Gold Price Prediction Using Time Series Forecasting

Keikiet Pham

EC Utbildning

Projekt i Data Science

202411

Abstract

This project leverages time series forecasting techniques to predict gold prices, focusing on the Seasonal Autoregressive Integrated Moving Average (SARIMA) and SARIMAX models. The data used spanned from 2009 to 2023 on daily gold prices were resampled to monthly averages to handle volatility and cyclicity in the dataset. Initial analysis revealed significant trends and seasonality, requiring differencing to achieve stationarity, confirmed by the Augmented Dickey-Fuller test. Autocorrelation plots guided model selection, resulting in a SARIMAX (1, 1, 1) × (1, 1, 1, 12) model, evaluated using AIC and BIC. The model effectively captured trends and seasonality but struggled with sudden market shifts. Future work could improve accuracy by incorporating economic indicators and advanced machine learning methods.

Table of Contents

1 Introduction

Time series forecasting is a key tool in predictive analytics, widely used in fields like finance, economics, and health care. It helps us predict future values of a variable based on patterns observed in past data. This method is especially useful in areas like stock market analysis, sales forecasting, and weather predictions. In this project, the goal was to forecast gold prices by using advanced time series models that can capture both trends and seasonal patterns in the data (www.geeksforgeeks.org).

Gold is often seen as a safe investment, especially in times of economic uncertainty, and its price has varied greatly over the years due to economic, political, and financial events. Accurately predicting these price changes can provide significant benefits for investors and financial analysts (economicobservatory.com).

To improve our model's accuracy, we incorporated external economic factors, specifically Year-on-Year (YoY) inflation, using a Seasonal Autoregressive Integrated Moving Average with exogenous factors (SARIMAX) model. By including inflation as an exogenous variable, we aimed to better understand the relationship between macroeconomic trends and gold prices.

This report explains the methods we used, including the theoretical reasoning behind SARIMA and SARIMAX models, the significance of incorporating inflation data, and the key insights gained from our analysis.

2 Background and Concepts

2.1 What is Time Series Forecasting?

Time series forecasting involves analyzing data points collected over time to predict future values. This approach is widely used in fields such as finance, economics, and resource management. Understanding the components of a time series—trend, seasonality, cyclic patterns, and noise—is crucial for effective modeling.

Trend refers to the general direction in which data moves over a long period, either upward or downward. For example, a steady increase in gold prices over several years indicates a trend. Seasonality, on the other hand, describes regular patterns that repeat at consistent intervals, such as monthly or annually. Cyclic patterns differ from seasonality as they do not follow a fixed period and are often influenced by broader economic cycles. White noise represents random variations that are unpredictable and not explained by trends, seasonality, or cycles (www.geeksforgeeks.org).

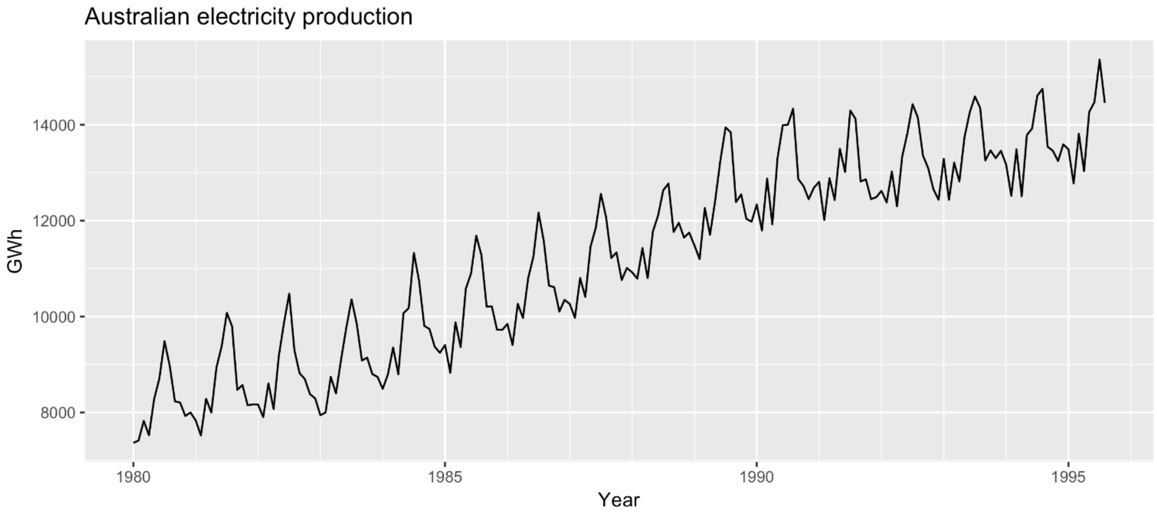


Figure 1. Example of time series patterns. The Australian electricity production is clearly trended with a change in the slope of the trend around 1990. Notice the seasonal pattern where there seems to be a spike in production once every year.

2.2 SARIMA and SARIMAX Models

The SARIMA model (Seasonal Autoregressive Integrated Moving Average) is a powerful tool for capturing both non-seasonal and seasonal patterns in time series data. It is characterized by parameters (p, d, q) for non-seasonal components and (P, D, Q, m) for seasonal components. The SARIMAX model extends SARIMA by incorporating exogenous variables, such as year-on-year inflation, to account for external influences (www.geeksforgeeks.org).

2.3 Stationarity

Stationarity is a key assumption in time series modeling, where the mean, variance, and autocorrelation must remain constant over time. Non-stationary data often needs transformations like differencing to achieve stationarity. The Augmented Dickey-Fuller (ADF) test is commonly used to assess stationarity and determine whether further differencing is needed(www.geeksforgeeks.org).

2.4 Differencing, Integrated (I) component

Differencing stabilizes the mean of a series by calculating the differences between consecutive observations, thereby removing trends or seasonal effects. In some cases, a second differencing or seasonal differencing is necessary to fully stabilize the data. The "Integrated" (I) component in SARIMA refers to this differencing process (Hyndman & Athanasopoulos, 2021).

2.5 Autocorrelation and Partial Autocorrelation

Autocorrelation measures how a time series relates to its past values, while partial autocorrelation isolates the relationship between a series and its lagged values. These concepts are crucial for selecting appropriate AR and MA terms in a model. The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots guide the determination of lag orders for autoregressive and moving average components (Hyndman & Athanasopoulos, 2021).

2.6 Autoregressive (AR) Component & Moving Average (MA) Component

The Autoregressive (AR) component predicts future values based on past observations, while the Moving Average (MA) component models the influence of past errors. Together, they enable SARIMA to handle trends and random variations effectively. Seasonal extensions allow the model to adapt to patterns that repeat over time, making SARIMA suitable for data with pronounced seasonal cycles (Hyndman & Athanasopoulos, 2021).

2.7 Seasonality

Seasonality is the presence of regular and predictable patterns in a time series that occur at consistent intervals, such as every year, month, or week. In the case of gold prices, seasonality might be driven by recurring global economic events or traditional investment cycles. Models like SARIMA account for these patterns by including seasonal components, which help the model make accurate forecasts by recognizing and adapting to these periodic behaviors (Hyndman & Athanasopoulos, 2021).

2.8 Exogenous Variables

Exogenous Variables are external factors that influence a time series but are not part of its own historical values. For example, in predicting gold prices, year-on-year inflation can be used as an exogenous variable, as economic conditions like inflation often impact gold's market value. Incorporating exogenous variables through the SARIMAX model may enhance the prediction accuracy by allowing the model to consider additional relevant information beyond the time series data alone (Hyndman & Athanasopoulos, 2021).

3 Methodology

This project adopted a traditional approach, consisting of data preparation, exploratory analysis, and model development using advanced time series forecasting techniques. The main steps involved are detailed as follows.

3.1 Data Collection and Preparation

The dataset used for this analysis consisted of historical gold prices spanning from 2009 to 2023, with initial daily closing prices. Given the objective to forecast long-term trends while minimizing the impact of short-term fluctuations, the daily data was resampled into monthly averages.

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was crucial for uncovering the inherent patterns and characteristics of the gold price data from 2009 to 2023. The initial visual inspection of the time series revealed pronounced long-term trends and potential seasonal variations, likely influenced by major global economic events. These observations were represented using time series plots, which highlighted both the overall upward movement and periods of fluctuation.

The visual evidence strongly indicated non-stationarity, characterized by clear upward and downward trends, suggesting that the mean and variance of the series were not constant over time. Additionally, the data displayed patterns that might be associated with economic cycles, white noise, and seasonal effects. The effects of significant global events, such as the financial crises and the COVID-19 pandemic, were particularly notable, manifesting as sharp price movements.

Due to the complexity of capturing cyclical and noisy patterns, we chose to limit the analysis to data starting from 2015. This decision was made to minimize the impact of earlier cyclical behavior and to simplify the model's assumptions. Furthermore, to reduce the noise inherent in daily observations and to highlight more stable trends, we resampled the data to monthly averages. This resampling also helped in modeling the data more efficiently by smoothing out short-term fluctuations, thus serving as a form of moving average.

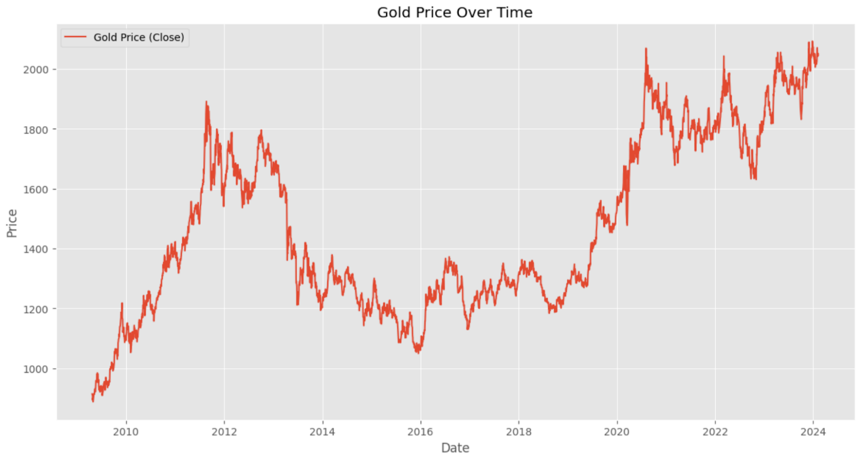


Figure 2 Gold Price Over Time before data processing

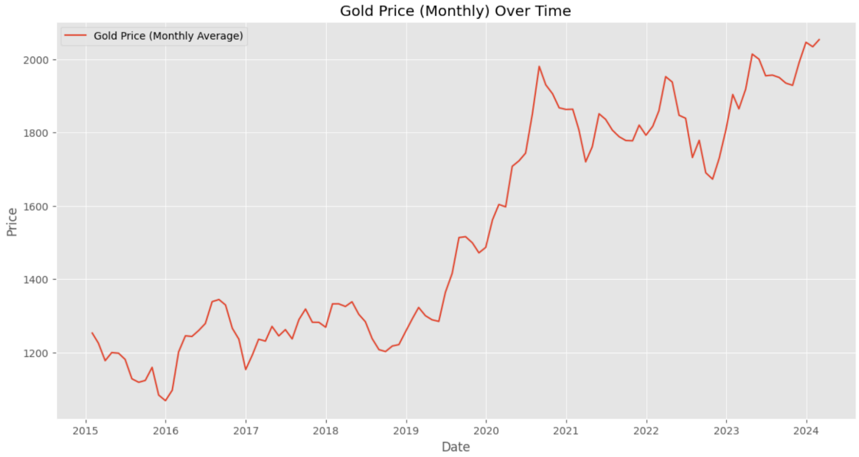


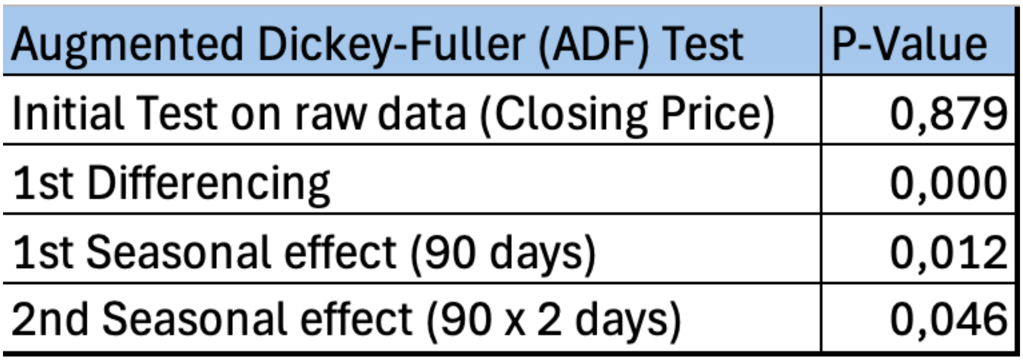
Figure 3 Gold Price Over Time after data processing

3.3 Stationarity Tests and Differencing

To ensure that the dataset met the crucial requirement of stationarity for effective time series modeling, the Augmented Dickey-Fuller (ADF) test was utilized. The initial ADF test performed on the raw closing price data indicated non-stationarity, evident from a high p-value of 0.879. This result suggested that the null hypothesis of non-stationarity could not be rejected.

To address this issue, first-order differencing was applied. This transformation stabilized the mean and significantly reduced the p-value to 0, indicating that the series had become stationary. Furthermore, to handle yearly recurring patterns, seasonal differencing was implemented. The first seasonal differencing yielded a p-value of 0.012, demonstrating stationarity at the seasonal level.

While a second seasonal differencing was considered, it resulted in a p-value of 0.046, which is below the significance level of 0.05. However, implementing further differencing risked over-differencing the data, which could have introduced noise and diminished the model's predictive power. Therefore, the chosen differencing strategy effectively captured and removed the trend and seasonal components, preparing the data for robust time series modeling.



Augmented Dickey-Fuller (ADF) Test	P-Value
Initial Test on raw data (Closing Price)	0,879
1st Differencing	0,000
1st Seasonal effect (90 days)	0,012
2nd Seasonal effect (90 x 2 days)	0,046

3.4 Determining AR and MA components

The ACF and PACF plots were used to understand the patterns in the data and guide the choice of model parameters. The non-seasonal ACF plot (figure 4) shows a quick drop-off after the first lag, indicating that the series doesn’t have strong dependencies beyond the immediate past value. The PACF plot also has a clear spike at lag 1, suggesting that only a simple autoregressive term might be needed.

The seasonal ACF plot (figure 5) reveals a slow decay, pointing to strong seasonal patterns over time. The PACF plot has a couple of clear spikes, suggesting the need for a seasonal autoregressive component to capture these regular cycles. Overall, these observations helped determine the necessity of both non-seasonal and seasonal terms in the SARIMA model.

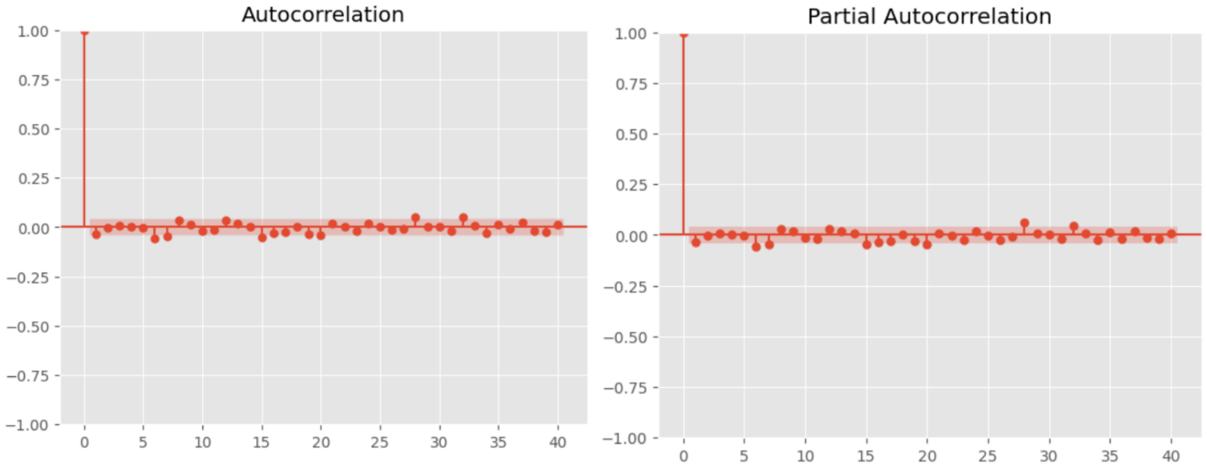


Figure 4 Non-Seasonal ACF and PACF – The sharp decline after lag 1 in the ACF and a single significant spike in the PACF suggest minimal non-seasonal dependencies.

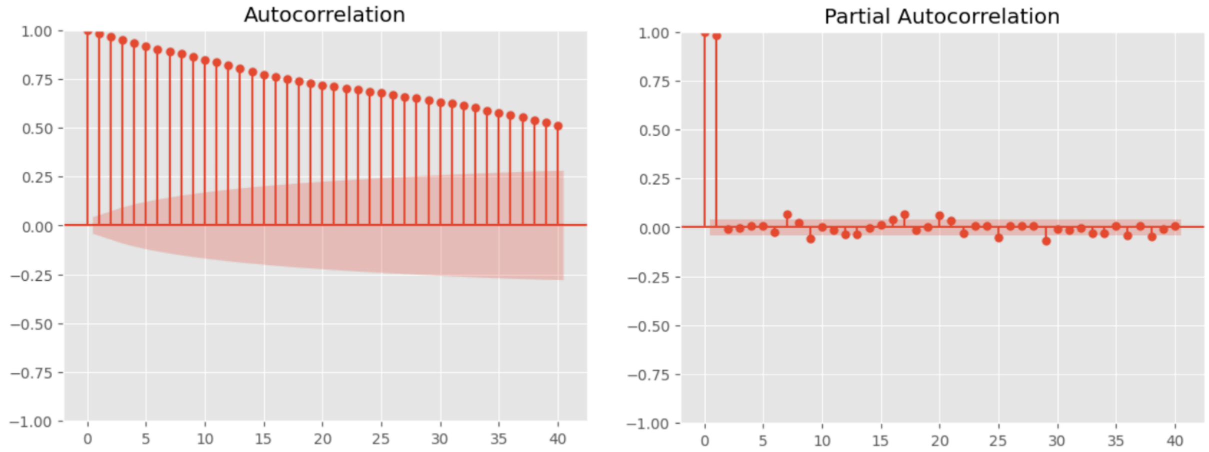


Figure 5 Seasonal ACF and PACF – The slow decay in the ACF and clear spikes in the PACF indicate strong seasonal patterns, requiring seasonal modeling components.

3.5 Data Splitting and Model Fitting

The dataset was divided into training and testing sets to evaluate the model's forecasting performance effectively. A split ratio of 80:20 was used, allocating 80% of the data for training and 20% for testing. Specifically, the training set contained 88 data points, while the test set comprised 22 data points. This approach ensured that the model had sufficient data to learn from historical trends and patterns while retaining enough unseen data for performance assessment.

The SARIMA model was constructed using the SARIMAX function from the statsmodels library. The chosen parameters for the model were set as follows: p = 0, d = 1, q = 0, for the non-seasonal component, and P = 1, D = 1, Q = 1, m = 12 for the seasonal component, where m represents the seasonal period of 12 months.

4 Modeling and Results

4.1 Building the SARIMA Model

First we built a SARIMA model. Given the observed seasonality, a SARIMA model was chosen, with parameters (p, d, q) for non-seasonal components and (P, D, Q, m) for seasonal components. The final model configuration was SARIMA(1, 1, 1) x (1, 1, 1, 12). The model was trained using the monthly gold prices, and various performance metrics, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), were used to evaluate its fit. AIC and BIC was 1223 and 1229

4.2 Building the SARIMAX Model

Recognizing that economic indicators like inflation can impact gold prices, we extended the SARIMA model to a SARIMAX model. Year-over-Year (YoY) inflation was used as an exogenous variable to improve forecasting. The SARIMAX model was configured with the same seasonal and non-seasonal parameters, (0, 1, 0) x (1, 1, 1, 12), and the exogenous inflation data. This model aimed to account for external economic factors that could influence gold price trends. The summary of the SARIMAX model showed a significant coefficient for the inflation variable, with a p-value of 0.040, suggesting that inflation has a meaningful impact on gold price predictions. The AIC for this model was slightly lower at 846, and 855 for BIC indicating a better fit compared to the SARIMA model.

4.3 Forecasting Results

Both models were used to forecast gold prices for the next 12 months. The forecasted values from the SARIMAX model, which included inflation, showed slightly better alignment with the observed trends compared to the SARIMA model. The forecast plots (appendice 1 & 2) highlighted that while both models captured the general seasonal pattern and trend, sudden market shifts or economic shocks could still cause discrepancies.

5 Discussion and Insights

The inclusion of YoY inflation in the SARIMAX model provided valuable insights into the role of economic indicators in gold price forecasting. The significant coefficient for inflation suggests that macroeconomic factors play a crucial role in shaping gold price movements. While the SARIMA model captured the core patterns well, the SARIMAX model demonstrated the advantage of incorporating external variables. However, challenges remain in dealing with the inherent volatility of gold prices, especially during periods of economic uncertainty. These findings underscore the importance of considering both internal patterns and external influences in financial time series modeling.

6 Conclusion and Future Work

This project applied time series forecasting techniques to predict gold prices, focusing on understanding both trend and seasonal behaviors. The SARIMA model served as a baseline, effectively capturing core seasonal and trend components. However, integrating external economic variables like Year-on-Year (YoY) inflation into a SARIMAX model proved to be a superior approach. The SARIMAX model showed better alignment with observed values, underscoring the importance of including exogenous factors that influence market behavior.

Despite the SARIMAX model's improved performance, challenges like gold price volatility remain. This is especially evident during economic disruptions, which traditional models struggle to predict. Moving forward, future work could explore additional exogenous variables, such as global interest rates or geopolitical risk indicators. Transformations, like logarithmic scaling, could also be considered to stabilize the data further. Experimenting with machine learning approaches, including LSTM networks, may offer a more robust solution for capturing non-linear relationships and handling complex data patterns.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Under projektet stötte jag på flera utmaningar, särskilt att förstå och använda de avancerade Time Series modellerna SARIMA och SARIMAX. Det var svårt att hitta rätt parametrar och göra datan stationär. Jag hanterade detta genom att läsa in mig på teorin, utföra många tester och använda verktyg som ACF- och PACF-plots för att justera modellen.

2. Vilket betyg du anser att du skall ha och varför.
   Jag tycker att jag förtjänar ett VG. Jag har lagt ner mycket tid på att förstå och använda avancerade metoder inom Time Series Analysis/Forecasting, ett komplext område som kräver fördjupning. Även om jag har lärt mig mycket känner jag att det finns mycket mer att utforska och förstå.

3. Något du vill lyfta fram till Antonio?
   Jag tycker det var roligt att få välja ämne för projektet själv. Time Series Analysis/Forecasting är väldigt intressant, även om det har varit frustrerande när jag inte har förstått allt direkt.

Appendix 1

SARIMA (p, d, q) = (0, 1, 0) (P, D, Q, m) = (1, 1, 1, 12)

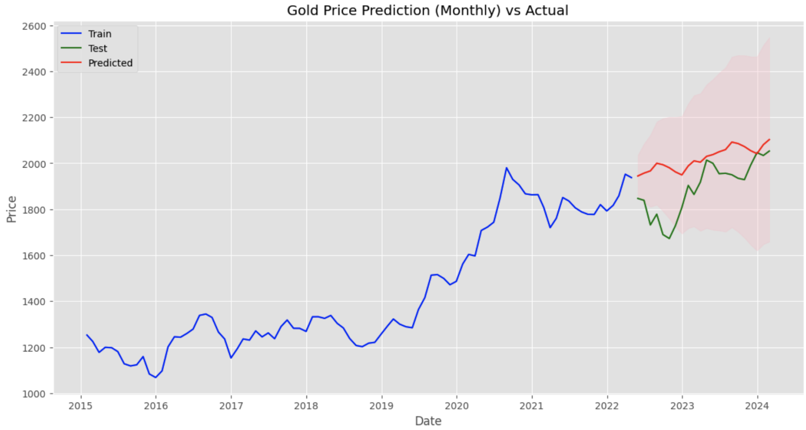


Table 1 Sarima Model Forecast: The blue line is the training data, the green is the test set, and the red is the model's prediction, with shaded areas representing the confidence interval.

Appendix 2

SARIMAX (p, d, q) = (0, 1, 0) (P, D, Q, m) = (1, 1, 1, 12)

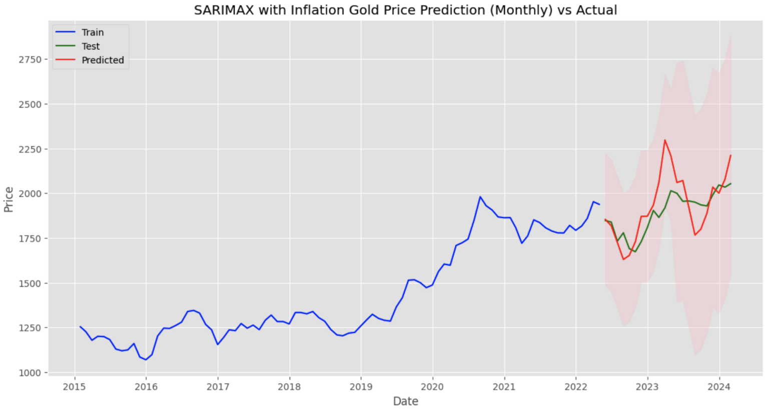


Table 2 SARIMAX Model Forecast with Inflation The blue line is the training data, the green is the test set, and the red is the model's prediction, with shaded areas representing the confidence interval.

Sources

1. **Geekforgeeks**. (2024). Time Series Analysis and Forecasting. Retrieved from https://www.geeksforgeeks.org/time-series-analysis-and-forecasting/
2. **EconomicObservatory**. (Philip Fliers, 2024) Retrieved from https://www.economicsobservatory.com/is-gold-a-safe-haven-for-investors
3. **Hyndman, R.J., & Athanasopoulos, G.** (2021). *Forecasting: principles and practice* (3rd edition). OTexts: Melbourne, Australia. Retrieved from https://otexts.com/fpp3 [2024.10].

Additional Sources Used for Inspiration and Learning

- Effective Time Series Forecasting: Gold Future Prices on Kaggle - https://www.kaggle.com/code/gvyshnya/effective-ts-forecasting-gold-future-prices
- Ritvik Math's Time Series Analysis Playlist on YouTube - https://www.youtube.com/playlist?list=PLvcbYUQ5t0UHOLnBzl46_Q6QKtFgfMGc3
- Michael Cortes' Tutorial on Time Series Forecasting - https://www.youtube.com/watch?v=tr8PF2v9Wgo&ab_channel=MichaelCortes