2020 | By: Aishwarya S. Acharya

Machine Learning Capstone Proposal

Gold ETF price predictor

Gold ETF price predictor

Using machine learning to predict gold ETF prices

Domain Background

Gold metal is universally deemed valuable. There is strong global market demand for gold. Gold is considered to be an ideal hedge for financial market risk. Gold also has been used to back money because of its limited volatility in price. Gold is particularly in high demand in Asian countries like India where gold is an indicator of opulence.

Gold exchange traded fund represents physical gold in its dematerialized or on paper form. This allows investors to invest in gold online without having to physically gold any gold. This prevents any convenience costs or storage costs and therefore have been gaining popularity. Gold ETFs are more beneficial than in jewelry form as pure gold would have to be mixed with impurities to make hard. Due to the current uncertainty, it may seem like Gold could be a potential option for risk free investing.

Gold ETFs can be termed as open-ended mutual fund schemes, which will invest the investors' money in standard gold bullion of 99.5% purity. Because of its open-ended nature, gold ETFs are traded on stock exchange just like the shares of any company. For many years investment firms have used modeling methodologies to be able to understand the market movements and predict reactions. Algorithms have the power to make a trade happen in a fraction of the time it takes for a human to click a button.

Problem Statement

This project aims in utilizing Deep learning models to predict Gold ETF prices using historical information as a time series. I will analyze the market movements over 5 years and predict the price in the future. We will compare this against a regression model. I will use linear, long-short term memory (LSTM), and based on the accuracy I may try some other models as well.

The inputs will contain multiple metrics such as opening price (Open), Highest price (High), Volume, Adjusted closing price. We will try to predict the future adjusted closing price.

Dataset and inputs

In this project we are using prominent Gold ETFs in India. The list of the ETFs are as below:

- Axis Gold ETF (AXISGOLD)
- Birla Sun Life Gold ETF (BSLGOLDETF)
- Canara Robeco Gold ETF (CANGOLD)
- HDFC Gold Exchange Traded Fund (HDFCMFGETF)
- ICICI Prudential Gold Exchange Traded Fund (IPGETF)
- IDBI Gold ETF (IDBIGOLD)
- Kotak Gold Exchange Traded Fund (KOTAKGOLD)
- Quantum Gold Fund (QGOLDHALF)
- Reliance Gold Exchange Traded Fund (RELGOLD)
- Religare Gold Exchange Traded Fund (RELIGAREGO)
- SBI Gold Exchange Traded Scheme (SBIGETS)
- UTI GOLD Exchange Traded Fund (GOLDSHARE)
- Reliance ETF Gold BeES (GOLDBEES)

I will be using historical data provided from Yahoo Finance which has 7 columns Date, Open, High, Low, Close, Adjusted close, Volume. The adjusted close takes into consideration dividends, stock splits and new offerings while close is simply the end of the day closing price of the ETF. I will pick a period of 5 years from 2014 to 2019 for training while 2019 to 2020 will be used to validate the data.

Solution Statement

In this project I would like to predict the share price of Gold ETFs. Our benchmark metrics will be using linear regression model. We will use 2014-2019 data to train them model and then predict 2019. Further we will use the predictions to forecast beyond 2020 as well.
We will compare this against a regression model or a decision tree regression model, I intend to use a variety of models to get the highest accuracy. I will use linear, long-short term memory (LSTM) model and tune the hyper parameters to get a good accuracy. I will then apply this to other models like random forest or gradient boosting as well incase accuracy is not met.

Benchmark Model

We will be using linear regression as the benchmark. The benchmark will have the same input at the training model and will provide a benchmark for its performance.

.Evaluation Metrics

As this is a regression model, we will be using root mean squared error (RMSE) and R-squared to evaluate the model. The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

Project Design

I: Query data from Yahoo Finance and clears data to remove N/As and other columns are not required. I will be focusing on the date, ticker, and adjusted closing price from 1st January 2014 to 30th June 2020 that is, a period of over 6 years.

II: I will preprocess the data and split it into training and validation data. The first 5 years (1st January 2014 to 31st December 2018) will be to train, while 1st January 20119 to 30th June 2020 will be used to validate the predictions.

III: I will train this model using LSTM model. The model will have 6 features per stock and total 13 stocks. This brings the total feature count to about 78 features. I might have to do feature selection using correlation.

IV: I will run the input through the benchmark model.

V: I will then compare the two models and then made improvements to the model by parameter tuning

VI: One implemented, I will deploy my model and create a web app using Lambda function and API gateway.

The webapp should be able to predict the future Gold ETF prices.

References

Megan Potoski, "Predicting Gold Prices," CS229, Autumn 2013.

Siddharth Banyal, Pushkar Goel, Deepank Grover "Indian Stock-Market Prediction using Stacked LSTM AND Multi-Layered Perceptron" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020.