GOLD PRICE PREDICTION - MARKET SENTIMENT

ANBUSELVAN MAHALINGAM

BELLEVUE UNIVERSITY

DSC 630 PREDICTIVE ANALYTICS

ANDREW HUA

Introduction

GOLD is a precious metal, with the chemical symbol Au and atomic number 79. All the gold on earth came from meteorites that bombarded the planet over 200 million years after it formed. This project aims to predict gold prices by analyzing their correlation with the U.S. Dollar Index (DXY) and incorporating market sentiment analysis based on historical market-impacting events (e.g., global pandemics, geopolitical tensions, political events).

To develop the prediction model for gold prices, we will analyze and understand the relationship between gold prices and the DXY and assess the impact of global events on market sentiment and subsequently on gold prices.

Data Selection

We will be using historical GOLD prices, the historical U.S. Dollar Index, and global news headlines to predict market sentiment using sentiment analysis. The following data sources will be used for this analytical purpose:

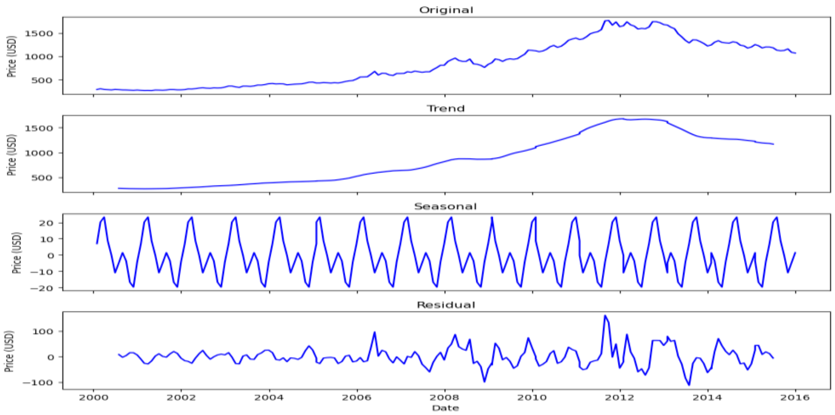
- Historical GOLD Price per Ounce, available from AURONUM gold price dataset for public usage.
- Historical US dollar index, available from Market Watch, for public usage.
- Market Sentiment dataset, news headlines Kaggle Stock Market Sentiment Analysis using NLP

Modeling & Methods

The purpose of this analytical project is to predict the GOLD price based on the historical gold price data and its relationship with dollar index and market sentiment, here due to the purpose and the nature of the dataset time series analysis and prediction model is recommended approach.

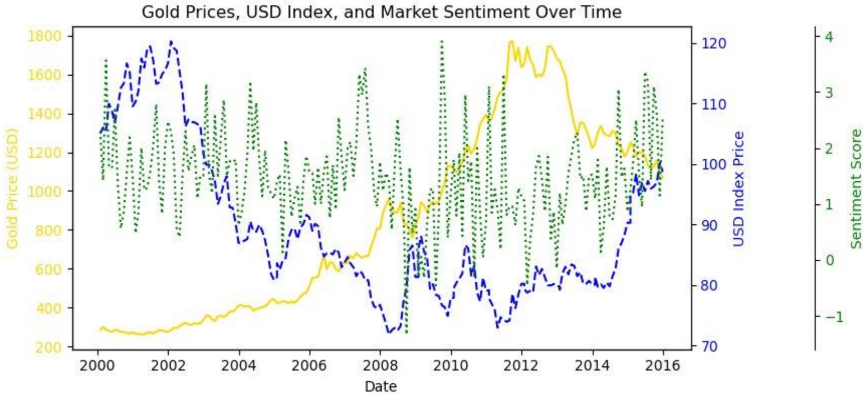
Exploratory Data Analysis (EDA)

To start with, we perform an Exploratory Data Analysis (EDA) to gain insights into the dataset. EDA involves summarizing the main characteristics of the data and visualizing trends, patterns, and anomalies. This helps in understanding the distribution, identifying missing values, and detecting outliers in the gold price and DXY datasets, below visualization to describe time series decomposition. There is a long-term upward trend and yearly seasonal cyclic fluctuations in GOLD price.



Correlation Analysis

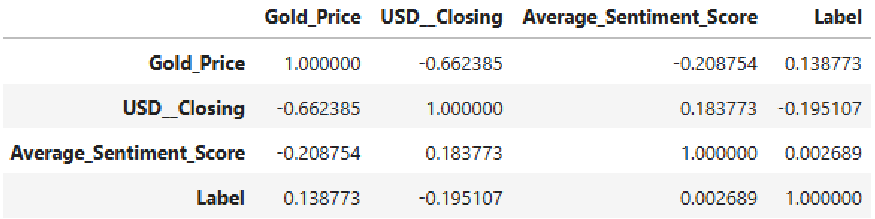
We conduct a correlation analysis to examine the relationship between gold prices and the U.S. Dollar Index (DXY). By calculating the correlation coefficient, we assess the strength and direction of the linear relationship between these two variables. This analysis helps in understanding how changes in the DXY may influence gold prices. Following are the correlation matrix details.



Gold Price and USD Closing (-0.192682) indicates a weak negative correlation. Typically, gold is viewed as a hedge against currency devaluation.

Gold Price and Average Sentiment Score (-0.202827) a weak negative correlation suggests that when market sentiment is positive gold prices might slightly decrease. Gold is often considered a "safe haven" during times of economic uncertainty, so better sentiment can reduce the demand for gold.

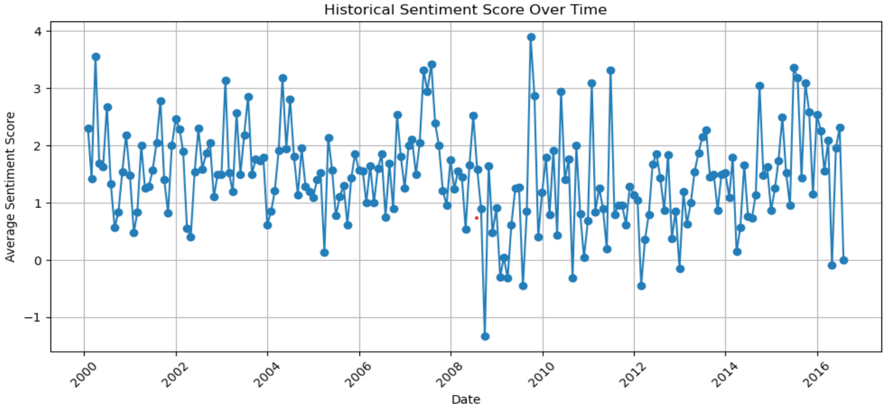
USD Closing and Average Sentiment Score (0.184210) a weak positive correlation indicates that a stronger U.S. dollar might coincide with better market sentiment.



	Gold_Price	USD__Closing	Average_Sentiment_Score	Label
Gold_Price	1.000000	-0.662385	-0.208754	0.138773
USD__Closing	-0.662385	1.000000	0.183773	-0.195107
Average_Sentiment_Score	-0.208754	0.183773	1.000000	0.002689
Label	0.138773	-0.195107	0.002689	1.000000

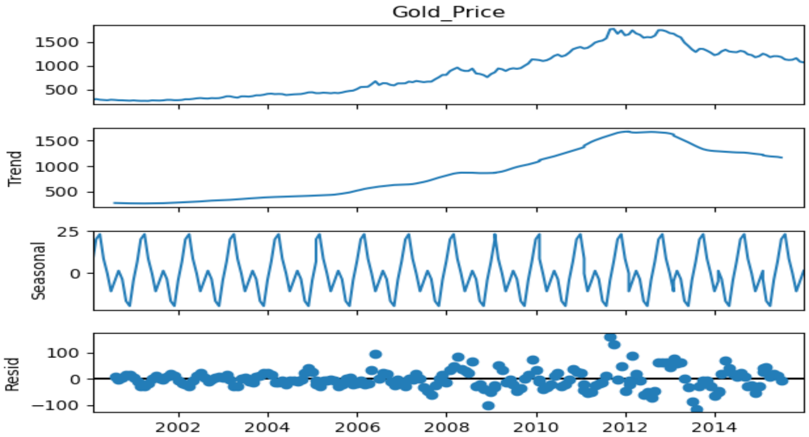
NLP for Sentiment Analysis

We leverage Natural Language Processing (NLP) techniques to analyze market sentiment from global news headlines. Sentiment analysis involves processing textual data to classify the polarity of opinions (positive, negative, or neutral). This analysis helps in quantifying market sentiment and incorporating it as a factor in our gold price prediction models. Below visualization is to describe historical aggregated sentiment score per month.



Stationarity Validation

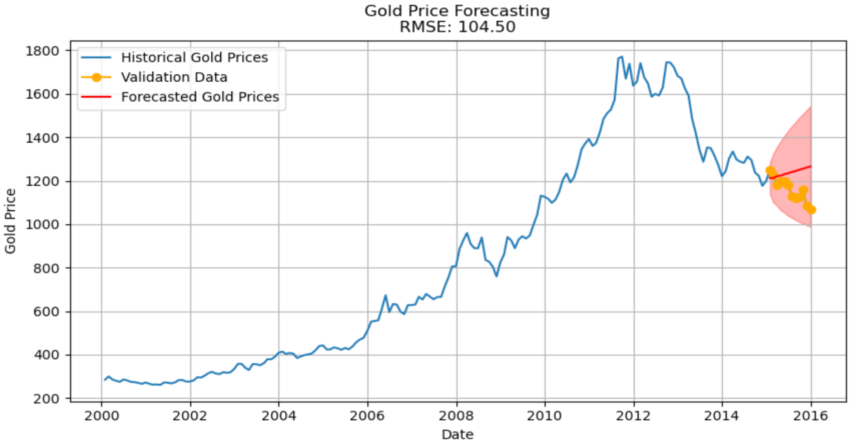
Stationarity is a key assumption in time series modeling. We validate the stationarity of the gold price data using statistical tests such as the Augmented Dickey-Fuller (ADF) test. If the data is non-stationary, we apply transformations (e.g., differencing) to achieve stationarity. Below visualization to confirm the data is stationary and good for time series forecasting.



```
ADF Statistic: -5.984356
p-value: 0.000000
```

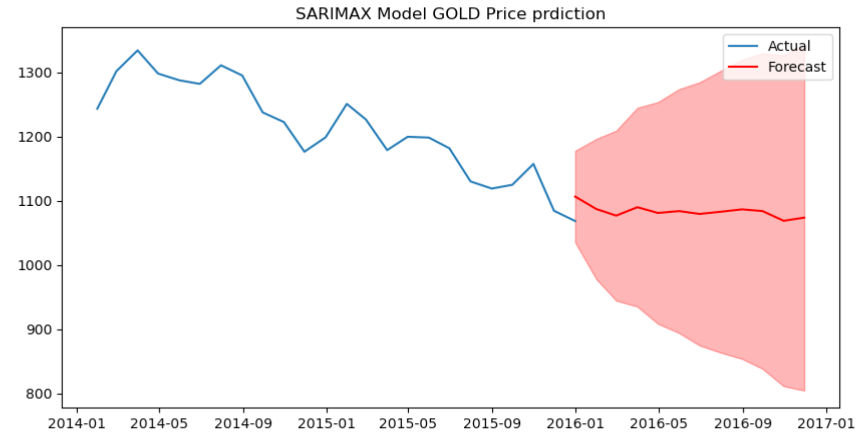
Approach 1: ARIMA Model for Time Series Analysis

The AutoRegressive Integrated Moving Average (ARIMA) model is used for predicting gold prices. ARIMA captures the autocorrelation within the time series data and is suitable for stationary datasets. We identified the optimal parameters (p, d, q) (1,1,1). Below visualization to describe the forecasting of ARIMA model.



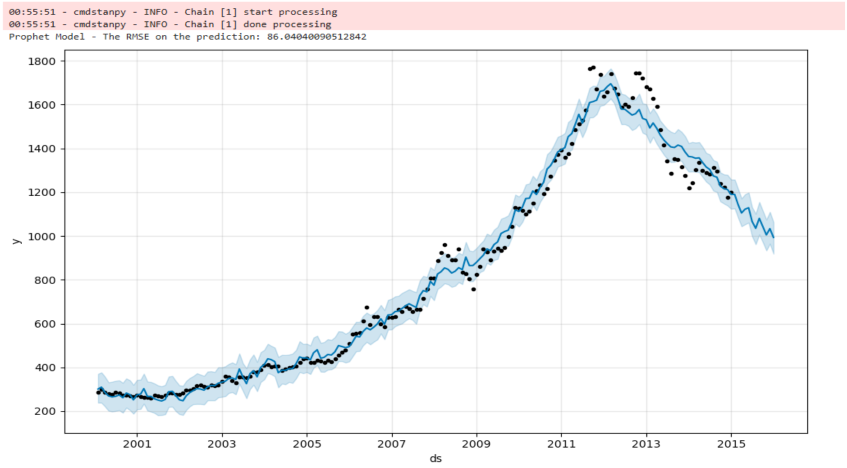
Approach 2: SARIMAX Model for Time Series Analysis

The Seasonal AutoRegressive Integrated Moving-Average with eXogenous factors (SARIMAX) model extends ARIMA by incorporating seasonality and external regressors (DXY, sentiment scores). SARIMAX is particularly useful for capturing both seasonal effects and the influence of external variables on gold prices. Below visualization to describe the forecasting of SARIMAX model.



Approach 3: Prophet Model for Predicting Gold Price

Prophet, developed by Facebook, is a robust forecasting tool designed for handling time series data with strong seasonal effects and missing data. It decomposes the time series into trend, seasonality, and holidays, making it suitable for predicting gold prices with seasonal patterns and incorporating external events. Below visualization to describe the forecasting of Prophet model.



```
00:55:51 - cmdstanpy - INFO - Chain [1] start processing
00:55:51 - cmdstanpy - INFO - Chain [1] done processing
Prophet Model - The RMSE on the prediction: 86.04040090512842
```

Approach 4: VAR + LSTM Approach

The Vector AutoRegression (VAR) model is combined with Long Short-Term Memory (LSTM) neural networks to predict gold prices. VAR captures the interdependencies between multiple time series (gold prices and DXY), while LSTM, a type of recurrent neural network, excels in modeling sequential data and capturing long-term dependencies. This hybrid approach leverages the strengths of both VAR and LSTM to enhance prediction accuracy.

Results Interpretation

We use Root Mean Square Error (RMSE) and prediction visualization. The following are the RMSE values for each of the selected models:

- ARIMA Model: The RMSE for the predictions is 104.50.
- SARIMAX Model: The RMSE for the predictions is 89.96.
- Prophet Model: The RMSE for the predictions is 86.04.
- Hybrid VAR+LSTM Model: The RMSE for the predictions is 343,517,591.62.

A lower RMSE value implies higher prediction accuracy, for example the RMSE value for the SARIMAX model, at 89.96, indicates that the model's predictions of gold prices have an average deviation of approximately 89.96 units from the actual observed values. This relatively low RMSE value suggests that the SARIMAX model performs well in capturing the underlying patterns in the data, including seasonal variations and the influence of external factors like the U.S. Dollar Index (DXY) and market sentiment.

Conclusion

Based on the RMSE values, the Prophet model emerges as the best-performing model, offering the most accurate predictions of gold prices. Its ability to handle seasonal patterns and external events effectively sets it apart.

The SARIMAX model also demonstrates strong performance, especially when accounting for external factors influencing gold prices. While it is slightly less accurate than the Prophet model, it provides valuable insights by incorporating market sentiment and the U.S. Dollar Index.

The ARIMA model, although effective, lags both Prophet and SARIMAX in terms of prediction accuracy. Its simpler approach does not capture external influences as well as the other models.

The Hybrid VAR+LSTM model presents significant challenges, as indicated by its high RMSE. This suggests that further refinement and addressing potential issues are necessary before it can be deemed a reliable forecasting method.

Integrating superior market sentiment data holds great potential for enhancing the accuracy of gold price prediction models. The SARIMAX model, in particular, has shown improvement with basic sentiment data, and incorporating higher quality, more granular sentiment information could further refine its predictions. By focusing on obtaining and utilizing advanced sentiment data, we can develop a more accurate and responsive forecasting system, leading to improved decision-making and better financial outcomes.

Reference

- Market Sentiment dataset, news headlines Kaggle- Stock Market Sentiment Analysis using NLP, obtained from Stock Market Sentiment Analysis using NLP (kaggle.com)
- ARIMA time series analysis and prediction, reference Machine Learning plus - ARIMA
- SARIMAX time series analysis and prediction, reference Geeks for Geeks - SARIMAX
- PROPHET Forecasting at scale, reference facebook github io - prophet
- Vector AutoRgressive (VAR) Machine learning model, reference Machine Learning plus - VAR
- Long Short-Term memory recurrent neural network machine learning – reference Machine Learning Mastery - LSTM