CS230

Gold price prediction using CNN-LSTM, with sentiment analysis (Finance, NLP)

Mia Sarojini Kynadi
SCPD
Stanford University
kynadi@stanford.edu

Abstract

This project uses a CNN-LSTM model to predict the price of gold, given a history of gold prices over the previous few days. It builds on previous related work by adding other time series information which are known to affect the price of gold. It also explores the option of predicting the percentage change in the price of gold from the previous day.

1 Introduction

Gold is one of the most traded commodities in the world. The price of gold fluctuates on a daily basis, and being able to predict the price of gold, as well as indicate whether the price is likely to go up or down from one day to the next can hugely benefit gold traders, and has the potential to make or break fortunes. Several machine learning approaches have been applied to this problem with varying degrees of success.This project explores a CNN-LSTM approach to predicting gold prices and proposes to build on previous related work. In addition to using past prices of gold and other commodities that are known to affect the price of gold as inputs, this project also proposes using news sentiment analysis specifically on news related to gold, to see if this can boost prediction performance.

The input to the CNN-LSTM model is a gold price taken from Yahoo Finance avialable in the public domain. In a similar manner, other indices which are known to affect the price of gold [3] are also taken from Yahoo Finance.

For the news sentiment analysis, another dataset available in the public domain will be used, which contains over $10,000$ manually annotated news headlines relating specifically to gold from 2000 to 2021, and classifies the headlines across multiple dimensions into various classes. This project explores one method of adding news sentiment as an input, although implementation and analysis of adding news sentiment is earmarked for future work.

The output from the CNN-LSTM model will be the predicted price of gold. Alternatively the model is set up to predict the percentage change in gold price from the previous day.

CS230: Deep Learning, Fall 2022, Stanford University, CA.

2 Related work

The paper [1] describes the use of a CNN-LSTM architecture for predicting the price of gold and this is taken as the baseline for this project. The model used by the authors consists of two convolutional layers followed by a max pooling layer. The output from the max pooling layer is flattened and fed as input to an LSTM layer. The LSTM layer's output is fed to a fully connected layer and finally an output layer which gives the regression output value. The reasoning for this architecture is that the CNN layers extract useful features from the time series input and this feature rich vector is fed to the LSTM layer. The authors experiment with forecasting horizons of 4, 6, and 9 days and found that the prediction accuracy was best with a 9 day window.

The authors use only past gold prices as input to the algorithm for predicting the future price of gold. In this project, other factors that are known to affect the price of gold such as Standard and Poor's (S&P) 500 index, US Dollar Index etc will also be added as inputs to the model. The authors also use the model as a classifier to predict if the price of gold will go up or down compared to the previous day, given a history of the past prices of gold. Instead of this approach, this paper uses past history of percentage change in the price of gold to predict the next percentage change in the price of gold as a regression problem.

The paper [2] describes the effect of news sentiment on gold price prediction. The dataset used is also released by the authors. The authors experiment with multiple architectures including simple and bidirectional LSTM, simple and bidirecional RNN, and BERT to get embeddings for the news headlines, and use these for classifying the headlines across multiple dimensions. These are specifically implemented for news headlines pertaining to gold by the authors.

Paper [3] explores the effect of using other indices to predict the price of gold such as other commodity prices, stock market indices and currency indices using a random forest method. They conclude that the two factors that were useful additions to predict the price of gold are Dow Jones Index and Standard & Poor's 500 Index. This project explores performance improvements by adding S&P500 and US Dollar indices to the gold price time series.

3 Dataset and Features

Gold price data was downloaded from the Yahoo Finance website that contains gold price from Aug 29th 2000 to November 8th 2022. The raw price data is normalized and converted to a time series data with a forecasting horizon of 9 days. In a similar manner, data for US Dollar Index and Standard and Poor's (S&P) 500 index was taken from Yahoo Finance, normalized and converted to a time series format. After combining the three time series inputs, it is split as 90% training data and 10% test data. The final dataset consisted of 4845 rows. Out of this, 4360 rows consisting of data from Sep 11, 2000 to Aug 12, 2020 is taken as training data and 485 rows consisting of data from Aug 13, 2020 to Nov 8, 2022 is taken as test data.

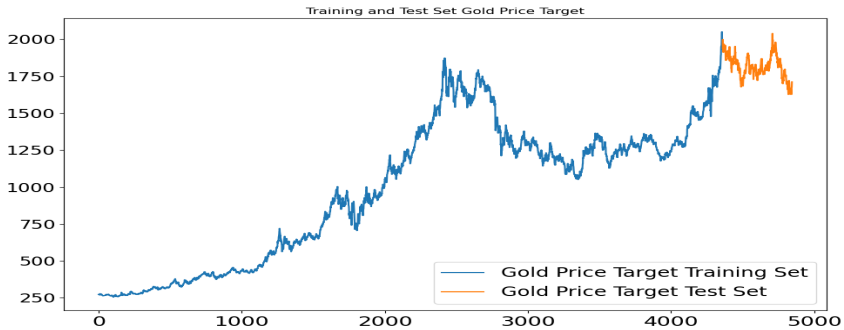

Figure 2: Training and Test Gold Price target values

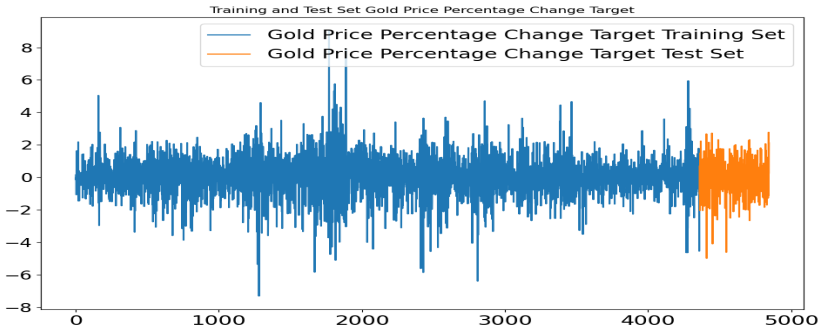

Figure 1: Training and Test Gold Percentage Change target values

The dataset for news sentiment analysis contains 11,412 human-annotated news headlines from around the world in the period 2000 to 2019, which are specifically about the gold commodity. This dataset was built by [2] by scraping news items from various financial news provider sites. The human annotations include the following flags: price direction (whether the headline indicates that prices goes up or down), past information (whether the headline talks about the past price of gold), future information (whether the headline talks about future price of gold) and price sentiment (whether the overall sentiment of the headline regarding gold price is positive, negative or neutral). From this dataset, the headlines which coincide with dates in the gold price data can be used.

4 Methods

The gold price prediction model is a CNN-LSTM model. In paper [1] the gold price time series inputs are fed to two convolutional layers followed by a max pooling layer and the reduced dimension output from the pooling layer is fed to an LSTM layer. The final output of the LSTM layer is passed through two fully connected layers to generate the regression output. ReLU non linearity is used for the convolution layers and LSTM layers.

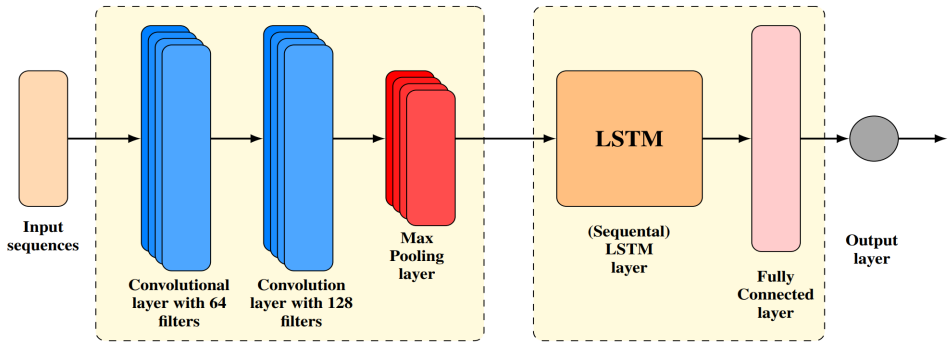

Figure 2: CNN-LSTM model architecture [1]

For predicting the price of gold, the inputs are the prices of gold for the past 9 days to predict the price on the 10th day. This is the baseline performance. The project also adds US Dollar index prices and S&P500 index prices for the past 9 days to the model to evaluate how performance is impacted by the addition of these inputs.

For predicting the percentage change in the price of gold, the inputs are the percentage change in the price of gold for the past 9 days, to predict percentage change on the 10th day compared to the 9th day. In addition to this, the percentage change in S&P 500 for the past 9 days is also added to the model to evaluate performance impact.

The model was trained for 50 epochs in each case, with ADAM optimization with default parameters and batch size of 32.

The project proposes to add the news sentiment indicator as an additional input to the fully connected layer. This is earmarked as future work.

5 Evaluation

The accuracy of the model's gold price prediction and the price percentage change prediction on the test set will be evaluated using Mean Squared Error as a metric.

The MSE of the model for the gold time series input will be taken as baseline performance. The effect on MSE after adding the other time series inputs will be compared against this baseline.

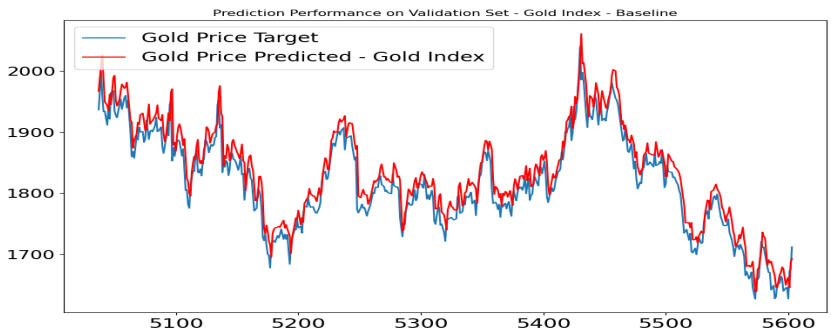

Figure 3: Baseline gold price prediction performance, MSE = 688.604

Figure 2 shows the baseline performance of the model on gold price prediction which uses only the gold price time series as input. The Mean Squared Error between the target gold price values and the predicted values is 688.604.

6 Results and Analysis

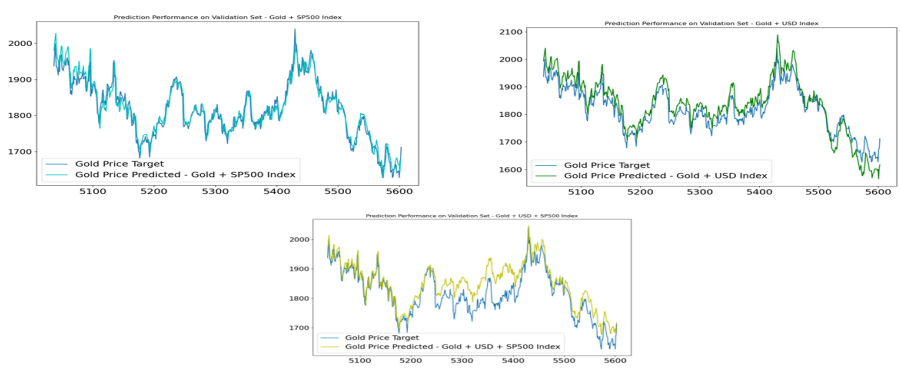

Figure 4: Gold price prediction performance with added inputs

Figure 3 shows the performance of the model when US Dollar Index and S&P500 indices are added as inputs. The additional time series are added as added channels to the convolutional layer. The Mean Squared Error values are shown in Table 1.

Inputs	MSE
Gold	688.608
Gold + USD	1601.45
Gold + SP500	**460.047**
Gold + USD + SP500	2083.24

Figure 5: Mean Squared Error Performance of different approaches