth, when it increases it is obvious pushes the gold price higher, while when having a low supply of any commodity, the price of that commodity increases. moreover, when countries fear the value of the dollar will be fallen since the dollar is the world's marketing currency, the gold price will eventually increase since many demands for gold will be available. Because of its importance, other literature has termed a safe haven during financial crises [2][3].

Therefore, the price of gold moving up and down and very uncontrollable. Figure1 shows the movement of the gold price for the period of about 41 years thus from 1979 to December 2019. Nevertheless, the gold price can be forecasted ahead [4], and that makes possible to make the future decision. The movement of the gold price is time-series order means changing with time, therefore doing the forecasting with such kind of data has been challenging for a while until the application of machine learning and deep learning introduced in the game of economics and statistics.

Figure 1. Showing the price movement from 1979 to 2019. [Source Author Computation]

In this study article, ARIMA models and SVM model have been deploy in determine the future price of the gold. Arima is one of the most traditional models and widely used method. This model is compared with the scientific one of SVM which is one of supervised machine learning technology.

The study is organized in sections where as in section 1, introduction of the article has been explained well as shown above. The rest of the paper has been arranged as Section2 explain and discuss about the past similar works, Section3 deals with methodology used and Dataset, Section 4 explain about the results and discussion of the results, conclusion and recommendation can be seen in section 5.

2. Literature Review

Prediction analysis has become more challenge due to increasing of availability of data which are not stable. Researchers and academicians are still working on the best way in finance and economics to conquer this challenge. Banhi and [5] in their research paper analysis uses Arima model to make forecasting of future gold price in India. Based on the data from November 2003 to 2014 they conclude the Arima model performed well and it is the best when forecasting short time range.

In addition [6] studied the forecasting of gold prices in India by using ARIMA model using the data from July 1990 to Feb 2015. This study suggested that ARIMA (0, 1, 1) is the best to be used in forecasting gold prices in India since the ARIMA (0, 1, 1) has the least value of RMSE, MAPE, and MAE. In this study, they only used one method of ARIMA with different parameters. The comparison with other model did not occur.

Arima can be used in prediction of livestock product as in 2019, Joseph [7] uses Arima technique to predict the livestock product consumption in Tanzania. In their research analysis their findings shows that consumption of all livestock product will increase and on top of that they concluded that the demand for animal feed should be expected soon.

Hossain, abdulla and Zakiri compares the accuracy of forecasting jete production in Bangaladesh. They compared the two model Arima and ANN. The result of their paper shows that ANN performes better than the Arima model [8]. In similar manner of comparing the models, Ayodel and others use New York Stock Exchange data to compare Arima and ANN, and the result of their study clearly indicate the superior of ANN over the Arima Model [9].

Generally the past studies suggest that when Arima model is compared with the AI models such as deep learning and machine learning models, Arima model become very weak as explain more this previous work[10][11][12].

With growth of technology, research and academician have engaged themselves in using the better way to analyses and finding solution using the more technological means. In prediction analysis some researcher have done some research on the performance of the SVM in forecasting. One of the area is prediction of football results. [13]develop a model to predict the football match results in his findings and results he observe that only 53.3 prediction accuracy can be obtained from SVM. He concluded by suggesting that SVM is not good enough for prediction of match results.

Again Minglei et al [14], did perform the analysis of energy consumption of hotel building. The MSE value of their result agrees with many studies that SVM is one of the best techniques in prediction by having MSE of 2.2% and R-square of 94%.

Akash et al [15], use the data from Perth mint of Australia to predict gold price using SVM and ANFI. with the assist of performing tolls mechanism of MAE, RMSE and MAPE, it was observed that SVM performed better than ANFI. In addition to that [16] explain more about the application of SVM in different aspect such as medical.

The literature also suggest also about the importance of measuring the performance of the forecasting with different methods.[17][18] explain different about the measurements and the important of them in forecasting. Also [19] said that the good performance is vital aspect to the business and can have critical impact to an organization.

3. Data and Methodology
3.1 Dataset
This study is aimed at predicting the future price of gold using deep learning technology. Data used in this study research are daily gold price, that can be retrieve from the World Gold Council [20]. The prices from the World Gold Council are indicated as per troy ounce. The dataset consists the daily prices from the January 1979 to December 2019, thus makes the total number of years of observation be 41.

The cleaning of data follows, to make sure all the data are in proper manner and well arranged. Missed data are filled by the average value of the pas three days. After that, the split of data is done. Here in the step the dataset is divided into training data and testing data. the testing data is for training the model while the testing data is used to validate how the best, the model has been trained. This research study aims on using ARIMA method and SVM in determine the forecasting of the gold price, and therefore there is need to know these models.

3.2 ARIMA
Time series prediction has been using different methods in finding the best way of forecasting especially in business and econometrics. One of the best ways, is application of ARIMA model in forecasting. ARIMA stands for Autoregressive Integrated Moving Average. ARIMA model begins few years back when statisticians and economist analysed timeseries trend without considering the non - stationarities of data and they have impact in prediction results. George Box and Gwilym Jenkins came up with Forecasting and control in which they explain how nonstationary data can be stationary.

ARIMA is one of the popular model used in econometric and statistical analysis of the time series data. this model is mostly and widely used when dataset involve shows to be non-stationarity. This model is also known as Box- Jenkins method. Arima mode is presented as ARIMA (p, d, q) when there is seasonality involved. When there is seasonality it is denoted as ARIMA$(p,d,q)(PDQ)m$. p is the number of number of autoregressive also known as lag order, d is the number of non-seasonality difference, q is the number of moving average times and m is number of period in each season 2 3. When two value of the (p,d,q) become zero, the ARIMA model become either AR, I or MA. Example when the value of is $(p,0,0)$ then ARIMA becomes equal to AR(p).

When using Arima model in forecasting, it is very important to determine the ARIMA model to be used. this means that the value of pdq should be determined and computed. The model is uses fit() function in training the model. One thing to remember is that, the timeseries should be stationary when computing. Different method can be used to check the stationarity but in this research, study will use

the Augmented Dickey Fuller test. More to that, The Box-Jenkins Model forecasts data using three principles, autoregression, differencing and moving average. These three principles are known as p, d, and q respectively. Each principle is used in the Box-Jenkins analysis and together they are collectively shown as ARIMA (p, d, q) [21][22].

3.3 SVM

SVM is a supervised learning technique that simply analyses data for classification and regression which was introduced by vapnik and Boser in 1992 [23]. SVM has done good work in categorization of text, image classification and it has widely used in health and biological science [24]. In recent years it has become more popular in prediction analysis in econometric and economic studies. As explain earlier, the SVM model is based on linear regression. In this research article, three types of kernel is deployed for analysis, Linear kernel, poly kernel and RBF kernel. Given w as weight vector, $\varphi(x)$ as mapping function and b is bias SVM mathematically can be presented as:

$$f(x) = w^T \varphi(x) + b \tag{1}$$

3.4 Performance Measurement Tools

Determining the best forecasting model is a crucial part since it indicate how best the prediction is. Someone has to know if the prediction is good or not. Forecasting is an accurate prediction. Normally the good forecasting is determining by comparing the prediction value and the true values. Different tools have been discuss by different scholars in determining the performance of forecasting. This study uses RMSE and MAPE in determining the performance.

RMSE means Root Mean Square Error is a standard derivation of the prediction error. It measures how far the regression line is best fit. This means that RMSE shows how concentrated data is around the best fit line.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predict - true\ Valuel)^2}{n}} \tag{2}$$

MAPE is Mean Absolute Percentage Error. This metrics is one of the mostly applied metrics used to compare and measure the performance of forecasting. It measure the performance accuracy by calculating the average absolute percentage error minus actual value divided by actual value. Thus

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Actual - Forecast}{Actual}\right| \tag{3}$$

Therefore, determining the best performance measurement to use depend on the person who is performing the analysis. Each measurement tools can be better if applied properly and can lead to error if wrongly computed. Therefore, research and analyst can choose which measurement tool to be used as the wish and depend on what they want to address. This research article opts to use RMSE and MAPE due to the most common applied metrics. More to that this research has add one more metric of R-square which shows by how much percentage the forecasting resulting are true or similar to the true value of the data existing. If the forecasing results are exactly as the actual values then the R-Square of the metric should be 100%.

4. Results and Discussion

Let discuss the results in two way since two models are involved, let this article explain in similar manner. **Table 2**, shows the performance matrix of the two models, however the SVM model has been tested with the different kernel to obtain the best fit model.

Table 1. Showing performance of the models

MODEL	MAPE	RMSE	R-Square
Arima (2, 1, 2)(2, 1, 2, 12)	2897.59	36.1795	0.8615
SVM (Linear)	2.49	0.0276	0.9978
SVM (Poly)	2.49	0.0275	0.9978
SVM(RBF)	169.3	10.8773	0.9875

Fist le discuss about Arima results. One of the key elements in Arima model is to ensure the time series data are in stationary. Different techniques have been deployed to determine and making the dataset stationary. This study will use Augmented Dickey-Fuller test to determine the stationarity state of the dataset. The test has Hypothesis that state:

1. Null Hypothesis H_0: There is unit root (non stationary)
2. Alternative Hypothesis H_1: There is No unit root (stationary)

When P-Value is greater than critical value (mostly 5%), the Fail to reject Null Hypothesis. And when P value is less of equal to 0.05, reject the Null hypothesis. From the table 1, it shows that the value of the P is 0.099 which way larger than the values of critical values at all significant level. More to that the ADF statistic is large than the value of critical levels. For that reasons this study Fail to reject null hypothesis (in simple term accept the null hypothesis), that say the times series data has a unit root and it is non stationary.

There the need to make stationary is critical. Here we do first differential and check if it is stationary. If not the second differentiation is performed. The number of differentiations also referred as lag term in Arima (p, d, q) number of differentiations is d. after the first differentiation the results shows P-value become 0.00, which implies that the study reject null hypothesis and accept alternative hypothesis which state that the data series has no unit root and data are stationary.

Table 2. Showing the ADF results at Level and First differentiation

		Level (d=0)	First different (d=1)
ADF Statistic		0.939174	-17.382101
P-Value		0.993581	0.000
Critical Value	1%	-3.431	-3.431
	5%	-2.862	-2.862
	10%	-2.567	-2.567

The Arima model parameter has to be found. Here in this study automatic function in python platform where by the best fit Arima mode is chosen based on the smallest AIC. In our research the best fit Arima is **ARIMA(2,1,2)x(2,1,2,12)**. According to the results as shown in table 2, Arima model produces the RMSE of 36.179, and MAPE of 2897.59, which implies not bad. This can also be proven by the R-square where by the forecasting line is of 86.147% fit. Figure 1 shows the model results.
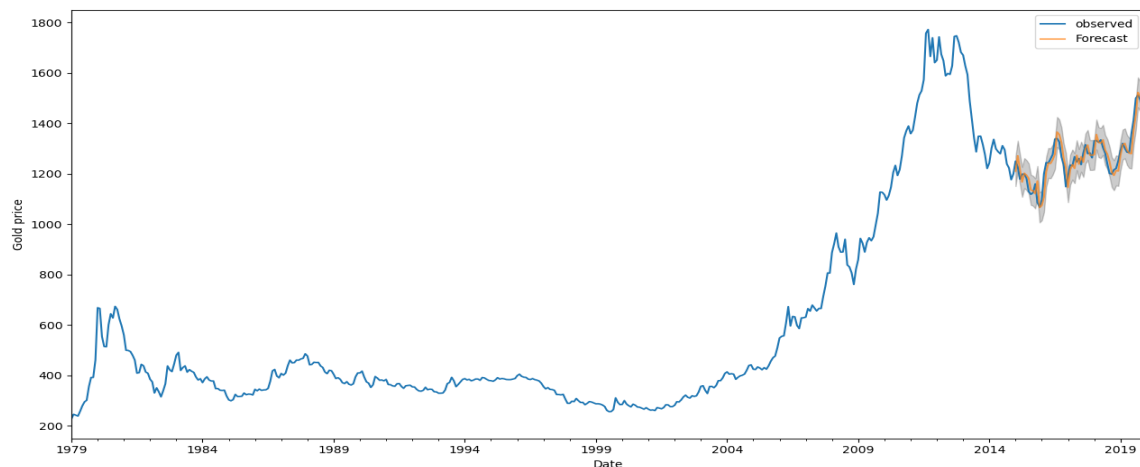
Figure 2. Showing the price prediction with Arima model

Supporting Vector Machine said to be one of the best models of the machine language. We introduce the training data in the model with aim to forecasting the gold price for a period of 2015 to 2019. Here in this model three kind of kernel are used and determine which is better by comparing with each other and Arima model. the results of SVM can be seen in **Figure 3** and also the performance measurement can be observed in **Table 2**.
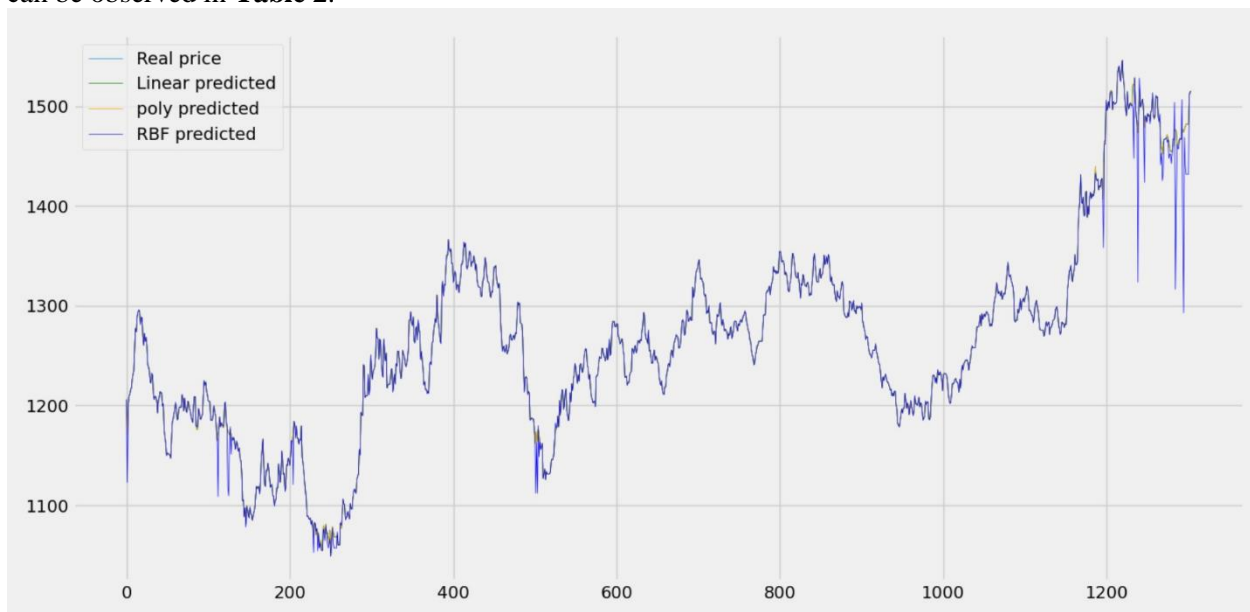

Figure 3. Showing the price prediction for SVM using Linear, Poly and RBF kernel

The three kernel used in forecasting are linear , where by linear uses tolerance for stopping criteria of 1e3, SVM poly uses degree of 1.9 and SVM RBF uses gamma of 0.15. the results can be seen in Table 3 where by the poly shows to performed much better than the three model by having 2.49 MAPE and RMSE of 0.00275. more to support that R-square of the SVM poly almost match the true value of the gold price of the period of 2015 to 2019. However the SVM linear did well similar to SMV(poly) .

From the above results it clear indicates that the SVM-poly performs very well compares to the other SVM and Arima. This result agrees with many researches about the better way of forecasting is use the machine learning techniques. The SVM can be applied to predict any other commodity's price as well as stock marketing price as well as temperate and weather forecasting.

4.1 Performance Measurement Results

To simplify the performance of the three method which involve two model can be seen using the performance metrics applied by this research as indicated in table1. Despite of using two method SVM and ARIMA, in SMV the research applies three different kernels to get what kernel is suitable of the gold price prediction. RMSE measure the accurate of forecasting where by the small value of RMSE shows the best forecasting. In this research results ARIMA has 36.2 and SVM-linear and SVM-Poly has RMSE of 0.027 each and lastly SVM-RBF has 10.87. the above performance clearly indicates SVM linear and poly has performed better using RMSE justification. Similar, using MAPE which measure the means percentage error, indicate that SVM poly and linear has very small result of 2.49, and the small value the better results. The results of this research is also supported by R square which shows how much the forecasting are similar to the actual results. This metrics shows SVM poly and Linear having 99% equivalent to the real value.

5. Conclusion

In this research work article, we have found out how to predict the gold price using the Arima model and SVM. The results of this study reveal that SVM(Poly) is found to performs much better compared to the other SVM(RBF) and Arima model. Support Vector Machine provides more promising choice to be used in forecasting and prediction analysis. However in this research SVM(linear) and SVM (poly) results are almost similar as shown by having 99% of R-square. Currently the machine learning has been widely used in finance and economics such as detection of fraud, customer service (robot chatting) and prediction. This research opens new room for scholars and academician to discuss and suggest the better way to improve the scientific model in forecasting.

References

[1]   Gangopadhyay K, Jangir A and Sensarma R 2016 Forecasting the price of gold: An error correction approach *IIMB Management Review* **28** 6–12

[2]   Baur D G and McDermott T K 2010 Is gold a safe haven? International evidence *Journal of Banking & Finance* **34** 1886–98

[3]   Anand R and Madhogaria S 2012 Is Gold a 'Safe-Haven'? - An Econometric Analysis *Procedia Economics and Finance* **1** 24–33

[4]   Khan M M A 2013 Forecasting of Gold Prices (Box Jenkins Approach) *International Journal of Emerging Technology and Advanced Enginreeing* **3** 662–70

[5]   Bandyopadhyay G 2016 Gold Price Forecasting Using ARIMA Model *Journal of Advanced Management Science* 117–21

[6]   Tripathy N 2017 Forecasting Gold Price with Auto Regressive Integrated Moving Average Model *International Journal of Economics and Financial Issues* **7** 324–9

[7]   Mgaya J F 2019 Application of ARIMA models in forecasting livestock products consumption in Tanzania ed F Yildiz *Cogent Food & Agriculture* **5**

[8]   Md. Moyazzem Hossain, Faruq Abdulla Z H 2017 Comparison of ARIMA and Neural Networks to Forecast the Jute Production in Bangladesh *Jahangirnagar University Journal of Science*

[9]   Adebiyi A A, Adewumi A O and Ayo C K 2014 Comparison of ARIMA and artificial neural networks models for stock price prediction *Journal of Applied Mathematics* **2014**

[10] Makala D and Li Z 2019 ECONOMIC FORECASTING WITH DEEP LEARNING: CRUDE OIL *MATTER: International Journal of Science and Technology* **5** 213–28

[11] Soy Temür A, Akgün M and Temür G 2019 Predicting housing sales in turkey using arima, lstm and hybrid models *Journal of Business Economics and Management* **20** 920–38

[12] Weytjens H, Lohmann E and Kleinsteuber M 2019 Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet *Electronic Commerce Research* 1–21

[13] Igiri C P 2015 Support Vector Machine – Based Prediction System for a Football Match Result

[14] Shao M, Wang X, Bu Z, Chen X and Wang Y 2020 Prediction of energy consumption in hotel buildings via support vector machines *Sustainable Cities and Society* **57** 102128

[15] Dubey A D 2016 Gold price prediction using support vector regression and ANFIS models *2016 International Conference on Computer Communication and Informatics, ICCCI 2016* (Institute of Electrical and Electronics Engineers Inc.)

[16] Kuusisto F, Costa V S, Nassif H, Burnside E, Page D and Shavlik J 2014 Support vector machines for differential prediction *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 8725 LNAI (Springer Verlag) pp 50–65

[17] Mehdiyev N, Enke D, Fettke P and Loos P 2016 Evaluating Forecasting Methods by Considering Different Accuracy Measures *Procedia Computer Science* vol 95 (Elsevier B.V.) pp 264–71

[18] Ouenniche J, Xu B and Tone K 2014 Forecasting models evaluation using A slacks-based context-dependent DEA framework *Journal of Applied Business Research* **30** 1477–84

[19] Klimberg R K, Sillup G P, Boyle K J and Tavva V 2010 Forecasting performance measures - What are their practical meaning? *Advances in Business and Management Forecasting* **7** 137–47

[20] Anon Goldhub | The Definitive Source for Gold Data and Insight

[21] Kamil R ;Janisha A 2018 Demand sensing using deep learning *International Research Journal of Engineering and Technology (IRJET)* **05** 3041–7

[22] Gordon Scott C 2019 Box-Jenkins Model Definition

[23] Vapnik V N 2000 *The nature of statistical learning theory* (Springer)

[24] Huang S, Nianguang C A I, Penzuti Pacheco P, Narandes S, Wang Y and Wayne X U 2018 Applications of support vector machine (SVM) learning in cancer genomics *Cancer Genomics and Proteomics* **15** 41–51