Machine Learning Engineer Nanodegree

Capstone Proposal

Manu Siddhartha

Jan 31st 2019

Gold Rates Prediction using Machine Learning Approach

Domain Background

Historically, gold had been used as a form of currency in various parts of the world including USA. 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 which results in reflecting the financial strength of the country. 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 gold.

Forecasting rise and fall in the daily gold rates, can help investors to decide when to buy (or sell) the commodity.

We in this project would forecast gold rates using the most comprehensive set of features and would apply various machine learning algorithms for forecasting and compare their results. We also identify the attributes that highly influence the gold rates.

We would use SPDR Gold Trust (GLD) Exchange Traded Fund data downloaded from https://finance.yahoo.com in the date range of **18/11/2004** to **01/01/2019**. We would try to predict **Adjusted Close price** of GLD ETF, the detail about the data would be described in the **Dataset and Input** section below.

Problem Statement

The challenge of this project is to accurately predict the future adjusted closing price of Gold ETF across a given period of time in the future. The problem is a regression problem, because the output value which is the adjusted closing price in this project is continuous value.

Various studies have been conducted by researchers to forecast gold rates using different machine learning algorithms with varying degrees of success but until recently the ability to build these models has been restricted to academics. Now with libraries like **Scikit-learn** anyone can build powerful predictive models.

For this project I will use different linear, ensemble and boosting machine learning models to predict the adjusted closing price of the SPDR Gold Trust (GLD) ETF using a dataset of past prices from **18/11/2004** to **01/01/2019**

Steps to be followed during Project

1. Exploring Gold ETF closing prices
2. Perform statistical analysis
3. Perform Feature Engineering
4. Normalizing the data
5. Splitting the dataset
6. Implement benchmark machine learning model and different solution models
7. Compare benchmark model with different machine learning models based on evaluation metrics described below.
8. Perform Feature Selection
9. Compare Feature selected models with full feature models

Datasets and Inputs

Data for this study is collected from November 18th 2011 to January 1st 2019 from various sources. The data has 1718 rows in total and 80 columns in total. Data for attributes, such as Oil Price, Standard and Poor's (S&P) 500 index, Dow Jones Index US Bond rates (10 years), Euro USD exchange rates, prices of precious metals Silver and Platinum and other metals such as Palladium and Rhodium, prices of US Dollar Index, Eldorado Gold Corporation and Gold Miners ETF were gathered. Table I lists the online sources from which this data was extracted.

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.We would use TimeSeriesSplit function in scikit-learn to split the data into training set and testing set, it could split the whole dataset into several packs and in each packs, the indices of testing set would be higher than training set. By doing this can prevent look ahead bias, which means the model would not use future data to train itself.

The historical data of Gold ETF fetched from Yahoo finance has 7 columns, Date, Open, High, Low, Close, Adjusted Close and Volume, the difference between **Adjusted Close** and **Close** is that closing price of a stock is the price of that stock at the close of the trading day. Whereas the adjusted closing price takes into account factors such as dividends, stock splits and new stock offerings to determine a value. We would use Adjusted Close as our outcome variables which is the value we want to predict. 'SP_open', 'SP_high', 'SP_low', 'SP_close', 'SP_Ajclose', 'SP_volume' of **S&P 500 Index**, 'DJ_open','DJ_high', 'DJ_low', 'DJ_close', 'DJ_Ajclose', 'DJ_volume' of **Dow Jones Index**, 'EG_open', 'EG_high', 'EG_low', 'EG_close', 'EG_Ajclose', 'EG_volume' of **Eldorado Gold Corporation (EGO)**, 'EU_Price','EU_open', 'EU_high', 'EU_low', 'EU_Trend' of **EUR USD Exchange rate**, 'OF_Price', 'OF_Open', 'OF_High', 'OF_Low', 'OF_Volume', 'OF_Trend' of **Brent Crude oil Futures,** 'OS_Price', 'OS_Open', 'OS_High', 'OS_Low', 'OS_Trend', of **Crude Oil WTI USD,** 'SF_Price', 'SF_Open', 'SF_High', 'SF_Low', 'SF_Volume', 'SF_Trend' of **Silver Futures**, 'USB_Price', 'USB_Open', 'USB_High','USB_Low', 'USB_Trend' of **US Bond Rate data**, 'PLT_Price', 'PLT_Open', 'PLT_High', 'PLT_Low','PLT_Trend' of **Platinum Price**, 'PLD_Price', 'PLD_Open', 'PLD_High', 'PLD_Low','PLD_Trend' of **Palladium price** 'RHO_PRICE' of **Rhodium Prices** 'USDI_Price', 'USDI_Open', 'USDI_High','USDI_Low', 'USDI_Volume', 'USDI_Trend' of **US dollar Index Price**, 'GDX_Open', 'GDX_High', 'GDX_Low', 'GDX_Close', 'GDX_Adj Close', 'GDX_Volume' of **Gold Miners ETF**, 'USO_Open','USO_High', 'USO_Low', 'USO_Close', 'USO_Adj Close', 'USO_Volume' of **Oil ETF USO.**

TABLE I. SOURCES OF DATA COLLECTION

Data	Source
SPDR Gold Trust (GLD) ETF	https://finance.yahoo.com
Oil Future Prices	https://in.investing.com
EUR USD Exchange rate	https://in.investing.com
Silver Futures Prices	https://in.investing.com/
Platinum Futures	https://in.investing.com/
Palladium Futures	https://in.investing.com/
Oil WTI USD	https://in.investing.com/
Rhodium Prices	https://www.quandl.com
US Dollar Index Futures prices	https://in.investing.com/
US Bond Rate Data	https://in.investing.com/
Oil ETF USO	https://finance.yahoo.com
Gold Miners ETF	https://finance.yahoo.com
SPY 500 Index	https://finance.yahoo.com
DowJones Index	https://finance.yahoo.com
Eldorado Gold Corporation (EGO)	https://finance.yahoo.com

Solution Statement

As the problem is the regression problem, so I would use supervised regression algorithms, such as **Support Vector Regressor, Random Forest Regressor, Laaso, Ridge, SGD** and **Gradient Boosting** and a model

that **ensemble** them together. I would apply them separately and choose the one which perform the best based on evaluation metrics.

The model would take input features as described above in dataset and inputs section in addition to this some technical indicators, like Simple Moving Average, Moving Average Convergence Divergence, Upper Band, Lower band, Relative Strength Index, DIFF, Open-Close and High-Low, the way to get them would be described in Project Design part. The model would return the forecast stock Adjust Close value in the chosen day range (5 days for example).

Benchmark Model

I would choose **Decision Tree Regressor** and use default arguments as the benchmark model. I would use the same data and features in my solution model detailed above. Further, I would compare simple benchmark model with hyper parameter tuned solution model based on the evaluation metrics described below.

Evaluation Metrics

I would use **Root Mean Square Error** and **R2 score** as my evaluation metrics. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. The formula for calculating RMSE is given below

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

Where f is the Predicted or forecasted value and o is observed or Actual value.

RMSE is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

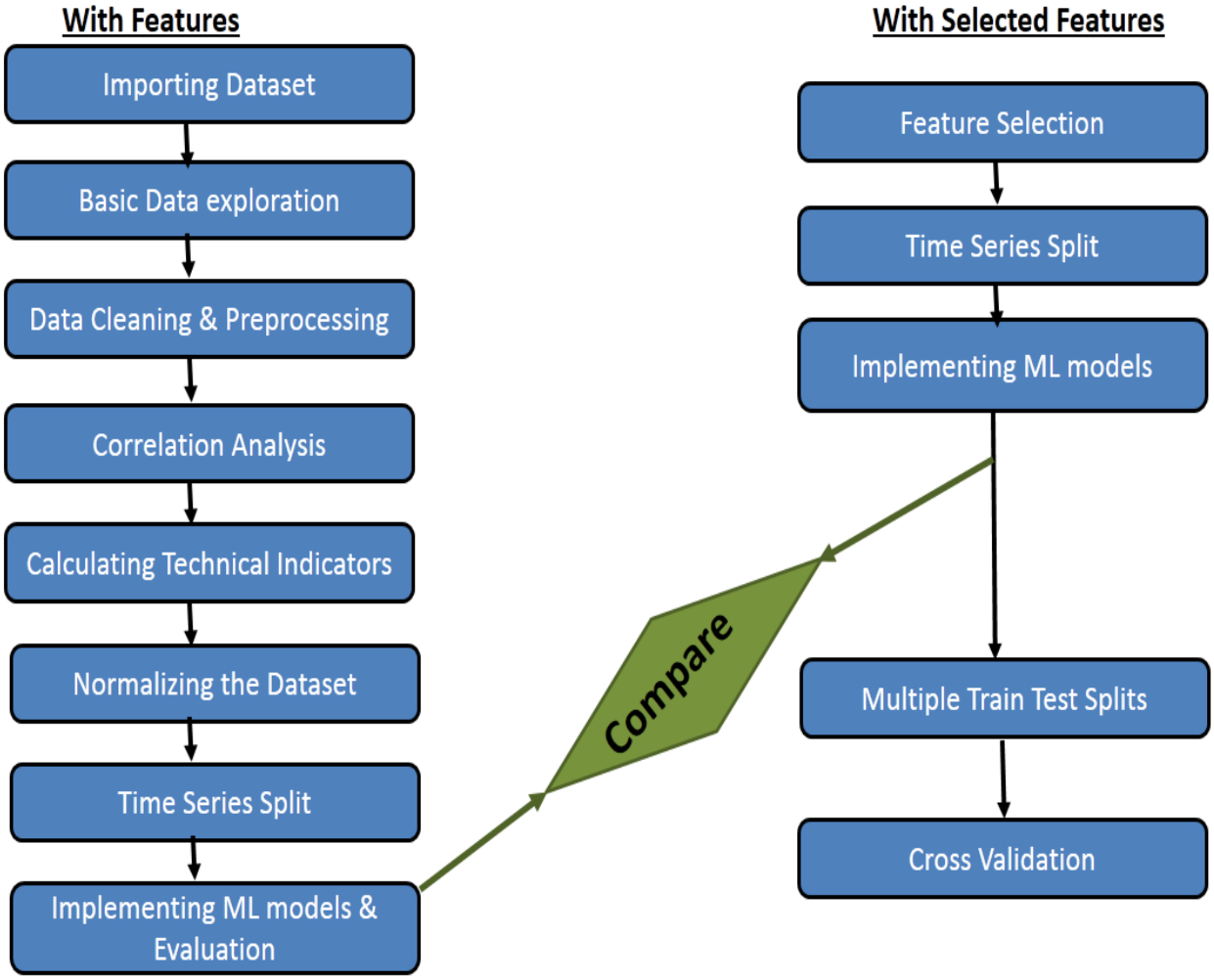
R-squared is always between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

So, when we apply both evaluation metrics to our benchmark and solution models , we would choose the one which has lower RMSE value and higher R2 score value.

Project Design

I would proceed the project into two parts. With all the Features and With Selected or best supporting features.



Platform & Libraries

I would use **Jupyter Notebook** to show the project work, code and description with complete model. Incorporate libraries such as sklearn, matplotlib, pandas and numpy.

Note :- The actual implementation may vary a bit from proposed design as I have not yet implemented the project. But the project workflow will remain same as mentioned above.

Refernces

- *Iftikhar ul Sami, Khurum Nazir Junejo, "Predicting Future Gold Rates using Machine Learning Approach", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 12, 2017*