

# Predicting Future Gold Rates using Machine Learning Approach

Iftikhar ul Sami, Khurum Nazir Junejo  
Graduate School of Science and Engineering  
Karachi Institute of Economics & Technology  
Karachi, Pakistan

**Abstract**—Historically, gold was used for supporting trade transactions around the world besides other modes of payment. Various states maintained and enhanced their gold reserves and were recognized as wealthy and progressive states. In present times, precious metals like gold are held with central banks of all countries to guarantee re-payment of foreign debts, and also to control inflation. Moreover, it also reflects the financial strength of the country. Besides government agencies, various multinational companies and individuals have also invested in gold reserves. In traditional events of Asian countries, gold is also presented as gifts/souvenirs and in marriages, gold ornaments are presented as Dowry in India, Pakistan and other countries. In addition to the demand and supply of the commodity in the market, the performance of the world's leading economies also strongly influences gold rates. We predict future gold rates based on 22 market variables using machine learning techniques. Results show that we can predict the daily gold rates very accurately. Our prediction models will be beneficial for investors, and central banks to decide when to invest in this commodity.

**Keywords**—Gold rates; prediction; forecasting; linear regression; neural networks; ARMA Model

## I. INTRODUCTION

Historically, gold had been used as a form of currency in various parts of the world including USA [5]. In recent times also, gold has maintained its value and has been used as a means for assessing the financial strength of a country. Big investors have also been attracted to this precious metal and invested huge amounts in it. Recently, emerging world economies, such as China, Russia, and India have been big buyers of gold, whereas USA, South Africa, and Australia are among the big seller of this commodity [8]. Chinese and Indian traditional events also affect the price of the gold. In that time more money is poured for purchase of this commodity. Small investors also find this commodity for safe investment rather than alternate investment options, which bear in-built investment risks. Internal financial conditions of the aforementioned countries play a vital role for setting spot rates for gold. Governmental investments in gold are largely decided by their financial conditions, and interest rates, as they are indicators of the strength of their economy. When US interest rates become lower, more economic activity is witnessed in US, thus capital inflows in gold market are observed. Similarly, when interest rates lowered in China from 5.31 (2010) to 4.35 (2016), it bought gold aggressively [8].

Global investors, either countries or giant companies, tend to invest elsewhere if they foresee a significant decline in gold prices. In such a scenario, some investors turn to some other form of investment, such as US Bonds, or stock exchange. Fig. 1 shows that New York Stock Exchange (NYSE) and S&P 500 tend to do better when gold rates are at their low. S&P 500 is an American stock market index based on the market capitalizations of 500 large companies listed on the NYSE or NASDAQ. It implies that capital flow was noticed from stock exchanges to gold market. On the other hand, some stakeholders convert their gold reserves to USD, therefore EuroUSD index (exchange rate from Euro to USD) tends to rise with the decline in gold rates. Value of USD itself is dependent on various factors including the interest rates decided by US Government. Performance of the leading stock exchanges such as NASDAQ, and Dow Jones also reflects the strength of the US economy. Therefore, various phenomena are interconnected with gold rates and affect the price also.

The spot price is the current market price at which commodity is purchased or sold for immediate payment and delivery. It is differentiated from the futures price, which is the price at which the two parties agree to transact on future date. Gold spot rates are decided twice a day based on supply and demand in gold market. Fractional change in gold price may result in huge profit or loss for these investors as well as government banks. Forecasting rise and decline in the daily gold rates, can help investors to decide when to buy (or sell) the commodity. Various studies have been conducted by researchers to forecast gold rates, each of them insightful in their own right. We in this study forecast gold rates using a) the most comprehensive set of features than any of the previous studies, which for the first time includes the performance indicators of Russian, Chinese, and Indian economies (as they are the biggest purchaser of gold) and as well as the stock price of leading gold producing/trading companies, and b) apply various machine learning algorithms for forecasting and compare their results. We also identify which attributes influence the gold rates the most, some of which were not even used before.

The rest of the paper is organized as follows: Section II covers the related studies that have been conducted in this problem domain. In Section III, we describe our data collection process, and the various attributes that we used. Results are presented in Section IV. We finally conclude in Section V.

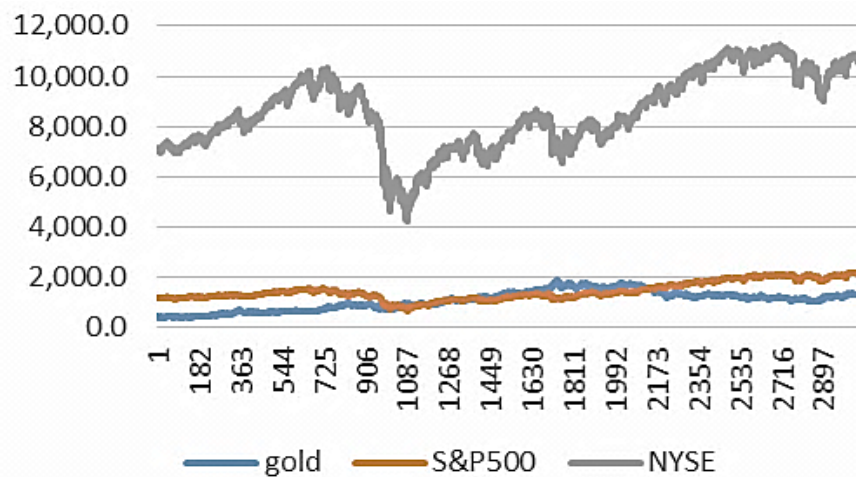


Fig. 1. Effect of index prices on gold rates.

## II. RELATED WORK

In [1], authors discuss the influence of US dollar on setting of crude oil and gold rates in the international market. They also analyze the influence of US dollar with respect to mutual funds, financial markets, interest rates of Federal Reserve, inflation, and economic recession between the period 1996 and 2009. They take in to consideration major world events which may have affected US dollar rates. In [2], authors use the simplest approach to predict future gold rates, as they do not take into consideration any attribute that may directly or indirectly influence the gold rates. Instead, they only use five attributes derived from gold rates itself. These five attributes are the opening, closing, highest, and lowest price of gold on a given day, and the volume of the commodity traded that day. They use decision trees and support vector regression algorithms for predicting gold rates, but do not report any results. On the other hand, [6] perform a very comprehensive empirical comparison of seventeen different approaches for time series modeling of the gold prices, and conclude that random walk approach is the best. The shortcoming of the study is that they only use few variables such as the price of other precious metals (palladium, silver, etc.) as input variables to the models. They do not take into consideration the economic conditions of major economies, or gold producing companies.

In [1], author use text mining and artificial neural networks (ANN) to forecast the gold prices and compare their results with the autoregressive-moving average (ARMA) model. ARMA model is the most frequently used statistical model for analyzing time series data. ARMA model consists of two parts, the first part AR, involves regressing the variable on its own past values. The second part MA, involves modeling the error term as a linear combination of error terms occurring simultaneously and at various times in the past. In [3] also, author use the ARMA model for predicting gold rates but use monthly rates of gold of past 124 months. They forecast actual gold prices and achieve an accuracy of 66.67%. In [10] also, author use ARMA model but compare their results with ANN and show that ANN performs better than ARMA. They used

data from 1990 to 2006 for training, whereas data from 2006 to 2008 was used.

For testing, Coefficient of determination ( $R^2$ ), mean absolute error (MAE), and root mean squared error were used for performance analysis. Cosine Amplitude Methods (CAM) test was also carried out for sensitivity analysis to determine relationship between related parameters.

In [4], author use extreme learning machines (ELM) algorithm, a variation of ANN. They compare the results of ELM with feed forward neural network without feedback, with back propagation, radial basis function, and ELMAN networks. They conclude that ELM performs the best with accuracy of 93.82%.

Variables considered by them include prices of gold, silver, and crude oil. They also consider Standard and Poor (S&P) 500 index and foreign exchange rate for preparing their model. In [5], authors take into consideration economic factors like inflation, currency prices, stock exchange performance, etc. to predict gold rates. They use multiple linear regression (MLR) model for forecasting gold prices based on eight independent variables. They conclude the most influencing parameters are Thomson Reuters Core Commodity (CRB) Index, EURUSD exchange rate, inflation rates, and money supply index (MI). Praise-Winsten procedure was used for removing correlated error terms. Using only these four attributes they achieved 96.92% accuracy. In [9], authors use logistic regression (LR) model and achieved 63.76% precision, 63.89% recall and 61.92% accuracy using eight years of data. They conclude that LR outperforms the SVM.

## III. PROPOSED METHODOLOGY

Table II lists the attributes considered by the studies discussed in Section II. The column labeled 'Proposed' lists the attributes used by us to build the models. Our attribute list is the most comprehensive and for the first time takes into consideration the performance indicators of Russia, China, and India as they are the biggest purchaser of gold. We do so, on the motivation that gold prices are constantly varying due to financial conditions of certain countries like USA, UK, China, and Russia among the other countries [7]. Their financial

strength lets them to invest more in gold, and when their economy becomes weak, and then they sell their gold reserves to strengthen their economy. Secondly, we also take into consideration the stock price of the major gold trade companies.

#### A. Dataset

Data for this study is collected from January 2005 to September 2016 from various sources. Data for attributes, such as Oil Price, NYSE, Standard and Poor's (S&P) 500 index, US Bond rates (10 years), EuroUSD exchange rates were gathered. Data of many government central banks and five large companies that have invested huge amounts in gold have also been collected. Price of precious metals during this period is also included in the analysis. Table I lists the online sources from which this data was extracted. Table II lists all these attributes.

The price of gold that we are trying to predict is taken in US Dollar. A lot of cleaning and preprocessing was performed on the dataset. The problem of missing values was handled in appropriate manner to complete the dataset.

Gold prices change on daily basis and are also affected by major world events. Current gold rates are much higher than a few years ago, as shown in Fig. 2. Keeping in view the huge difference in price, it was decided to split the dataset in a sequential fashion instead of random sampling. Therefore, the most recent 25% data is used as the test set, and the earliest 75% data is used for training. Thus, the first 2295 records make up the training set whereas the test set comprises of the last 770 rows. Due to major fluctuation in gold prices over the years, recent historical data would be more indicative of the future trend. Therefore, we further divide the training set into four versions. The first version contains all the records from 0% to 75%, the second version consists of records from 15% to 75%, the third version consists of records from 30% to 75%, and the last version from 45% to 75% of the whole data.

#### B. Correlation Analysis

Correlation analysis was performed to determine which of the twenty-two attributes collected by us are highly correlated to gold price. Fig. 3 shows the result of the correlation analysis. It gives some interesting insights. The attribute that has the highest correlation with gold rates is not the performance of US (or any other major) economy, or the price of other precious metals, but it is the stock price of Silver Wheaton Corporation (SLW), the world's largest precious metals streaming company. Other major gold producers, Eldorado Gold Corporation, and Compania de Minas Buenaventura, stand at the tenth and eleventh most correlated attributes. This is the first study to use the values of major gold producers to forecast gold price (see Table II).

Followed by SLW, as expected, the major correlated attributes are the price of precious metals (such as silver), and performance indicators of major world economies such as that of US and UK. A surprise is the seventh place of interest rate of Russia, first to be included in any study forecasting price of gold. Interest rate of China on the other hand does not have a major influence on gold price.

#### C. Machine Learning Models

We use two ML models, namely neural networks, and linear regression. Neural networks also known as artificial neural networks (ANN) are a family of models inspired by biological neural networks and are used to approximate functions that can depend on a large number of inputs. In addition to the input and output layers, they consist of one or more hidden layers of neurons that try to learn non-linear decision boundaries that separate different classes of data. It can also be used to predict continuous valued attributes such as gold prices in our case. Fig. 4 depicts a sample ANN.

Linear regression (LR) is an approach used in statistics to model relationship between dependent (class variable) and one or more independent variables (attributes). Linear regression can be used for predicting continuous valued attribute.

We use the implementation of LR and ANN that is provided by RapidMiner tool. Both models are optimized using the RMSE performance measure.

TABLE I. SOURCES OF DATA COLLECTION

Data	Source
Gold Spot Rates	www.gold.org
Silver Spot Rates	www.perthmint.com.au
Platinum Spot Rates	www.perthmint.com.au
Palladium Spot Rates	www.perthmint.com.au
Rhodium Spot Rates	www.markets.businessinsider.com
Oil Spot Prices	www.investing.com
S&P500 Index	www.finance.yahoo.com
US Bond Rates	www.treasury.gov
EuroUSD Index	www.finance.yahoo.com
NYSE Index	www.finance.yahoo.com
EGO Index	www.finance.google.com
SLW Index	www.finance.google.com
AU Index	www.finance.google.com
ABX	www.finance.google.com
BVN	www.finance.google.com
Interest Rate China	www.quandl.com
Interest Rate USA	www.treasury.gov.
Interest Rate UK	www.global-rates.com
Interest Rate Russia	www.global-rates.com

TABLE II. ATTRIBUTES USED FOR PREDICTING GOLD RATES

Variable	[1]	[4]	[5]	[6]	[9]	[10]	Proposed
Time Period	1996-2009	2000-2014	2003	1972-2013	2005-2013	1990-2008	2005-2016
Oil Spot Price	✓	✓				✓	✓
Oil Future Price	✓				✓		
Gold Spot Price	✓	✓	✓	✓	✓	✓	✓
Gold Future Price					✓		
S&P 500 Index	✓	✓	✓		✓		✓
Silver Price	✓	✓		✓		✓	✓
Silver Future Price					✓		
Copper Future Price					✓		
Platinum Price	✓			✓			✓
US Dollar Index	✓	✓	✓		✓	✓	✓
Base Interest Rate	✓						
Consumer Price Index (CPI)	✓						✓
Inflation	✓					✓	
Gold Production						✓	
US Bond (10 Years)	✓						✓
US Bond (5 Years)							
Dow Jones Index	✓				✓		✓
Commodity Research Bureau (CRB)			✓				
EUROUSD Rate			✓		✓		✓
Money Supply (M1)			✓				
NYSE Index			✓				✓
Treasury Bill			✓				
Palladium Price				✓			✓
Rhodium Price				✓			✓
Interest Rate (US)							✓
Interest Rate (UK)							✓
Interest Rate (China)							✓
Interest Rate (Russia)							✓
Eldorado Gold Corp. (EGO)							✓
Silver Wheaton Corp. (SLW)							✓
AngloGold Ashanti Limited (AU)							✓
Barrick Gold Corp. (ABX)							✓
Compania de Minas Buenaventura (BVN)							✓
NASDAQ					✓	✓	
FTSE					✓	✓	
Hang Seng Index					✓		
Nikkie Index					✓		
GBP-USD					✓		
JPY-USD					✓		
CNY-USD					✓		
Shanghai Composite					✓		
KOSPI Composite					✓		
Ibo Vespa					✓	✓	
DAX					✓	✓	
EuroNext 100					✓		

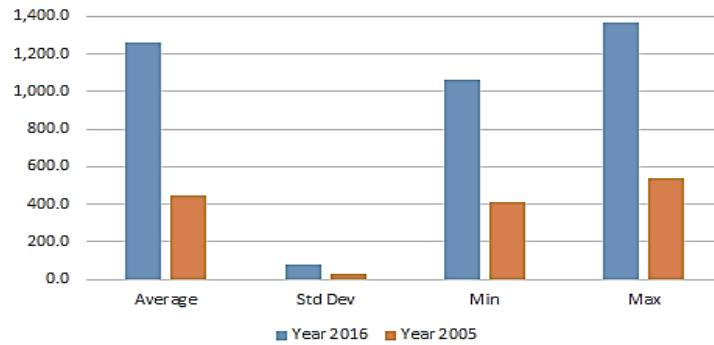


Fig. 2. Contrast in prices of gold.

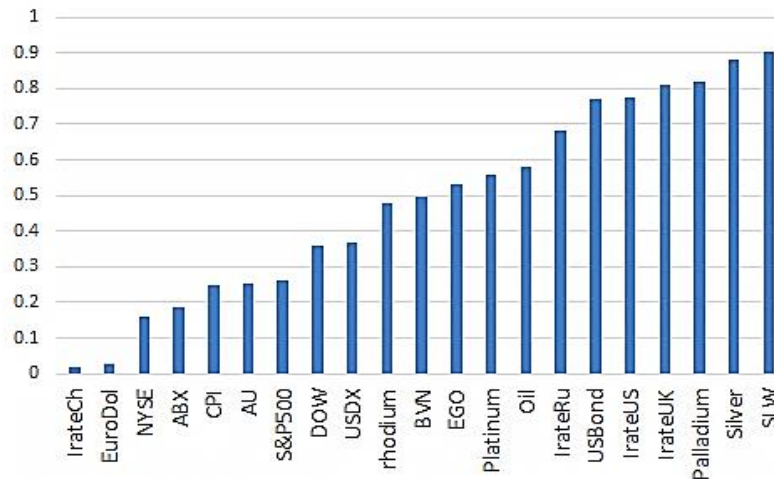


Fig. 3. Correlation analysis.

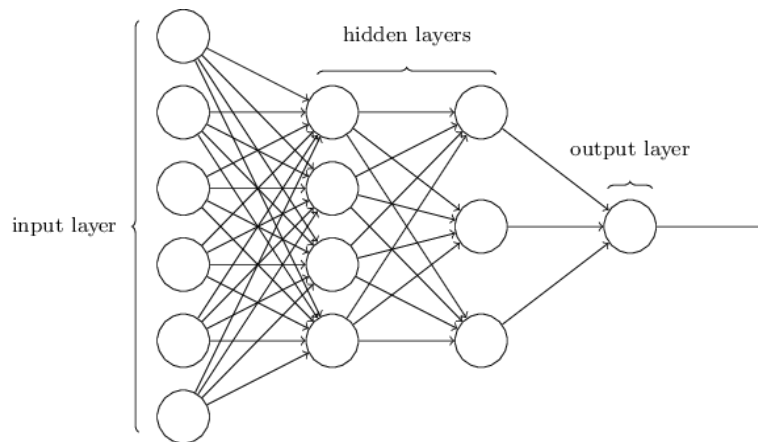


Fig. 4. Artificial neural network.

#### IV. RESULTS

ANN has various tuning parameters. We experimented with most of them and found two parameters, namely number of layers, and learning rate to be the ones having the greatest impact on its performance.

Therefore, Fig. 5(a)-(d) show the result of applying ANN on the test set while varying the values of these two parameters over the training set. The four figures correspond to the four

versions of the training set (Section III). Similarly, Fig. 6 depicts the performance of LR while varying its ridge parameter.

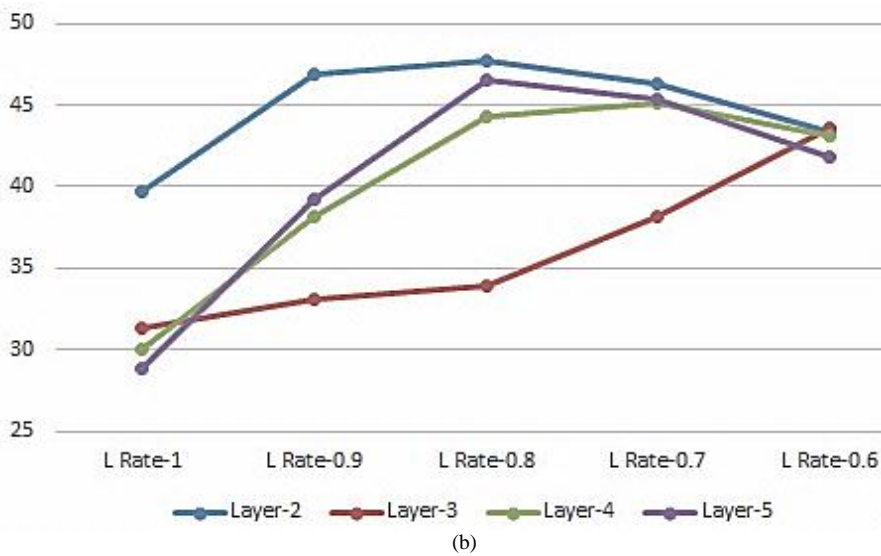
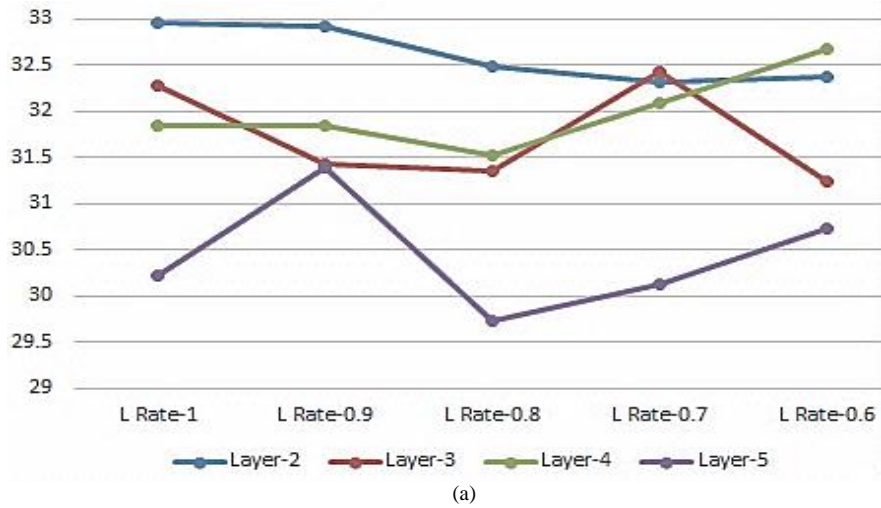
The results are very encouraging. Fig. 5(d) shows that while using as little as 920 days of data (i.e. 45-75%) for training, the root mean squared error is as low as 19. This is an extremely low error considering that the average rates of gold in the test data is above 1200\$ (Fig. 2). Performance of LR

even though is lower than that of ANN, but the difference is not significant. LR has the advantage of a faster training time than ANN.

Typically, the performance increases for ML algorithms when larger training data is used, but it is interesting to note that both the classifiers perform best when smaller training data (i.e. 45-75% training set version) is used. The intuition behind this phenomenon is that the bigger training data (i.e. 0-75% training data version) contains records of gold rates way back from 2005, while the gold rates have changed significantly in the last few years. For example, in 2005, average rate of gold was just above 400\$ whereas in 2016 it was above 1200\$. Thus, using too much history (way back from 2005) tends to deteriorate the performance of the classifiers.

It is much beneficial to use only recent history. The 45-75% training data consists of only a little more than two and half years of data prior to the data in the test set.

Fig. 5(a)-(d) also provide some insight about the working of ANN. ANN with two layers of neurons performs best when smaller training set is used, whereas in contrast, ANN with five layers of neurons performs best when a larger training set is used. This indicates that ANN with five layers for smaller training set, thus its relative performance overfitted the 45-75% training set, and thus its relative performance improved with the increase in the size of the training set. As for ANN with two layers, it was under fitting decreased with the increase in size of the training set. Whereas ANN with three layers best fits the data as it has the most consistent performance.



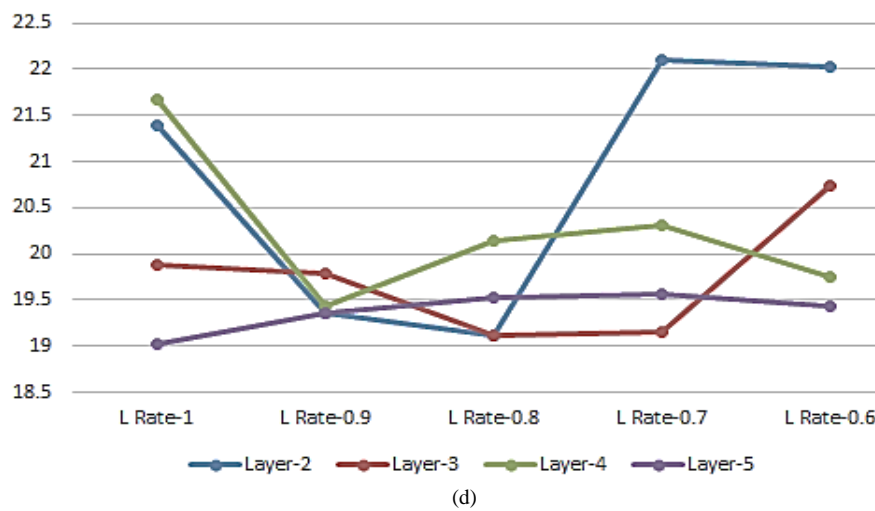
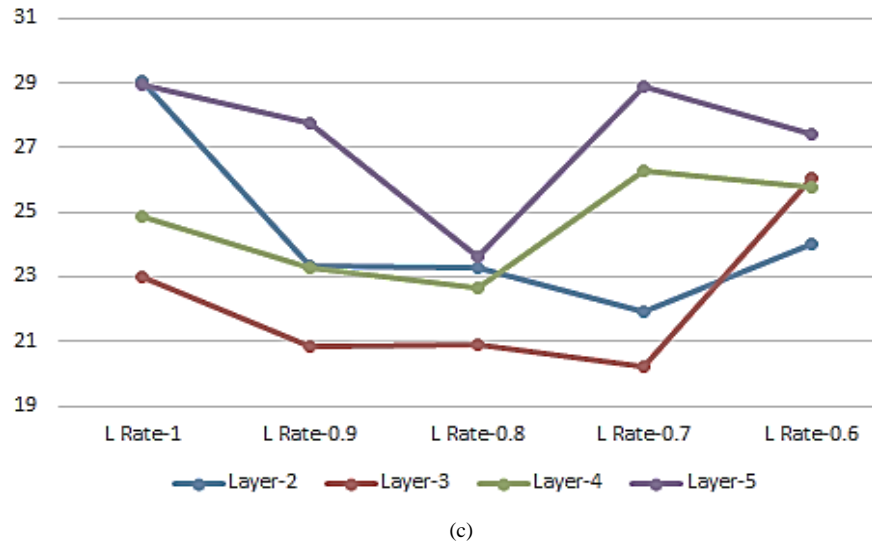


Fig. 5. (a) ANN error chart on 0-75% training dataset; (b) ANN error chart on 15-75% training dataset; (c) ANN error chart on 30 to 75% training dataset; (d) ANN error chart on 45 to 75% training dataset.

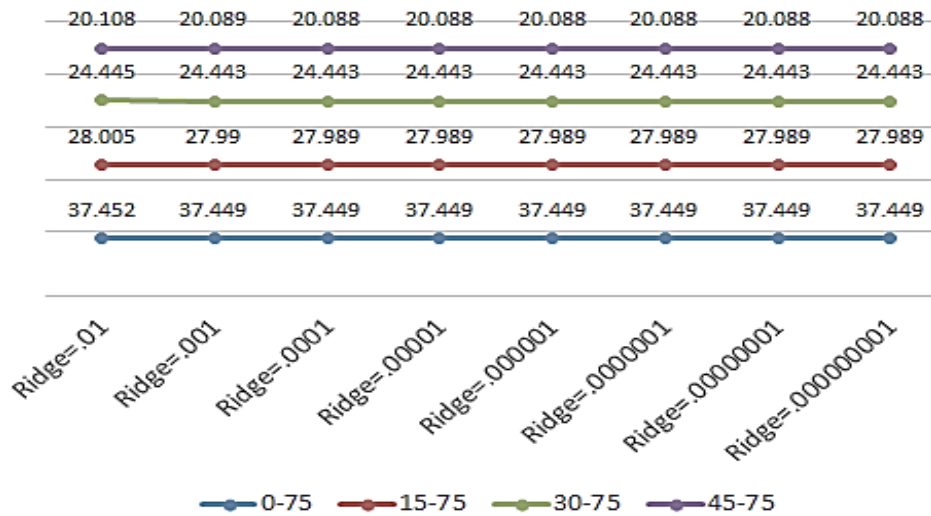


Fig. 6. Error chart of linear regression.

## V. CONCLUSION

Gold has been one of the most important commodities throughout history. Maintaining gold reserves by central banks is crucial to support the current economic structure of the world. Some major companies and investors also invest a huge amount of money in gold. Although not easy, predicting the rate of gold would help investors and central banks to better decide when to sell and buy it, thus maximizing their profits. In this study, we used machine learning algorithms to predict the gold rates very accurately. Our study is also the most comprehensive to date, thus taking into consideration various economic indicators of various countries and companies. It is the first time that the stock value of major gold trading/producing companies, and Russia's interest rates, have been successfully used as an indicator for forecasting of gold rates. To the contrary we show that stock value of a major company has more influence on the gold rates than US economy. In future, we intend to improve our results by using ensemble learning, and deep learning.

## REFERENCES

- [1] M. C. & C.-C. C. Hsin-Hung Chen, "The Integration of Artificial Neural Networks and Text Mining to Forecast Gold Futures Prices," *Communications in Statistics - Simulation and Computation*, pp. 1532-414, 2016.
- [2] D. G. Navin, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector," *International Journal of Science and Research*, pp. 2026-2030, 2015.
- [3] V. K. F. B. Rebecca Davis, "Modeling and Forecasting of Gold Prices on Financia Markets," *American International Journal of Contemporary Research*, pp. Vol 4, No 3, 2014.
- [4] M. S. S. S. Kumar Chandar, "Forecasting Gold Prices Based on Extreme Learning Machine," *International Journal of Computers Communications & Control*, pp. 372-380, 2016.
- [5] A. Y. A. S. Z. Ismail, "Forecasting Gold Prices Using Multiple Linear Regression Method," *American Journal of Applied Sciences*, pp. 1509-1514, 2009.
- [6] Hassani, H., Silva, E. S., Gupta, R., & Segnon, M. K. (2015). Forecasting the price of gold. *Applied Economics*, 47(39), 4141-4152.
- [7] I. S. G. a. A. Mody, "International Gold Price Movements, 1972-1982," *Economic and Political Weekly*, vol. Vol. 17, no. No. 46/47, pp. 1861-1870, 1982.
- [8] H. Long, "http://money.cnn.com," cnn.com, 10 february 2016. [Online]. Available: <http://money.cnn.com/2016/02/10/investing/china-gold/>. [Accessed 24 december 2016].
- [9] P. V. M. V. P. G. M. P. Sima P Patel, "Gold Market Analyzer using Selection based Algorithm," *International Journal of Advanced Engineering Research and Science (IAERS)*, vol. 3, no. 4, pp. 55-102, 2016.
- [10] A. Y.-C. Hossein Mombeini, "Modelling Gold Price via Artificial Neural Network," *Journal of Economics, Business and Mangement*, vol. 3, no. 7, pp. 699-703, 2015.