Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/100121

Time Series Forecasting Using Tree Based Methods

EL Houssainy A. Rady, Haitham Fawzy and Amal Mohamed Abdel Fattah*

Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Cairo, Egypt.

Received: 24 Feb. 2020, Revised: 23 Apr. 2020, Accepted: 3 May 2020
Published online: 1 Mar. 2021

Abstract: This paper aims to use the tree-based methods for time series data forecasting and compare between Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT) and ARIMA model to predict monthly gold prices. The time series data for the monthly gold prices was used during the period from Nov-1989 to Dec-2019, which represents 362 observations. ARIMA, DT, RF, and GBT models were fitted based on 90% of data as training set. Then, their accuracy was compared using the statistical measure RMSE. The results indicated that RF was better than DT, GBT and ARIMA (0,1,1) in predicting future gold prices, based on RMSE= 38.52.

Keywords: ARIMA, Decision Tree, Random Forest, Gradient Boosted Trees, ,Machine Learning, Tree Based Methods, Forecasting, Time Series.

1 Introduction

Machine learning (ML) is a branch of artificial intelligence which is related to the development of algorithms that can be learned and adapted based on data. Its name distinguishes it from human learning. It is a relatively new field that has evolved in the computer science community[1]. Statistics and machine learning share much work, but statistics is an older field that has evolved in the field of mathematics. The main difference between machine learning and statistics is the assumption about the data. In statistics, we usually assume that data model is the true data generating process and try to estimate its parameters. In machine learning, we often assume that the data has been generated from some unknown data generating process and different learning algorithms are used to approximate it. Thus, statistics addresses models, while machine learning deals with learning algorithms or procedures [2]. Supervised, unsupervised and reinforced learning are the general forms of machine learning. Tree-based methods are considered as one of the best and the most used supervised machine learning methods. Tree based methods provide higher accuracy and stability and enhance improved interpretation of results for predictive models. They are better than linear models in mapping non-linear relationships and are highly adaptable in solving both classification and regression problems. Time series analysis and forecasting have been a dynamic research area over the last few decades. Different types of forecasting models have been developed, and researchers have relied on statistical techniques to predict time series data. One of the most popular traditional approaches used to analyze stochastic time series is the Autoregressive Integrated Moving Average (ARIMA). It is commonly used due to the ease of understanding and interpreting the resulting models. The main assumption in the implementation of this model is to consider the time series to be linear and following a particular known statistical distribution like the normal distribution. However, time series data are often full of nonlinearity and irregularity, such as economic and financial time series. To address this, Tree based methods can be used as a modern technique to overcome the problems of forecasting non-linearity and non-stationary time series data. In this study, we will apply Decision Tree (DT), Random forest (RF) and Gradient Boosted Trees (GBT) as modern methods of forecasting techniques and see how they could be used as an alternative method to traditional methods. In this study, some comparisons among ARIMA, DT, RF and GBT methods will be performed. Python software was used for building the best model for forecasting and comparing the results of these techniques to determine the best one. These models were applied particularly in this study because comparing tree-based methods and statistics models, focuses on data classification more than on time series data forecasting. In particular, it focuses on extracting patterns and anomalies from data sets.

* Corresponding author e-mail: haitham.fawzy@yahoo.com

The rest of the study is organized, as follows: Section 2 covers the related studies that have been conducted in this study. Methodology is presented in Section 3. Results are presented in Section 4. Conclusion is presented in Section 5.

2 Related Works

The search for efficient time series forecasting techniques is profound in literature that is motivated partly by the dynamic nature of the problem as well as the need for better results. [3] compared the performance of four different machine learning methods, including Neural Network (NN), Decision Tree (DT), Naive Bayes (NB), and k-nearest neighbors (KNN) in stock market prediction. The study predicts variations of market index within the next 1, 6, and 21 days. In the mean times, the study claimed that although NN models have high accuracy, many of them are not put into practice due to the inability of neural networks to explain its reasoning. For the results, DT with 3 classes is considered the best model in the study. [4] compared different machine learning methods for forecasting time series from the M3 competition. The study showed better predictive performances using multilayer perceptron and Gaussian process methods. However, the study did not compare machine learning methods with traditional approaches, such as exponential smoothing or ARIMA. [5] compared the performance of two machine learning methods: Artificial Neural Networks (ANNs) and Support Vector Machine (SVM). Moreover, the Box-Jenkins Approach and the Autoregressive Integrated Moving Average (ARIMA) model were utilized to predict the demand for consumer products in modelling time series data based on modelling accuracy. The study revealed that the SVM method had a better forecast accuracy (in terms of MAPE) than ANN and ARIMA. [6] used decision tree and support vector regression (SVR) to develop a forecasting model for forecasting gold prices based on past historical prices of gold. The results showed that the decision tree takes less time to process the data and has less mean square error than the SVM. [7] developed a model for forecasting gold price using sample data in US$ per ounce from January 02, 2003 to March 1, 2012. Data was generated until January 02, 2012 to develop the model, whereas the rest the of data was used in forecasting the gold price and determining the accuracy of the model. The study used Box-Jenkins methodology for building ARIMA model and suggested ARIMA (0, 1, 1) being the most suitable model to forecast the gold price. Forecasting accuracy, Root Mean Square Error, Mean Absolute Error, and Mean Absolute Percentage Error were calculated to choose the best model. [8] elaborated on a study that used nine types of infectious disease data collected by a national public health surveillance system in China to evaluate and compare the performances of two decomposition methods of regression and exponential smoothing, ARIMA and support vector machine (SVM). The 2005 to 2011 and 2012 data were used for modeling and forecasting the samples respectively. Their performances were evaluated according to mean absolute error (MAE), mean absolute percentage error (MAPE), and mean square error (MSE) and the accuracy in forecasting future epidemic disease that proved their effectiveness in epidemiological surveillance. Although the comparisons found that no single method to be superior, the current study has emphasized that the SVM outperforms the ARIMA model and decomposition methods in most cases. [9] compared several statistical methods including ARIMA, naive, exponential smoothing, and theta, among others with machine learning methods, including different types of neural networks, the nearest neighbors method, a decision tree, support vector regression and Gaussian processes. The results suggest that most of the statistical methods outperform machine learning methods for univariate time series forecasting.

3 Proposed Methodology

In this section, the dataset and the proposed methods to predict future values for the same time series of the monthly gold price are introduced. First, the dataset will be introduced, then predictive models will be explained.

3.1 Dataset

The dataset used in our study is based on monthly gold prices. It was obtained from the index Mundi website www.indexmundi.com from Nov-1989 to Dec 2019.This means that we have 362 observations for the monthly gold price. To evaluate the out-of-sample forecasting ability of the various models, some observations at the end of the sample period are not used in estimating the models. Thus, there are two periods in the analysis: a training series which is composed of 90% of the dataset in the period (Nov-1989 to Nov-2016) and a test series which contains the remaining 10% of the set in the period (Dec. 2016 to Dec. 2019). Table 1 represents some descriptive statistics of the training series of a monthly gold price.

3.1.1 Descriptive Statistics for the Monthly Gold Price

Table 1 shows that the mean of the training series equals 673.95, the median of the time series is 393.06, the maximum number of gold price is 1772.14 per month, the minimum number of gold price is 256.08 per month, and the standard deviation of time series is 451.01. It also shows that the Skewness for gold price is small (=0.96) and close to zero which implies that the distribution of the data is approximate normal. Moreover, Kurtosis is a negative coefficient of -0.53, and the number of observations is 325.

Table 1: Descriptive Statistics for Training Series of the Monthly Gold Price

Statistics	Value
Observations	325
Min	256.08
Median	393.06
Mean	673.95
Max	1772.14
Standard deviation	451.01
Skewness	0.96
Kurtosis	-0.53

Table 2 represents some descriptive statistics of the testing series of a monthly gold price. The mean of the test series equals 1302.4, the median of the time series is 1283.04, the maximum number of gold price is 1510.58 per month, the minimum number of gold price is 1157.36 per month, and the standard deviation of time series is 90.93. Also, the skewness for gold price is small (=1.03) and close to zero which indicates that the distribution of the data is approximate normal. Moreover, kurtosis is a positive coefficient of 0.45, and the number of observations is 37.

Table 2: Descriptive Statistics for Test Series of the Monthly Gold Price.

Statistics	Value
Observations	37
Min	1157.36
Median	1283.04
Mean	1302.4
Max	1510.58
Standard deviation	90.93
Skewness	1.03
Kurtosis	0.45

3.1.2 Data Plot

Fig.1 depicts the price of gold on the y-axis against the equally-spaced time intervals (i.e. months) on the x-axis. It is used to evaluate patterns, knowledge of the general trend, and data behavior over time. The positive trend of data is clear.

3.2 Predictive Models

3.2.1 ARIMA models

ARIMA model is represented by three parameters: p order of an autoregressive component (AR), d order of differencing, and q order of a moving average component. ARIMA model takes historical data and decomposes that data into an

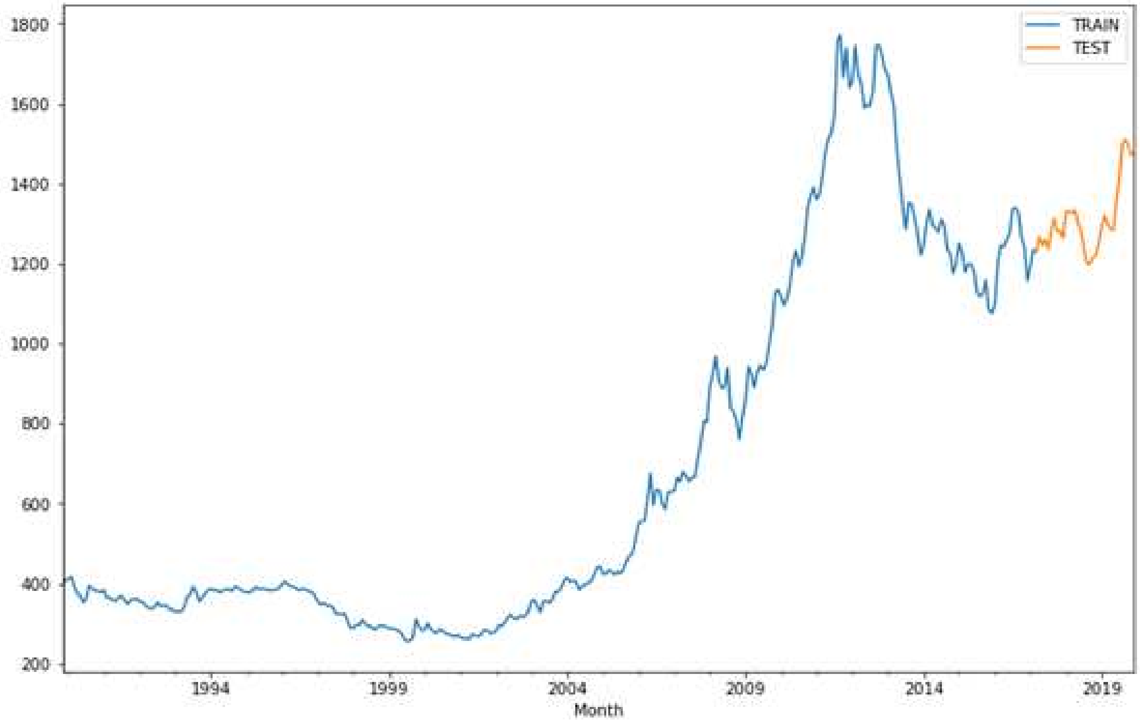

Fig. 1: Time Series Plot of the Monthly Gold Price - US Dollars per Troy Ounce 1989 to 2019

autoregressive (AR) process which maintains memory of past events, an Integrated (I) process which makes data stationary for easy forecast and a moving average (MA) process of forecast errors. The p order autoregressive process, AR(p), is simply the linear relationship between a dependent variable and its own lag [10]. The AR(p) can be expressed as:

$$y_t = \Phi + \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_p y_{t-p} + \varepsilon_t \tag{1}$$

where p is the lag order and t = (1,2,....,n)

Similarly, the q order moving average process, MA(q), can be expressed as:

$$y_t = \Phi + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - ... - \lambda_q \varepsilon_{t-q} \tag{2}$$

Where q is the lag order of the error term ε_t.

Combining the AR(p) and MA(q) models, we can express an ARMA(p,q) model as:

$$y_t = \Phi + \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_p y_{t-p} + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - ... - \lambda_q \varepsilon_{t-q} \tag{3}$$

The ARMA model assumes that the time series data is stationary (that is statistical properties of data do not change over time). But usually the real data are not stationary in nature. Differencing process is usually used to make time series data stationary. The first order differencing process of time series y_t is defined as $\triangle y_t = y_t - y_{t-1}$. For example, if y_t is non-stationary series, we will take a first-difference of y_t , so $\triangle y_t$ becomes stationary. Then, the ARIMA (p, 1, q) model is:

$$\Delta y_t = \Phi + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + ... + \theta_p \Delta y_{t-p} + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - ... - \lambda_q \varepsilon_{t-q} \tag{4}$$

*** The Box-Jenkins Approach**

Box-Jenkins is a set of approaches for time series analysis and for finding out the best fit for ARIMA models. This includes four stages in building and assessing Box-Jenkins time series model [11] . Fig. 2 shows the four stages of modeling according to this approach.

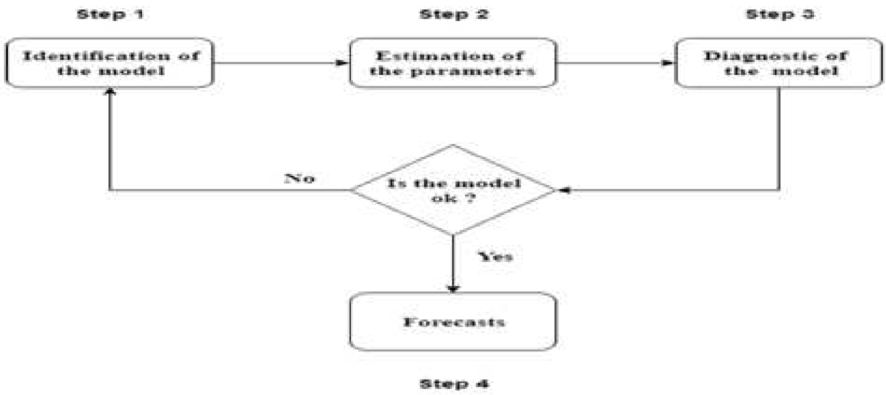

Fig. 2: Box-Jenkins Modeling Approach

• **The Box-Jenkins Approach to Fitting ARIMA Model for the monthly gold price**

1. identification:

The first step in developing a Box-Jenkins model is to make sure that the series is stationary and identifying seasonality in the series, and using the plots of the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) of the series to find out the appropriate values of p, d, q of the order of general ARIMA model. Fig.1 indicates that there is a general tendency in the data, so the time series is not a stationary. For further investigation of the stationarity of the time series, Augmented Dickey-Fuller (ADF), Dickey-Fuller (DF), Philips-Perron (PP), and Kwiatkowski-Philip-Schmidt-Shin (KPSS) tests were applied for the training series. The null hypothesis of ADF and PP tests is that the data are non-stationary. Therefore, large p-values indicate the data non-stationarity while small p-values imply the reverse. Using the usual 5% threshold, differencing is required if the p-value is greater than 0.05. The popular unit root test KPSS has reversed hypotheses, so its null hypothesis assumes the data are stationary. In this case, small p-values (e.g., less than 0.05) indicate that differencing is required. The results of applying KPSS, ADF, PP, and DF tests are shown in Table 3.

Table 3: P-values for KPSS, ADF, PP, and DF Tests for Training Series.

Test	ADF	DF	KPSS	PP
P-Value	0.91	0.81	0	0.91

As shown in Table 3, the p-value of KPSS test is 0.00 which is less than p = 0.05. On the other hand, the p-value of ADF test is 0.91, DF test is 0.81, and PP test is 0.91 which are greater than p = 0.05. The results indicate that the training series for the monthly gold price is not stationary. Fig.3 is also useful for identifying non-stationary time series where ACF and PACF of the data decreases slowly which indicates non-stationarity of data [12]. All the above results and plots confirm that the training series data are not stationary.
As shown in Fig.3, the identification of AR model is often best done with PACF, while MA model can be best identified with ACF. ACF has significant autocorrelations at lag 1 which leads to MA (1). The PACF plot shows definite significant values at lag 1 that leads to AR(1).

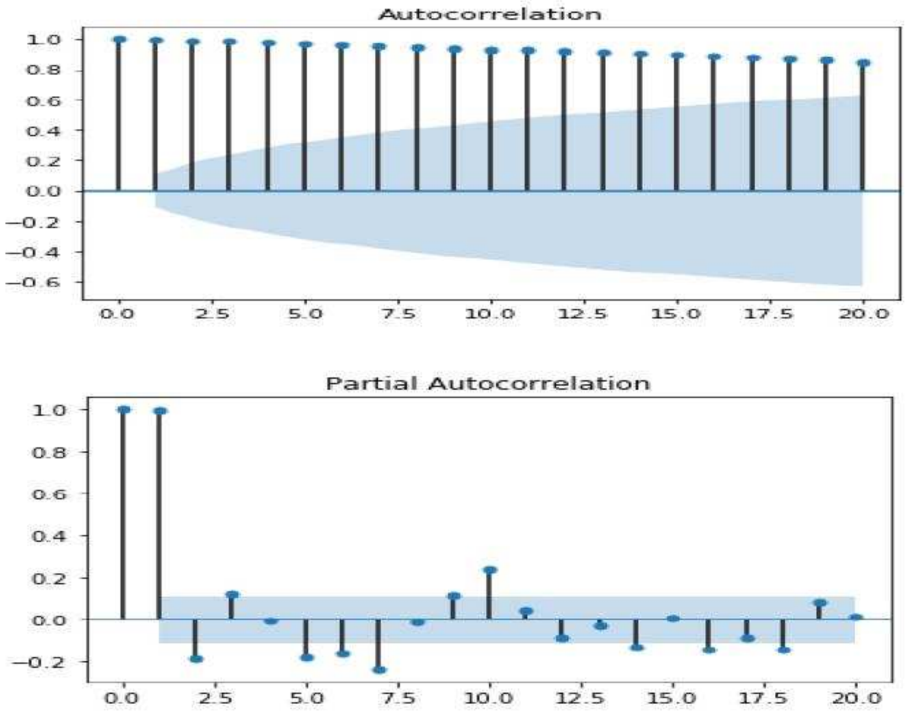

Fig. 3: ACF and PACF for Training Series for the Monthly Gold Price

• Achieving Stationarity of the Series

Table 3 illustrates that the training series of monthly gold price is not stationary. First, differencing is required for the training series of monthly gold price. To examine if another differencing is needed, ADF, DF, KPSS, and PP tests were applied. The results of these tests are listed in Table 4.

Table 4: P-values for KPSS, ADF, PP, and DF Tests for Training Series.

Test	ADF	DF	KPSS	PP
P-Value	0	0	0.14	0

As seen in Table 4, the p-value of KPSS is 0.14 which is greater than 0.05, while the p-values of ADF, PP, and DF are 0.00 which is less than 0.05, so there is no need for a second differencing. Thus, the first difference removes the trend and becomes stationary. The general upward trend has also disappeared. The plot of the data after taking the first difference of the training series for the monthly gold price is shown below in Fig. 4.

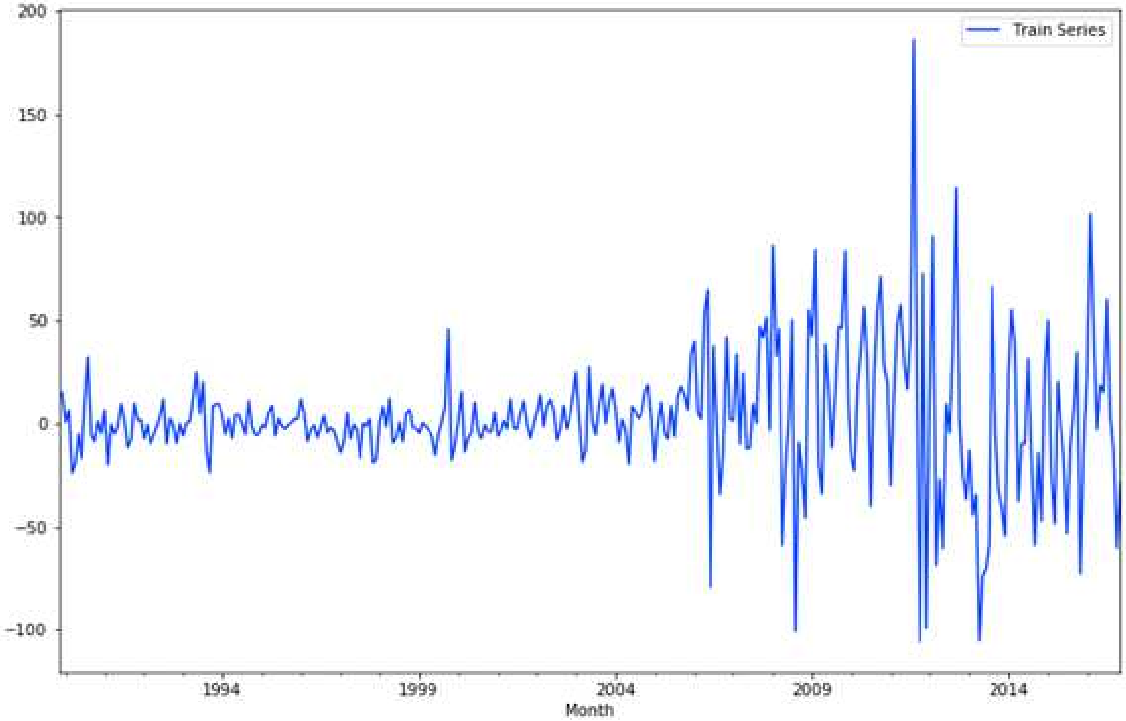

Fig. 4: Time Series Plot of the first difference of the training series for the monthly gold price

To decide the degree of the model, values of p, d, q, two different methods were used as illustrated below:

Degree of Differencing: From the above results and as shown in Table 4 and Fig. 4, we have to differentiate the data only once. Thus, ARIMA (p,d,q) model with d = 1 will be used to represent the process.

Finding p and q: for determining the values of p and q, Fig. 3 depicts that ACF has significant autocorrelations at lag 1 which leads to MA (1). The PACF plot shows definite significant values at lag 1 that leads to AR(1). These tools are important in the identification stage since they evaluate the statistical relationship between observations in a univariate time series. The tentative ARIMA model based on the ACF and PACF plots is ARIMA (1, 1, 1). After selecting the model parameters values to be ARIMA (1, 1, 1), different ARIMA (p, d, q) models are fitted to find out the best model for the monthly gold price. Moreover, to choose the best model for the data, the estimated model should be compared with other ARIMA models. The two common criteria, Bayesian Information Criteria (BIC) and Akaike's Information Criterion (AIC), are defined by:

$$BIC = ln(n)m - 2ln(\hat{L}) \tag{5}$$

and

$$AIC = 2n - 2ln(\hat{L}) \tag{6}$$

Where $(\hat{L})$ denotes the maximum value of the likelihood function for the model, m is the number of parameters estimated by the model, and n is the number of observations. The best model that has smaller AIC and BIC because of the number of parameters is the smallest. Different models associated with accuracy criteria are listed in Table 5.

Table 5: The Values of (AIC, BIC) for Different ARIMA Models.

Model	AIC	BIC
ARIMA (0, 1, 0)	3153.97	3161.47
ARIMA (1, 1, 0)	3145.82	3157.17
ARIMA (0, 1, 1)	3144.9	3156.24
ARIMA (1, 1, 1)	3145.55	3160.67
ARIMA (0, 1, 2)	3146.49	3161.62
ARIMA (1, 1, 2)	3146.13	3165.03

According to the results in Table 5, the best model is ARIMA (0, 1, 1) because the model has the smallest values of AIC and BIC criteria suggesting that the ARIMA (0, 1, 1) model is the best one. Therefore, the ARIMA (0, 1, 1) is the most suitable model that can be obtained for the training series of a monthly gold price.

2. Parameters Estimation:

After getting the appropriate value of p, d, q, the next stage is to find the values of the coefficients that best fit the selected ARIMA model. The most common methods use non-linear least-squares estimation or Maximum Likelihood Estimation (MLE).

We concluded that the ARIMA (0, 1, 1) model is the best model with the smallest values of AIC and BIC. Modeling results of an ARIMA (0, 1, 1) process have been estimated by MLE and the following model was obtained: ARIMA (0, 1, 1); coefficient MA1= 0.1923. Investigating the results of these estimates shows that all the coefficients are significant and the diagnostic check of the parameter estimates suggests that this model is suitable. Thus, the tentatively identified ARIMA model is as in Eq. 7.

Table 6: Parameter estimates of ARIMA (0, 1, 1) model.

Variable	Estimate	Standard Error	t-statistic	P-value
Constant	2.6017	0.217	2.379	0.019
MA(1)	0.1923	0.057	3.372	0.001

$$y_t = C + y_{t-1} + \varepsilon_t + \theta \varepsilon t - 1, \qquad \varepsilon_t \sim WN(0,1), \quad |\theta| < 1 \tag{7}$$

From Table 6 and Eq. 7, we can write the estimated model as:

$$y_t = 2.6017 + y_{t-1} + \varepsilon_t - 0.1923\varepsilon_{t-1} \tag{8}$$

3. Diagnostic of the Model:

It is the most important step to modeling ARIMA model in time series. In this step, we test whether the estimated parameters and residuals of the fitted ARIMA model are significant. The estimated errors of an estimated ARIMA model should resemble a white noise process if the model is correct.

Before we accept a fitted model and interpret its findings, it is essential to check whether the model is correctly specified , i.e. whether the model assumptions are supported by data. We must check the residuals of our model.

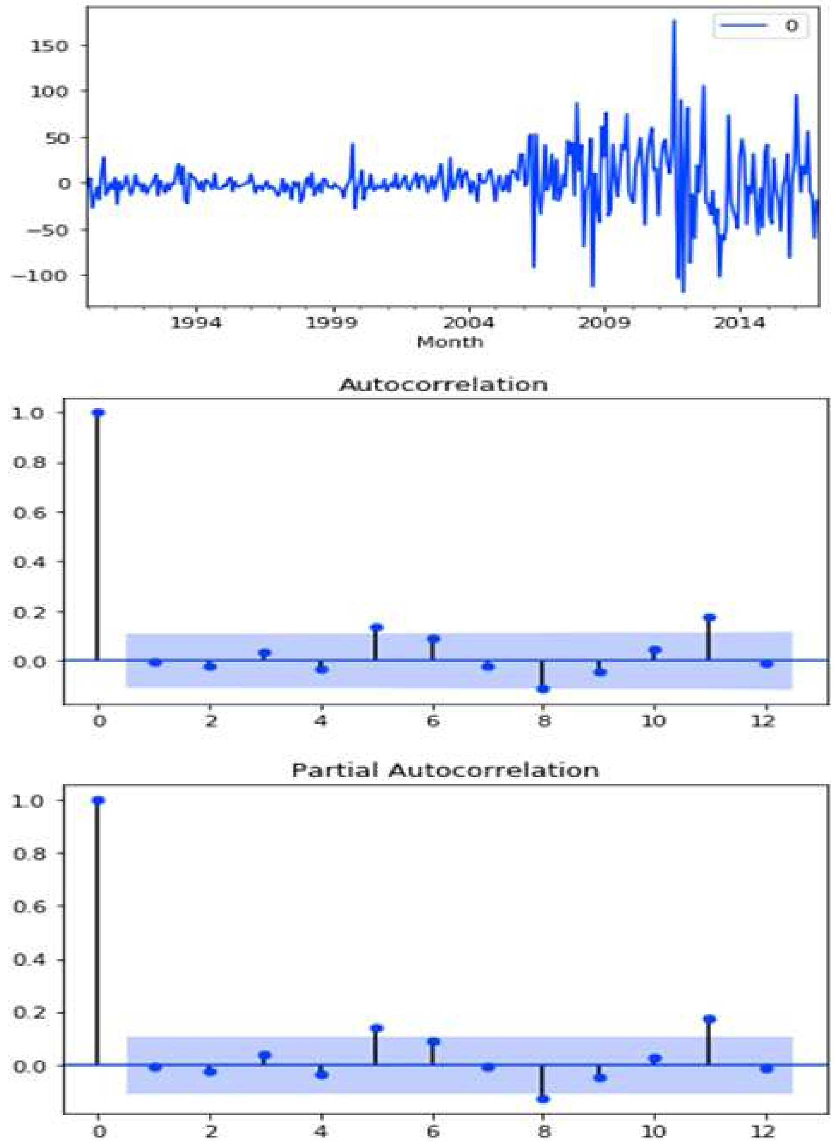

Fig. 5: Diagnostic Plots for the Residuals of the ARIMA (0, 1, 1) Model

Fig. 5 depicts three diagnostic tools for the fitted ARIMA (0, 1, 1) model that are the residuals, sample ACF and PACF of the residuals, for a whole range of values of K from 2 to 12. These suggest that the ARIMA (0, 1, 1) model fits the gold price time series. It can be concluded that the model of ARIMA (0, 1, 1) is the best model and can forecast the gold price quite good.

4. Forecasting:
Since ARIMA (0, 1, 1) model is fitted to the gold price data, we can use Eq. 8 directly to forecast gold price data for the testing data. In Table 7, the comparison of forecast values with actual data for the gold price for 37 months is displayed. Fig. 6 represents the plot of data and forecasts with 95% confidence interval where the series of the forecasted values follow the same behavior as the original series of the monthly gold price. Table 7 shows the results which illustrate the forecasts of 37 values of the time series and compare them with the last 37 actual values with 95% forecast limits.

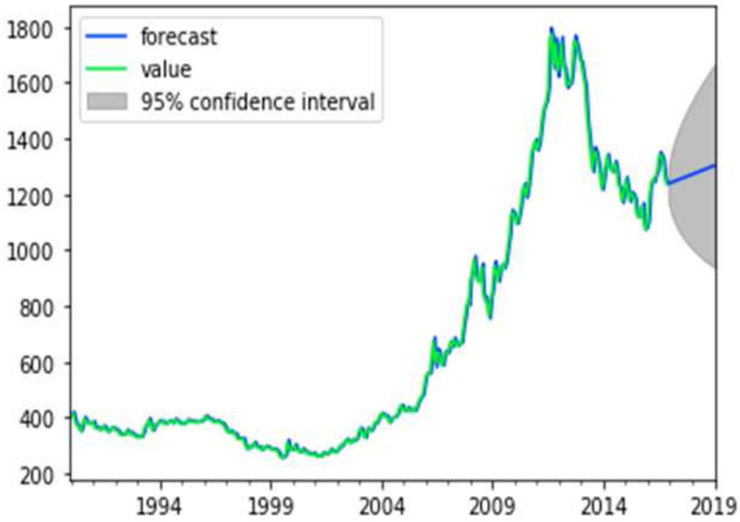

Fig. 6: Plot of the Data and the Forecasts with 95% Confidence Interval

Table 7: Observed and Forecasted Values of Gold Price using the ARIMA Model.

Month	Forecast	Actual	Month	Forecast	Actual
Dec-2016	1237.24	1157.36	Jul-2018	1286.68	1237.71
Jan-2017	1239.85	1192.1	Aug-2018	1289.28	1201.71
Feb-2017	1242.45	1234.2	Sep-2018	1291.88	1198.39
Mar-2017	1245.05	1231.42	Oct-2018	1294.48	1215.39
Apr-2017	1247.65	1266.88	Nov-2018	1297.08	1220.65
May-2017	1250.25	1246.04	Dec-2018	1299.68	1250.4
Jun-2017	1252.85	1260.26	Jan-2019	1302.28	1291.75
Jul-2017	1255.46	1236.84	Feb-2019	1304.89	1320.07
Aug-2017	1258.06	1283.04	Mar-2019	1307.49	1300.9
Sep-2017	1260.55	1314.07	Apr-2019	1310.09	1285.91
Oct-2017	1263.26	1279.51	May-2019	1312.69	1283.7
Nov-2017	1265.86	1281.9	Jun-2019	1315.29	1359.04
Dec-2017	1268.46	1264.45	Jul-2019	1317.9	1412.89
Jan-2018	1271.07	1331.3	Aug-2019	1320.5	1500.41
Feb-2018	1276.67	1330.73	Sep-2019	1323.1	1510.58
Mar-2018	1276.27	1324.66	Oct-2019	1325.7	1494.81
Apr-2018	1278.87	1334.76	Nov-2019	1328.3	1470.79
May-2018	1281.47	1303.45	Dec-2019	1330.9	1479.13
Jun-2018	1284.07	1281.57	-	-	-

3.2.2 Decision Tree

A decision trees (DT) is one of the simplest and most useful supervised machine learning structures that use a tree like model for decisions. Decision trees build regression or classification problems in the form of a tree structure. It breaks down the data into smaller subsets that contain similar values, while at the same time an associated decision tree is incrementally developed [13]. The final result is a tree with a root node, internal nodes and leaf nodes. Internal nodes contain one of the possible input variables (features) available at that point in the tree. The selection of input variable is chosen using information gain or impurity for classification problems and standard deviation reduction for regression problems. The leaves represent labels/predictions. In this study, decision tree method is applied for regression problems where variance reduction is employed for selection of variables in the internal nodes. First, variance of root node is calculated using Eq. 9, then variance of features is calculated using Eq. 10 to construct the tree.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{9}$$

In Eq. 9, n is the total number of samples and is the mean of the samples in the training set. Calculating variance of the root node, we calculate variance of input variables, as follows:

$$\sigma_X^2 = \sum_{c\varepsilon X} P(C)\sigma^2 \tag{10}$$

In Eq. 10, X is the input variable and P(c) is probability of the distinct values of this feature. Input variable that has the minimum variance or largest variance reduction is selected as the best node as shown in Eq. 11:

$$Vr_x = \sigma^2 - \sigma_x^2 \tag{11}$$

Finally, leaves represent the average values of instances with bootstrapping method. This process continues recursively until variance of leaves gets smaller than a threshold or all input variables are used. Once a tree has been constructed, new instance is tested by asking questions to the nodes in the tree. When getting a leaf, value of that leaf is taken as prediction.
To construct the DT model for gold price, the largest variance reduction was achieved using decision trees over 5-class classification.

*** Forecasting using Decision Tree algorithm**

In Fig. 7 and Table 8, the comparisons of the forecast values with actual data for gold price are shown.

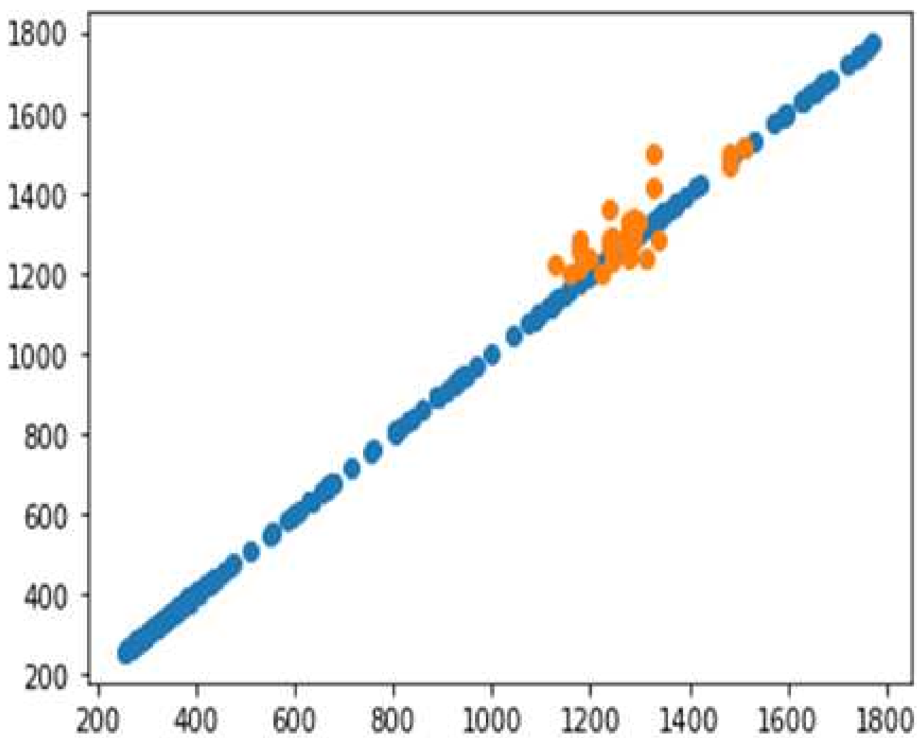

Fig. 7: Actual and Forecast results by Decision Tree

Table 8: Observed and Forecasted Values of Gold Price using Decision Tree.

Month	Forecast	Actual	Month	Forecast	Actual
Dec-2016	1209.15	1157.36	Jul-2018	1283.35	1237.71
Jan-2017	1128.87	1192.1	Aug-2018	1209.15	1201.71
Feb-2017	1209.15	1234.20	Sep-2018	1209.15	1205.92
Mar-2017	1209.15	1231.42	Oct-2018	1209.15	1215.39
Apr-2017	1209.15	1266.88	Nov-2018	1209.15	1220.65
May-2017	1283.35	1246.044	Dec-2018	1209.15	1250.4
Jun-2017	1283.35	1260.26	Jan-2019	1283.35	1291.75
Jul-2017	1283.35	1236.84	Feb-2019	1283.35	1320.07
Aug-2017	1283.35	1283.04	Mar-2019	1283.35	1300.9
Sep-2017	1283.35	1314.07	Apr-2019	1283.35	1285.91
Oct-2017	1283.35	1279.51	May-2019	1283.35	1283.7
Nov-2017	1236.55	1281.9	Jun-2019	1283.35	1359.04
Dec-2017	1283.35	1264.45	Jul-2019	1352.70	1412.89
Jan-2018	1283.35	1331.3	Aug-2019	1352.70	1500.41
Feb-2018	1283.35	1330.73	Sep-2019	1507.61	1510.58
Mar-2018	1283.35	1324.66	Oct-2019	1507.61	1494.81
Apr-2018	1283.35	1334.76	Nov-2019	1507.61	1470.79
May-2018	1283.35	1303.45	Dec-2019	1507.61	1479.13
Jun-2018	1283.35	1281.57	-	-	-

The results shown in Fig. 7 and Table 8 indicate that the tendencies of the predicted value curve are basically near to those of the actual value one, and the predicted values fit the actual ones very well.

3.2.3 Random Forest

Random forest (RF) is a type of meta learner that uses number of decision trees for both classification and regression problems [14]. The features and samples are drawn randomly for every tree in the forest that are trained independently. Each tree is generated using bootstrap sampling method. Bootstrapping relies on sampling with replacement. Given a dataset D with N samples, a training data set of size N is created by sampling from D with replacement. The remaining samples in D that are not in the training set are separated as the test set. This kind of sampling is called bootstrap sampling. The probability of an example not being chosen in the dataset that has N samples is:

$$Pr = 1 - \frac{1}{N} \tag{12}$$

The probability of being in the test set for a sample is:

$$Pr = (1 - \frac{1}{N})^{N} exp^{-1} = 0.3673 \tag{13}$$

Every tree has a different test set and this set consists of totally 63.27% of data. Samples in the test set are called out-of-bag data. On the other hand, every tree has different features which are selected randomly. While selecting nodes in the tree, only a subset of the features is selected and the best one is chosen as separator node from this subset. Then this process continues recursively until a certain error rate is reached. Each tree is grown independently to reach the specified error rate.
To construct the RF model for gold price, minimum root mean squared error was achieved using random forest with 200 trees in the first level.

Forecasting using Random forest algorithm
In Fig. 8 and Table 9, the comparisons of the forecast values with actual data for gold price are shown.

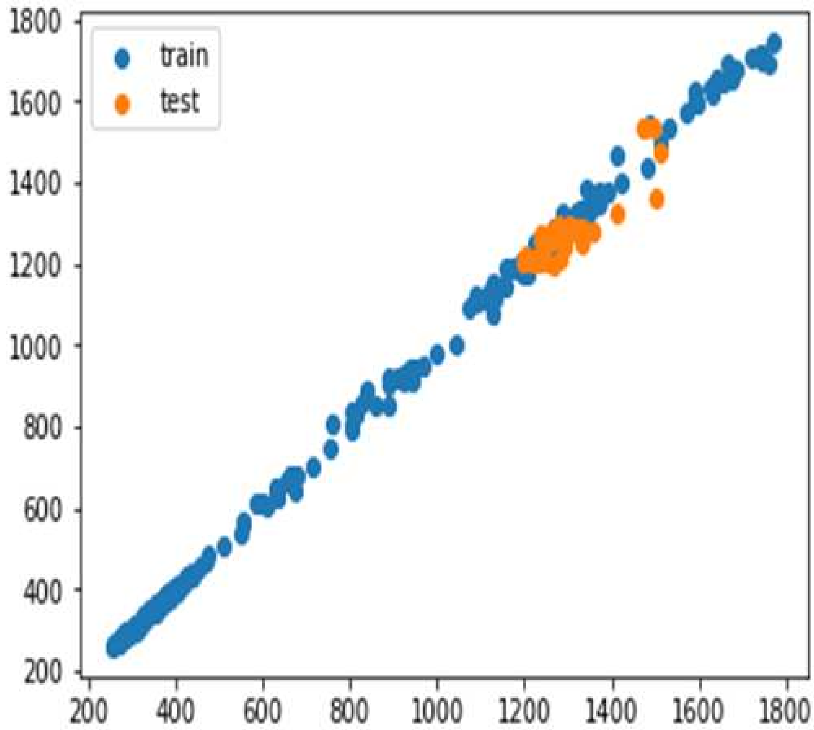

Fig. 8: Actual and Forecast results by Random forest

Table 9: Observed and Forecasted Values of Gold Price using Random forest.

Month	Forecast	Actual	Month	Forecast	Actual
Dec-2016	1273.11	1157.36	Jul-2018	1296.53	1237.71
Jan-2017	1163.79	1192.1	Aug-2018	1234.38	1201.71
Feb-2017	1205.86	1234.2	Sep-2018	1223.61	1198.39
Mar-2017	1224.82	1231.42	Oct-2018	1206.44	1215.39
Apr-2017	1222.31	1266.88	Nov-2018	1222.76	1220.65
May-2017	1262.24	1246.04	Dec-2018	1221.45	1250.4
Jun-2017	1253.79	1260.26	Jan-2019	1251.30	1291.75
Jul-2017	1263.11	1236.84	Feb-2019	1295.30	1320.07
Aug-20178	1230.09	1283.04	Mar-2019	1299.35	1300.90
Sep-2017	1296.24	1314.07	Apr-2019	1299.57	1285.91
Oct-2017	1299.29	1279.51	May-2019	1296.55	1283.7
Nov-2017	1295.60	1281.9	Jun-2019	1296.24	1359.04
Dec-2017	1296.95	1264.45	Jul-2019	1345.56	1412.89
Jan-2018	1264.78	1331.3	Aug-2019	1395.85	1500.41
Feb-2018	1301.67	1330.73	Sep-2019	1517.15	1510.58
Mar-2018	1301.46	1324.66	Oct-2019	1534.31	1494.81
Apr-2018	1299.79	1334.76	Nov-2019	1507.05	1470.79
May-2018	1305.22	1303.45	Dec-2019	1507.05	1479.13
Jun-2018	1299.63	1281.57	-	-	-

The results shown in Fig. 8 and Table 9 indicate that the tendencies of the predicted value curve are basically near to those of the actual value one, and the predicted values fit the actual ones very well.

3.2.4 Gradient Boosted Trees

The gradient boosted trees (GBT) method is an ensemble learning method that combines a large number of decision trees to produce the final prediction [15]. Boosting indicates that the model is built using a boosting process. Boosting is built on the principle that a collection of weak learners can be combined to produce a strong learner, where a weak learner is defined as a hypothesis function that can produce results only slightly better than chance and a strong learner is a hypothesis with an "arbitrarily high accuracy". The hallmark of all boosting methods is the additive training method which adds a new weak learner to the model in each step. In the case of gradient boosted tree, the weak learner is a new decision tree. This is shown in Eq.14, where F(x) is our full model after t-1 rounds and h(x) is the new tree we add to the model.

$$F_t(x) = F_{t-1}(x) + h(x) \tag{14}$$

GBT are similar to random forest models, but the difference is that trees are built successively. With each iteration, the next tree fits the residual errors from the previous tree to improve the fit.
The results shown in Fig.9 and Table 10 indicate that the tendencies of the predicted values are identical to those of the actual value one, and the predicted values fit the actual ones very well.
Forecasting using Gradient Boosted Trees algorithm
In Fig. 9 and Table 10, the comparisons of the forecast values with actual data for gold price are shown.

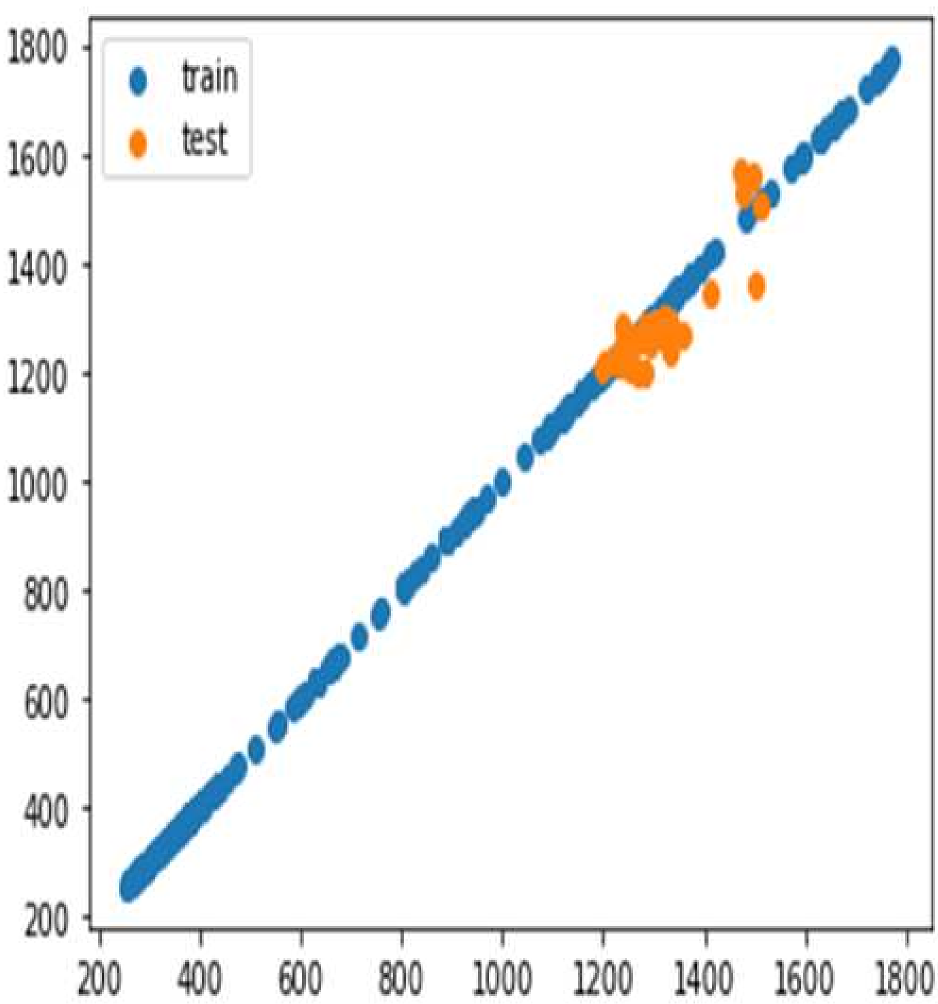

Fig. 9: Actual and Forecast results by Gradient Boosted Trees

Table 10: Observed and Forecasted Values of Gold Price using the GBT Model.

Month	Forecast	Actual	Month	Forecast	Actual
Dec-2016	1250.43	1157.36	Jul-2018	1297.22	1237.71
Jan-2017	1189.72	1192.1	Aug-2018	1234.44	1201.71
Feb-2017	1208.91	1234.2	Sep-2018	1222.44	1198.39
Mar-2017	1202.86	1231.42	Oct-2018	1205.15	1215.39
Apr-2017	1203.46	1266.88	Nov-2018	1207.90	1220.65
May-2017	1255.52	1246.04	Dec-2018	1215.28	1250.40
Jun-2017	1247.87	1260.26	Jan-2019	1241.44	1291.75
Jul-2017	1259.28	1236.84	Feb-2019	1287.00	1320.07
Aug-2017	1222.68	1283.04	Mar-2019	1285.59	1300.90
Sep-2017	1280.88	1314.07	Apr-2019	1279.68	1285.91
Oct-2017	1293.17	1279.51	May-2019	1278.60	1283.70
Nov-2017	1288.12	1281.9	Jun-2019	1285.09	1359.04
Dec-2017	1293.35	1264.45	Jul-2019	1355.06	1412.89
Jan-2018	1256.08	1331.3	Aug-2019	1385.44	1500.41
Feb-2018	1303.89	1330.73	Sep-2019	1490.04	1510.58
Mar-2018	1290.86	1324.66	Oct-2019	1536.68	1494.81
Apr-2018	1282.79	1334.76	Nov-2019	1516.42	1470.79
May-2018	1294.02	1303.45	Dec-2019	1510.06	1479.13
Jun-2018	1301.04.57	1281.57	-	-	-

The results shown in Fig. 9 and Table 10 indicate that the tendencies of the predicted value curve are basically near to those of the actual value one, and the predicted values fit the actual ones very well.

Evaluation methods for gold price forecasting

We used Root Mean Squared Error (RMSE) to evaluate model performances. It is square root of the summation of differences between actual and predicted values. RMSE is frequently used in time series analysis. RMSE can be calculated as shown in Eq. 15.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{15}$$

Where y_i is the actual value and $\hat{y}_i$ is the predicted value and n is the number of observations [16].

4 Results

Table 11 presents accuracy measure results for the ARIMA, DT, and RF Models. The smaller the (RMSE) values, the better the performance and the predicted values are closer to the actual values.

Table 11: A Comparison of the Performance Criteria for the Models ARIMA, DT and RF.

Model	ARIMA (0,1,1)	DT	RF	GBT
RMSE	75.46	43.70	38.52	43.26

The minimum accuracy measure (RMSE) of monthly gold price time series define the best model.The above-mentioned table indicates the following:
1. RF model performs better than the GBT model using (RMSE).
2. RF model performs better than the DT model using (RMSE).

3. RF model performs better than the ARIMA model using (RMSE).
4. GBT model performs better than the DT model using (RMSE).
5. GBT model performs better than the ARIMA model using (RMSE).
6. DT model performs better than the ARIMA model using (RMSE).

5 Conclusion

The present paper aimed to construct the best (ARIMA, DT, RF and GBT) models for the time series data of a monthly gold price from Nov-1989 to Dec 2019 and compare between models to see which one is better in forecasting the monthly gold price. The results of applying the ARIMA, DT, RF and GBT methods were compared through the (RMSE) results. From this study, it can be concluded from the presented discussion that results of RF were more accurate (with the lowest RMSE) and RF was more efficient forecasting technique for monthly gold price than DT, GBT and ARIMA models. In the future, we intend to improve our results using a hybrid method of ARIMA and Tree Based Methods to benefit from the qualities of both models.

Acknowledgement

The authors are grateful to the editor and the reviewers for the careful checking of the details and the constructive comments that improved this paper.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Mitchell, R., Michalski, J., and Carbonell, T. An artificial intelligence approach. Berlin: Springer, (2013).
[2] Taieb, S. B. Machine learning strategies for multi-step-ahead time series forecasting. Universit Libre de Bruxelles, Belgium, 75-86, (2014).
[3] Paliyawan, P. Stock market direction prediction using data mining Classification. future, 5, 6, (2006).
[4] Ahmed, N. K., Atiya, A. F., Gayar, N. E., and El-Shishiny, H. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6), 594-621, (2010).
[5] Kandananond, K. A comparison of various forecasting methods for autocorrelated time series. International Journal of Engineering Business Management, 4, 4, (2012).
[6] Navin, G. V. Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector Regression (SVR). International Journal of Science and Research (IJSR), 4(3), 2026-2030, (2015).
[7] Khan, M. M. A. Forecasting of gold prices (Box Jenkins approach). International Journal of Emerging Technology and Advanced Engineering, 3(3), 662-670, (2013).
[8] Zhang, X., Zhang, T., Young, A. A., and Li, X. Applications and comparisons of four time series models in epidemiological surveillance data. PLoS One, 9(2), (2014).
[9] Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 34(4), 802-808, (2018).
[10] Shumway.R.H. and Stoffer D.S. Time Series Analysis Using the R Statistical Package. Free Dog Publishing, (2018).
[11] Bleikh, H. Y., and Young, W. L. Time series analysis and adjustment: Measuring, modelling and forecasting for business and economics. CRC Press, (2016).
[12] Hyndman, R. J., and Khandakar, Y. Automatic time series forecasting: the forecast package for RJ Stat. Softw. 26 (3), 1âĂŞ22, (2008).
[13] Milunovich, G. Forecasting Australia's real house price index: A comparison of time series and machine learning methods. Journal of Forecasting, 39(7), 1098-1118, (2020).
[14] Breiman, L. Random forests. Machine learning, 45(1), 5-32, (2001).
[15] Friedman, J. H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232, (2001).Aladag, C. H., and Egrioglu, E. Advanced time series forecasting methods. Advances in time series forecasting. Bentham Science Publishers, Netherlands. 978-1, (2012).
[16] Aladag, C. H., and Egrioglu, E. Advanced time series forecasting methods. Advances in time series forecasting. Bentham Science Publishers, Netherlands. 978-1, (2012).