

**INTERNATIONAL BURCH UNIVERSITY
FACULTY OF NATURAL SCIENCES**

GOLD PRICE PREDICTION

SEMINAR PAPER

Subject: CEN 359: Introduction to Machine Learning

Professor: Elma Avdić

Student: Ida Bajrami
ID: 301017023

Sarajevo, June 2020.

ABSTRACT

Gold is one of the most valuable metals we know about. Back in the history, it was used as a mean of trade, besides the other methods of payment. In some countries, the gold is considered as gift or souvenir. Nowadays, the gold is much more valuable than before and it represents the financial strength for some countries. Like other precious metals, gold is measured by troy weight and by grams. The price itself is determined through trading in the gold and derivative markets and is constantly varying based on different variables.

In this project, I will try to most accurately predict and present the future gold price using the linear regression model. I believe my prediction can be somehow beneficial for banks, investors or anyone who wants to invest in this metal and follow its price range. The value of gold price may sometimes result in the huge profit or huge loss for the investors as well as for the Government banks. Gold price prediction can help in these kind of situations and can help the investors to more precisely decide when to buy or sell the gold.

Keywords: Gold price; Prediction; Linear regression model; Test dataset; Training dataset; Python.

Content table

1. Introduction.....	1
2. Related work	2
3. Proposed method.....	3
4. Experiments	4
4.1 Dataset.....	4
4.2 Implementation of necessary libraries.....	4
4.3 Implementation of necessary functions.....	5
4.4. Steps for implementation.....	5
5. Results and Discussion	7
6. Conclusion	9

1. Introduction

In this paper I will explain the details and usage of gold price prediction and what is the precise method used for that. The gold price has changed a lot through the history and nowadays is one of the most valuable metals which is used in different spheres of lives.

If we take a look at the history, the gold has been used as a form of a currency in a lot of countries. Now, the gold is usually used to determine the strength of a country. This metal has brought a lot of investor's attention. Besides them, some people started using gold as a measure of how rich it makes them and as a form of safe and long-lasting investment. At the time, the USA was one of the main countries that used gold as a currency and financial assets, but nowadays, world's economies such as Russia, India, China, Australia and others are among the biggest sellers of this luxury metal. Many government investments are decided by their financial condition and strength, which is mainly measured in gold. Whenever the interest rates are lowered, the most aggressive buying was seen in the USA and China. Other countries which are not active sellers of the gold, keep the gold bars in the international banks. Some of them have a historical approach mixed with modern technologies in order to secure the gold bars. This is just one of the examples how countries value the gold and how it can be helpful to the economy of a country.

On the other hand, some investors buy gold and convert them to USD or Euro. The values of these two currencies depend on many factors, including the interest rates decided by US or some other country Government.

The inputs to my algorithm are SPX, GLD, USO, SLV, EUR/USD (five columns of numerical datatypes and one column in date type) against the dates in the date column. I then use linear regression model to output a predicted price.

2. Related work

In some of the literature regarding this topic and machine learning in general, „The Integration of Artificial Neural Networks and Text Mining to Forecast Gold Futures Prices“ article, has stated and discussed the influence of the US dollar on setting the oil and gold prices in the international market. Some of the most significant world events were taken into consideration when it comes to the change of US dollar rates. Text mining and artificial neural networks are used to forecast the gold prices, and at the end compare their results with the *ARMA* (autoregressive-moving average) model. ARMA model is one of the most frequently used statistical models for analyzing the time series data. It has two parts: the first part AR, which involves regressing the variable on its own past values; and the second part MA, which involves modeling the error term as a linear combination of error terms occurring simultaneously and at various times in the past.

In „Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector“ article, some of the very simplest approaches were used to predict the gold rates and prices, meanwhile they did not take into consideration any of the attributes that can either directly or indirectly influence the gold rates. They used only five attributes derived from the gold rate itself: *opening*, *closing*, *highest* and *lowest price* of gold on a given day, and the *volume* of the commodity traded that day. Their major methods were the decision trees and support vector regression algorithms for prediction, but at the end none of the results were accomplished or reported. Some of the drawbacks for this kind of approach and methodology may be that they did not take economic conditions of the country, economy of gold producing companies and similar, into consideration.

Some other literatures have a similar approach and methodologies as stated above in these two articles. Some of the authors used the ARMA model for predicting gold rates and prices but used monthly rates of gold of past 124 months. They achieved the accuracy of 66,67%.

In „Forecasting Gold Prices Based on Extreme Learning Machine“ article, the author used an *extreme learning machines* (ELM) algorithm which is basically the variation of ANN. After they gained the results, they compared them with feed forward neural network without any feedback, with propagation, radial bias function and *ELMAN* networks. At the end, they came to the result of accuracy of 93,82%.

As the conclusion after all these work and experiments were done, the logistic regression model gave the accuracy of 63,76% and 61,92% accuracy using eight years of data. The final conclusion was that the linear regression models performs better than SVM.

3. Proposed method

Linear regression, in machine learning, is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. It is the type of supervised¹ learning, which is commonly used for predictive analysis. Regression is a technique that displays relationship between two or more variables. One explanatory variable case is called *simple linear regression*. When there are multiple variables, then the process is called *multiple linear regression*. In linear regression, all relationships are achieved using linear predictor functions, whose unknown model parameters are estimated from the data.

Linear regression model can be used to fit a predictive model to an observed dataset of values of response and explanatory variables. The most important uses of the linear regression algorithm are: determining the strength of predictors², forecasting an effect and trend forecasting.

The hypothesis function used in linear regression model is represented by:

$$y = \theta_0 + \theta_1 x,$$

where θ_0 and θ_1 are parameters. The main goal is to minimize the cost function:

$$J(\theta_0, \theta_1) = \frac{1}{2} * m \sum_{i=1}^n (h\theta(x^i) - y^i)^2$$

The process how the algorithm is applied and how the dataset is used:

- take the data from the dataset,
- apply the learning algorithm,
- if your algorithm is not working properly, you will easily notify it,
- do not get too close to the data,
- always ask the questions about the data, but do not get too close to dataset

The main goal of the learning is to find a model along with its parameters, so that the resulting predictor will perform well on data which is different from the training one.







¹ Supervised learning is learning from unlabeled data. Idea is that we are going to teach the computer how to do something while we are given the “right” answer. It is based on linear combination of fixed nonlinear basis functions.

² A predictor is a function that is given a particular input example and it produces the output. It can be represented in two ways; as a function or as a probabilistic model.

4. Experiments

4.1 Dataset

The dataset that will be used for this algorithm is from Kaggle. The dataset for this algorithm contains all necessary and depending factors needed for gold price prediction. The columns in this dataset are: Date, SPX, GLD, USO, SLV, EUR/USD. The price of gold is stated in US Dollar. Prices are affected on a daily basis with different world events, so current gold rates are much higher than a few years ago.

 Date	# SPX	# GLD	# USO	# SLV
 2Jan08 16May18	 677 2.87k	 70 185	 7.96 117	 8.85
1/2/2008	1447.160034	84.860001	78.470001	15.18
1/3/2008	1447.160034	85.57	78.370003	15.285
1/4/2008	1411.630005	85.129997	77.309998	15.167
1/7/2008	1416.180054	84.769997	75.5	15.053
1/8/2008	1390.189941	86.779999	76.059998	15.59
1/9/2008	1409.130005	86.550003	75.25	15.52
1/10/2008	1420.329956	88.25	74.019997	16.061001
1/11/2008	1401.02002	88.580002	73.089996	16.077
1/14/2008	1416.25	89.540001	74.25	16.280001
1/15/2008	1380.949951	87.989998	72.779999	15.834
1/16/2008	1373.199951	86.699997	71.849998	15.654

4.2 Implementation of necessary libraries

Since this algorithm and model will be used and implemented in Python, I will be using some of the libraries for better performance, implementation and visual representation:

- **Pandas** – is Python library to help load the dataset,

- **Numpy** – is Python library mainly used for scientific and mathematical computing. Since Python does not have a built-in support for arrays, using this library makes it possible to create multidimensional arrays,
- **Scipy** – one of the mathematical and scientific computing libraries similar to Numpy,
- **Sklearn** – is Python library, which is good for encapsulating multiple different transformers, which can be a class that can have fit or transform method,
- **Matplotlib** and **Seaborn** – libraries for visualizing the data and drawing all necessary graphs.

4.3 Implementation of necessary functions

After all libraries have been successfully installed and implemented, there are also some of the vital functions that have to be used in order to make the algorithm work:

- **DataFrame** – it represents two-dimensional size-mutable tabular data structure. Data is aligned in tabular fashion in rows and columns. When using it with Pandas library, it consists of *data*, *rows* and *columns*,
- **loc()** – this is method used within the DataFrame and it is used when selecting rows by label or index, which means that selection is based on index of the DataFrame,
- **iloc()** – also one of the methods used within the DataFrame, but it is slightly different from the previously mentioned loc() method. It is the Python indexer used for integer-location based indexing, which means selection by position. More precisely, when used within both DataFrame and Pandas, it is used to select rows and columns by number in order that they appear in DataFrame,
- **hist()** – a method that is the part of the histogram plotting which is a good tool for quick representation of probabilistic distribution,
- **pairplot()** – this method will, by default, create a grid of axes, such that each numeric variable from dataset will be stored in the y-axis across a single row, and in x-axis across a single column.

4.4. Steps for implementation

Steps that will be performed in Python are:

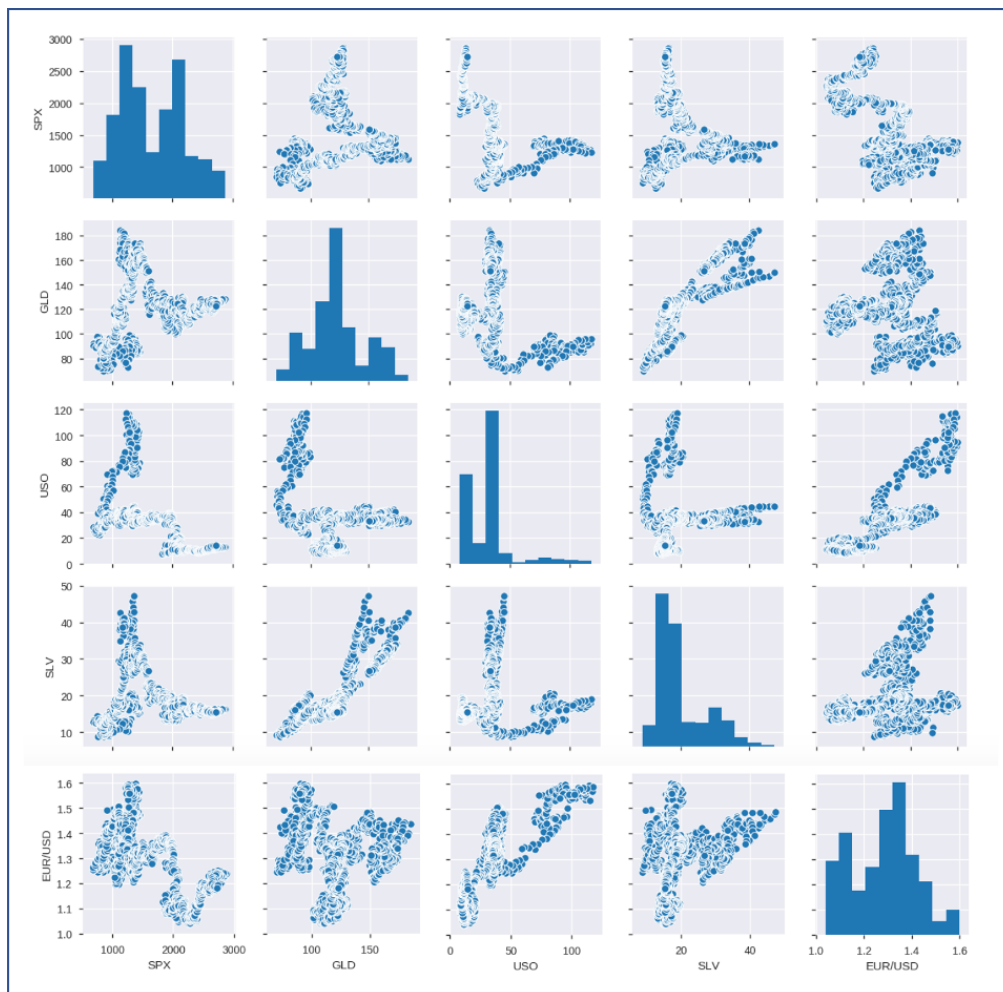
- import the libraries and load the dataset,

- define variables, plot the current data we have and sort it in descending order,
- split the data in training and testing sets,
- create linear regression model,
- train the model on a training set,
- check the error for regression,
- predict the gold price and plot the graph with the results.

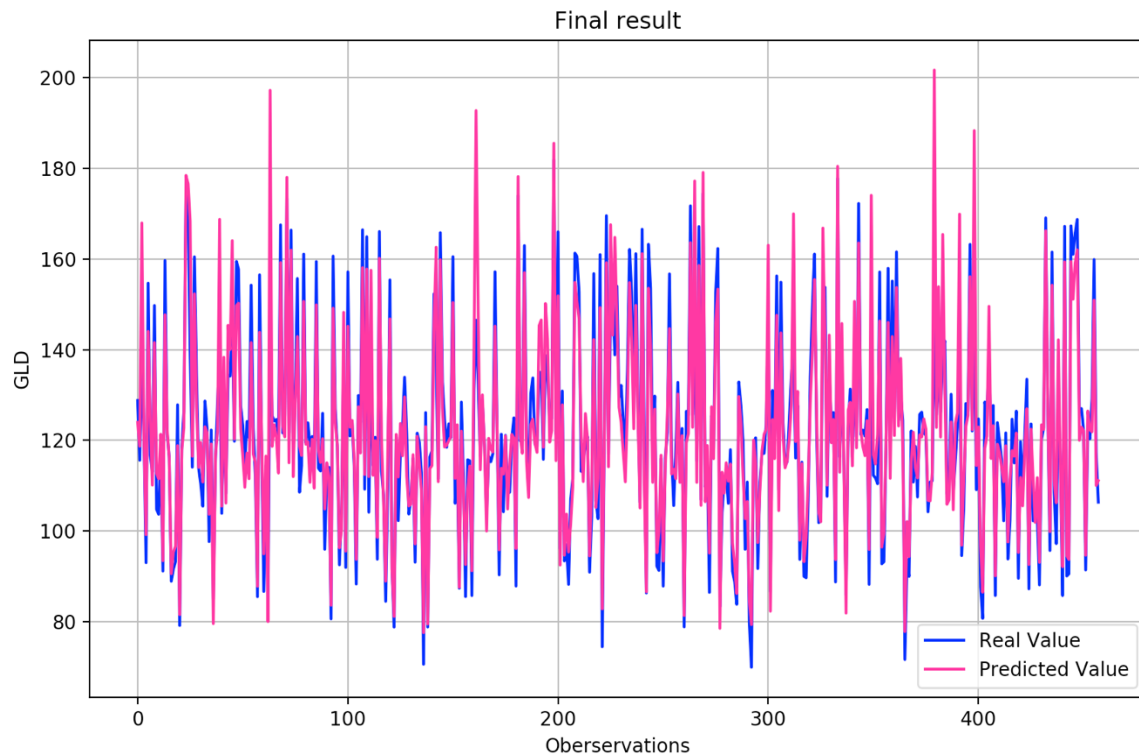
5. Results and Discussion

The main goal of this project is to predict the price of variable with column name GLD. The main objective is to predict the values of the variable GLD with holding the explanatory or predictor variables such as SPX, USO, SLV, EUR/USD, GLD. In this project, there are two approaches for better explanation of obtained results. The first approach is showing paired correlation and correlation of the predictors. The second approach is performing linear regression algorithm, splitting the data into two parts (test and train set), and finally using all necessary libraries for better statistical and visual representation.

The image below is about the correlation between the predictors, while the response variable is calculated and scatterplots are plotted accordingly. As shown below, the correlation between variables GLD and SLV is positive, which means that these two variables are highly correlated. The correlation between variables GLD and USO, which means that these two variables are slightly correlated. The correlation between variables GLD and EUR/USD, and variables GLD and SPX is almost zero, which makes these variables almost not correlated. *Heatmap* and *pairplot* are created mainly to show and understand the relations between variables better.



As the final result is shown on the picture below, the train and test sets are split in ratio of 80-20 and the random state is initialized to zero. After applying linear regression model, the *training accuracy* is 0.8877758904855643 and *testing accuracy* is 0.853012546687373. Also, the *mean absolute error* after running is 6.112232362040013 and the *mean squared error* is 75.61760337571606 (which implies that the square root of this error is 8.695838279068676).



6. Conclusion

The linear regression model itself tells us the probability, it estimates a continuous quantity and output the predicts. In other words, the aim of this approach is to model a continuous variable Y as a mathematical function of one or more x variables, so that we can use this model to predict Y when only x is known. As per results of calculations above, I can say that the final results are pretty good based on the training and testing accuracies (with 3,47% between these two accuracies). Based on the result of value of root mean squared error (8.695838279068676), we can say that the algorithm did a decent job.

If this project had more people and more precise dataset to work on, I believe that the huge update could be done regarding the gold price. I also think it is possible to create a system for this type of price prediction with high precision based on different parameters, such as main parameters for basic price calculation and maybe parameters provided by users or some investors. This system would be some kind of trading system, but with some changes and better performance. I believe that in the near future this project will expand and upgrade to a bigger level.