Proposal to use clustering and multiple tabular predictors to predict the 1-day future price of gold

MLND Capstone Project Proposal - N. Simon

Domain Background

Predicting changes in the value of gold has obvious financial benefit including knowing when to buy or sell and determining future volatility. Historically, gold has been used as a currency, either directly or as backing for a paper currency (Gold's history as a currency standard, 2010) and it has proven to be a useful asset to hold during times of economic downturn and uncertainty. This is despite the relatively small role it plays as a functional material in the modern world. The efficient market hypothesis (EMH) suggests that the market price reflects all information and that it should be impossible to consistently predict future prices of gold, and in particular, that it is impossible to consistently get returns greater than market level returns (Fama, E. F., 1970). Therefore, any system which is able to beat the market consistently will not only provide significant avenues for financial return, but also demonstrate that the EMH is not perfectly supported.

Problem Statement

The problem I will be attempting to solve is to predict the next day's gold price based on historical gold price data and S&P 500 prices. By comparing predicted price of gold with the actual market price I can determine how accurately the model predicts the value of gold and with what level of risk. Additionally, based on these predictions I can extrapolate an expected return of the model over a long period and show if it is significantly different to the market return or other naive trading strategies.

The Datasets

The datasets being used for this implementation is Nicholas Ward's Gold Historical Datasets as of 2021-12-19 (https://www.kaggle.com/nward7/gold-historical-datasets) and the S&P 500 Historical Data (https://ca.investing.com/indices/us-spx-500-historical-data). By combining the gold price with the S&P 500 price, with a set of previous returns and simple moving variance for both, I hope to add useful information for each row. I will create columns with the 16 day variance normalised return over the past day, previous 2 days, previous to that 4 days and previous to that 8 days (looking back a total of 16 days worth of prices), for each of the S&P 500 and gold price. After preprocessing the data, and removing empty rows or spurious data, I will divide it into 3 sets, a training set of size ~3000 (0.75 of the total dataset), a test set of size ~1000. In keeping with the usual practice for training on historical market price data, the training set will

not be randomly chosen but rather, will only have rows predating the rows in the test set. This is to prevent the training from "peeking" at future data and learning from it.

Solution Statement

I will implement a solution made up of a K-Means clusterising pre-processor with N centroids and N tabular predictors individually trained on the rows corresponding to a single cluster. Initially, I will train the K-Means categoriser on the entire training set using only the 8 normalised return values, and the 2 16 day variance values for a total of 10 dimensions. I will then use the trained K-Means categoriser to divide the training date into N datasets and use each to train an individual tabular predictor. It is hoped that by taking this approach the tabular predictors can each be trained to address a narrower set of circumstances, and ultimately give better results.

I will then test using the test dataset. I will use the K-Means categoriser to separate the test data into N sub-datasets. I will then measure the performance of the each of the tabular predictors using the appropriate test sub-dataset. For each row in the test dataset, I will compare the resulting predicted return to the actual historical market return for the appropriate day, and by analysing the result, determine if the system gets significantly better than market returns or not.

See Figure 1 for a flow chart illustrating the resulting system.

Benchmark Model

Since there are no public models that provide a significantly better than market level of return, I will be comparing the resulting system against 3 models, one which always buys gold, ultimately tracking the market return, another which always sells, and finally, another which randomly picks whether to buy or sell.

Evaluation Metrics

The usefulness of the resulting system will be in its ability to predict the price of gold with sufficient accuracy as to give a statistically significant potential return greater than the market itself. The EMH suggests that this will be impossible, and that it is only possible to, at best, match market level returns. To compare the market return to the predictions, I will determine if the prediction matches the market direction on a daily basis more than 50% of the time. In addition, I will compare how close the predicted return is to the market return when compared with the average return. Anything consistently better than chance is a success.

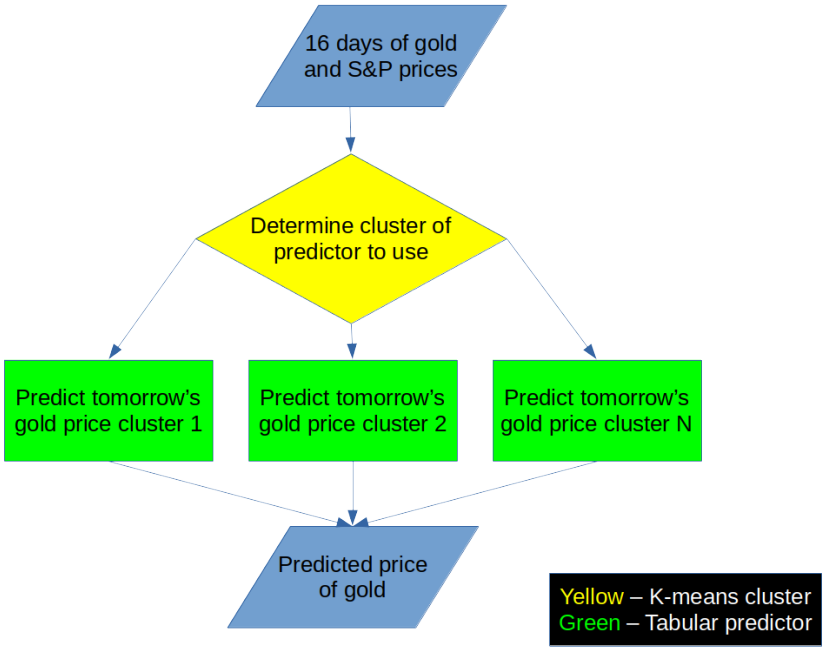

Figure 1: Model of resulting system

Project Design

I will implement the solution to the above system using AWS Sagemaker to simplify rapid prototyping and developing of the system. I will implement the K-Means categoriser by using the usual AWS training job workflow and then deploy to an AWS endpoint. I will use that endpoint to divide the training and test data into appropriate subsets, and then shutdown the K-Means categoriser endpoint. I will then train and test each of the N tabular predictors in turn. Finally, using the resulting categoriser and N tabular predictor models, I will implement production ready system using lambda functions to coordinate submission of a single input (16 days of gold and S&P 500 prices) and produce a predicted gold price and ultimately a buy or sell signal.

References:

- Gold's history as a currency standard. (2010). Retrieved from https://www.reuters.com/article/idINIndia-52748720101108
- Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486
- Raghuram, K. S. Statistical, Machine Learning predictive analytics and the impact of stock market indicators in predicting gold prices.
- Riazuddin, M. Predicting Gold Prices Using Machine Learning (2020). Retrieved from https://towardsdatascience.com/machine-learning-to-predict-gold-price-returns-4bdb0506b132
- Shah & Pachanekar (2021). Gold Price Prediction Using Machine Learning In Python. Retrieved from https://blog.quantinsti.com/gold-price-prediction-using-machine-learning-python/
- Datta P. (2021). Building A Gold Price Prediction Model Using Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2021/07/building-a-gold-price-prediction-model-using-machine-learning/