Nile University

Gold Price Prediction Project

CSCI417/ECEN425: Machine Intelligence

Prof. Ghada Khoriba

Problem Statement

The goal of this project is to build a model that can predict the gold price, using the provided dataset of gold price data from Kaggle[1]. The dataset includes the following features: SPX*, GLD*, USO*, SLV* and EUR/USD*. The target variable is the GLD price.

*GLD: stands for the SPDR Gold Trust, which is an exchange-traded fund (ETF) that tracks the price of gold.

*USO: stands for the United States Oil Fund, which is an ETF that tracks the price of crude oil.

*SLV: stands for the iShares Silver Trust, which is an ETF that tracks the price of silver.

*EUR/USD: stands for the exchange rate between the European Union's euro and the US Dollar.

Related Work

There have been several studies on gold price prediction using various machine learning techniques. In a study by Kumar et al. (2019)[2], the authors used a combination of feature selection and machine learning techniques to predict the gold price. They found that support vector machines (SVMs) and random forests achieved the best performance. Another study by Kasozi et al. (2019)[3] used artificial neural networks (ANNs) to predict the gold price and found that ANNs outperformed traditional machine learning algorithms such as linear regression and decision trees.

Data Preprocessing

The first step in this project was to load and preprocess the data. The following preprocessing steps were taken:

1. The data was loaded from a CSV file into a pandas DataFrame.
2. The Date column was converted to a datatime object and set as the index of the DataFrame.

Data Exploration and Visualization

To get a better understanding of the data, we plotted the time series of the different features. This allowed us to visualize the trends and identify any patterns or anomalies in the data. Some observations from the plots include:

- The gold price has generally trended upwards over the time period covered by the data.
- The SPX, USO, and SLV prices also show upward trends, although there are some periods of volatility.
- The Gold and Silver time series show a similar trend.

Model Architecture

For this project, we will implement and compare the performance of five different machine learning models: linear regression, random forest, support vector machine, XGBoost and Lasso regression. These models were chosen because they are commonly used for time series prediction and have achieved good performance in previous studies on gold price prediction.

Evaluation Results and Evaluation Strategies

To evaluate the performance of the different models, we will use the mean absolute error (MAE) between the predicted and actual gold prices. The MAE is a common metric for evaluating time series prediction models and is defined as the average absolute difference between the predicted and actual values.

To prevent overfitting, we will split the dataset into training and test sets and only use the training set for model fitting. The test set will be used to evaluate the performance of the models. Additionally, we will use cross-validation to fine-tune the hyperparameters of the random forest and XGBoost models. (Will be shown later why these two specific models are chosen to fine-tune them)

Evaluation Outputs

After implementing and evaluating the different models, the following results were obtained:

- Linear Regression: MAE = 5.69
- Random Forest Regressor: MAE = 1.26
- Support Vector Machine: MAE = 16.10
- XGBoost: MAE = 1.46
- Lasso Regression: MAE = 5.62

To further improve the performance of the random forest and XGBoost models, we fine-tuned their performance using grid search and cross-validation. The hyperparameters that were tuned for the random forest model included the number of estimators, the maximum depth of the trees, and the minimum number of samples required to split a node.

The fine-tuned versions of the random forest and XGBoost models achieved the best performance:

- Random Forest Regressor (Tuned): MAE = 1.22
- XGBoost (Tuned): MAE = 1.43

Based on the evaluation results, it can be concluded that the fine-tuned Random Forest Regressor is the best model of predicting the gold price in this dataset. The MAE of 1.22 indicates that the model is able to predict the gold price with an average error of around 1.22 units.

Conclusion

This project demonstrated that machine learning techniques can be effectively used to predict the gold price. The fine-tuned Random Forest Regressor was found to be the best model for this task, achieving an MAE of 1.22. Further studies with largfer and more diverse datasets may be needed to confirm the effectiveness of this model for gold price prediction.

References

1) https://www.kaggle.com/datasets/altruistdelhite04/gold-price-data
2) https://www.sciencedirect.com/science/article/pii/S1364815218308879
3) https://www.sciencedirect.com/science/article/pii/S1364815219300813