Zhi Qiang Zheng & Marckenrold Cadet

Zhizheng0889@floridapoly.edu

Abstract

Forecasting is an old but always important field of study for humanity. Forecasting is the study of predicting what can happen in the future. Forecasting with time series method is one of the most used techniques in the field of data science. In this project, we are applying the time series techniques to the domain of forecasting precious metal stock prices.

TIME SERIES FORECASTING OF PRECIOUS METAL STOCK PRICES

AN EMPIRICAL ANALYSIS USING ARIMA, SARIMA, AND EXPONENTIAL SMOOTHING MODELS

Zhi Qiang Zheng & Marckenrold Cadet

Professor Hajibabaee

CAP 5771 Data Mining and Text Mining

11/21/2023

**TIME SERIES FORECASTING OF PRECIOUS METAL STOCK PRICES: AN EMPIRICAL ANALYSIS USING ARIMA, SARIMA, & EXPONENTIAL SMOOTHING MODELS**

1. **Introduction**:

Throughout human history, precious metals have held a timeless allure, symbolizing wealth and enduring value. Even in the modern civilized world, the quest for precious metals continues. Consequently, the forecasting of precious metal stock prices has become a captivating and valuable area of study. In this project report, we embark on a journey into the fascinating realm of predicting precious metal prices.

Our adventure begins with the discovery of a dataset on Kaggle—'all\_commodities\_data.csv,' generously provided by Guillem Servera (Servera). As we prepare for this expedition, we will delve into essential data preprocessing and Explanatory Data Analysis (EDA). Subsequently, we will venture into the modeling and forecasting phase of the project. Finally, we will thoroughly analyze our results and evaluate our findings.

This report seeks to shed light on the intricate world of precious metal price forecasting, with the ultimate goal of providing valuable insights into this captivating financial domain. The forecasting goal of this project is to forecast three months into the future for the monthly mean closing precious metal price.

1. **Data Preprocessing and Exploratory Data Analysis (EDA)**

Before going on the journey of forecasting precious metal prices, it is vital to prepare and understand the dataset. This section outlines the key steps taken during data preprocessing and exploratory data analysis.

**2.1 Data Loading**

The adventure begins with the acquisition of our dataset, which is stored in a CSV file. We leverage the Pandas library to load the data into our analysis environment, making it accessible for further exploration.

**2.2 Data Transformation**

To effectively work with time series data, we convert the 'date' column into a datetime data type. This transformation ensures that we can handle chronological data effortlessly. Then, convert the daily data into monthly data for predicting the mean closing price of precious metals prices.

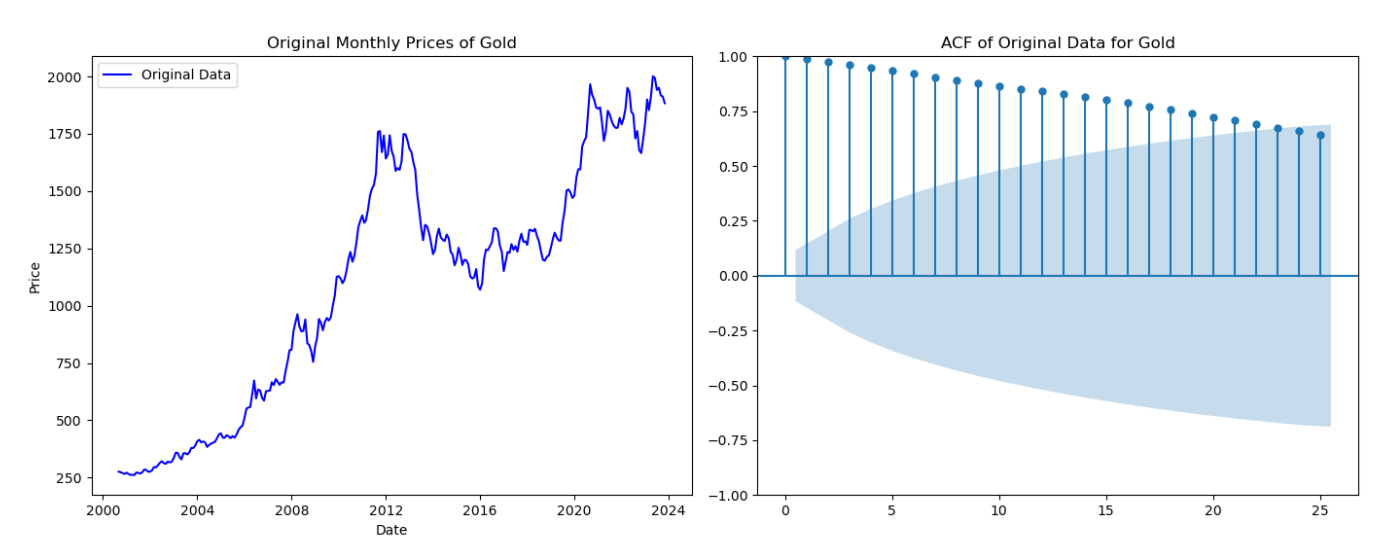
**2.3 Data Sorting**

Chronological order is paramount in time series analysis. Thus, we meticulously sorted the dataset based on the 'date' column, guaranteeing that our data is arranged in chronological order.

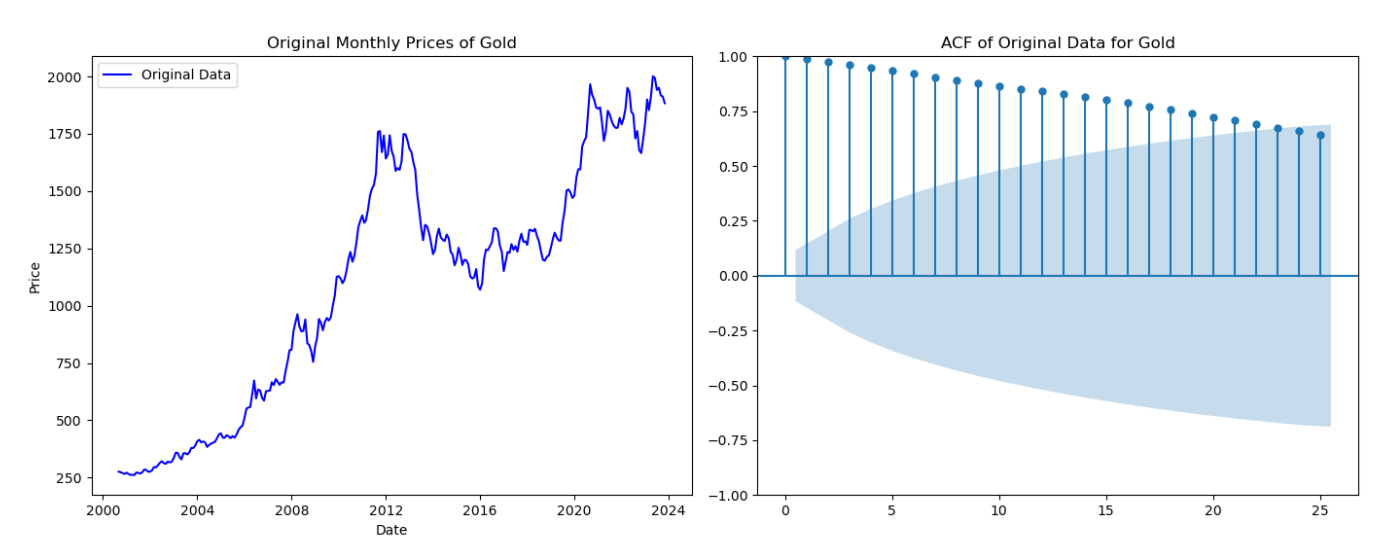
**2.4 Time Series Decomposition and Visualization**

To gain a deeper understanding of each precious metal's behavior, we employ time series decomposition. This process dissects the data into its fundamental components, including trend, seasonality, and residual.

Visualizations play a pivotal role in our exploration. Line plots are employed to reveal historical price trends for each commodity, shedding light on their past performance. Additionally, the autocorrelation function (ACF) is utilized to uncover potential seasonality patterns within the data.



Plot 1: The Times Series Line For Gold Price



Plot 2: Autocorrelogram of Gold Price

The two charts displayed here represent a portion of the comprehensive data visualization available, with additional plots included in the appendix at the conclusion of the report. These charts illustrate the dynamic nature of gold prices over time without any obvious seasonal patterns. This observation is also consistent across the other four precious metals, for which detailed visualizations are provided in the appendix.

**2.5 Stationarity Analysis**

Earlier, EDA led us to check the stationarity for all five precious metal prices. Stationarity is a crucial concept in time series analysis. To analyze stationarity, we employ the Augmented Dickey-Fuller Test (ADF Test). Both the original time series data and differenced time series data are subjected to this test. The results guide us in determining whether further transformations are necessary for modeling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precious Metal | Original Data Test Statistic | Differenced Data Test Statistic | Original Data p-value | Differenced Data p-value |
| Gold | -0.8928 | -13.8116 | 0.7903 | 8.21e-26 |
| Silver | -1.7862 | -6.2981 | 0.3873 | 3.47e-08 |
| Copper | -2.0812 | -4.9577 | 0.2521 | 2.69e-05 |
| Platinum | -1.9760 | -7.6571 | 0.2972 | 1.73e-11 |
| Palladium | -1.8479 | -4.4107 | 0.3569 | 0.0003 |

Table 1: Augmented Dickey-Fuller Test

As Table 1 above shows, the original data p-values are all greater than 0.05, indicating that all the data are non-stationary. After differentiating all five of the precious, all the data became stationary with the p-value less than 0.05.

**2.6 Additional Data Preprocessing**

In addition to the previously mentioned data preparation steps, further data preprocessing has been done for the modeling and forecasting process. We split the dataset into two parts: one for training (from January 2018 to December 2022) and another for testing (from January 2023 to March 2023). This separation helps us rigorously evaluate our models, just like they will be used in the real world.

Lagged features were constructed for specific models to capture time-dependent relationships within the data. Additionally, data normalization and standardization techniques were applied to selected models to ensure uniform scales among variables, thereby improving model robustness and accuracy.

1. **Methodology**

Finally, we can start the exploration of using different forecasting models to forecast monthly mean closing prices. In this process, we explore the forecasting with Naïve, ARIMA, SARIMA, and Exponential Smoothing models.

**3.1 Naïve Models**

Naïve model is the simplest form of forecasting, serving as the baseline for our project to compare performance. The naïve model assumes the most recent data observation to be the best predictor for the forecasting values. The surprisingly good performance of the Naïve model in the case of short-term forecasting establishes it as an effective baseline measurement tool. However, it does not account for any data pattern like trends or seasonality. Therefore, it is an underfitted model for forecasting, making it beneficial as a benchmark but not trustworthy for any serious forecasting problem.

**3.2 ARIMA**

The ARIMA (Autoregression Integrated Moving Average) model is an outstanding and commonly used statistical forecasting model in the field of time series analysis. It is a combination of:

* **Autoregressive (AR) Component (p)**: This part captures the relationship between an observation and a specified number of lagged observations (Harper).
* **Integrated (I) Component (d)**: This component represents the differencing of observations to make the time series stationary, meaning it has constant properties over time, like mean and variance.
* **Moving Average (MA) Component (q)**: This component models the relationship between an observation and a residual error from a moving average model applied to lagged observations (Farr).

The ARIMA model is helpful for data that shows a clear trend or autocorrelation when these elements are stationary. It is widely applied in financial time series forecasting due to its flexibility and ability to model complex temporal structures. However, ARIMA does not inherently handle seasonal patterns, and it assumes a linear relationship between variables.

**3.3. SARIMA**

The SARIMA Model (Seasonal Autoregressive Integrated Moving Average) is an extension of the ARIMA model, specifically designed to handle the seasonality in time series forecasting. Thus, the model is a combination of:

* + **Autoregressive (AR) Component (p):** This component allowed the model to learn the feedback mechanism within the time series data. It was determined by examining significant spikes in the PACF plot up to a certain lag, which indicates the number of lag observations included in the model.
  + **Differencing (I) Component (d):** The differencing steps were applied to make the time series stationary. The order of differencing was chosen based on the number of transformations required to stabilize the series' mean.
  + **Moving Average (MA) Component (q):** This component was included to model the noise or shock effects in the time series, which was observed in the ACF plot as significant spikes.
  + **Seasonality (S) Component (s):** The seasonality component in a SARIMA model accounts for the recurring patterns or cycles that occur at regular intervals in the time series data. It is essential to identify and model seasonality correctly to make accurate forecasts for data with seasonal patterns.

**3.4 Exponential Smoothing**

Exponential Smoothing is a straightforward yet effective method for forecasting time series data, particularly useful when the data shows trends or seasonality. We used triple exponential Smoothing (Holt-Winters) in the project that models for data with both trends and seasonal patterns as an additional exploration into the forecasting domain.

**3.5 Modeling and Tuning**

The journey of forecasting commodity prices unfolded as an exploratory expedition through the realm of time series analysis, each phase marked by different models and methodologies that progressively built on each other.

Initially, the journey began with the Naïve model, which served as the first step into this intricate world. This simple model provided initial guidance by establishing benchmark performance metrics such as RMSE, MAE, MSE, and MAPE for commodities such as gold, silver, copper, platinum, and palladium. The data from January to March 2023 were the terrain on which this model tested. While the Naïve model offered valuable initial insights, its simplicity was akin to navigating through complex terrain with a basic compass, revealing the need for more sophisticated tools to delve into the nuances of the data.

The journey then progressed to vector autoregression (VAR) modeling, a more advanced technique capable of navigating the multivariate nature of commodity data. Spanning the period from January 2018 to December 2022, the VAR model embarked on a quest to decipher the interdependencies between various market indicators. This phase of the journey was akin to mapping the intricate relationships within the data. However, despite its analytical prowess, results from the VAR model indicated that the path to accurate forecasting was still surrounded by complexity.

In search of greater precision, the expedition turned to looping through different parameters with Auto ARIMA, using the pmdarima library and manual code for exponential Smoothing. This phase represented a significant leap forward, similar to the use of advanced navigation tools that automatically selected the best route through the complex landscape of ARIMA modeling. Auto ARIMA embarked on a systematic exploration of parameter combinations, searching for the statistical sweet spot for the most accurate model. This sophisticated approach marked a transition from a broad terrain scan of data to a specific data-driven path, focusing on the most promising paths for accurate forecasting.

Throughout this journey, each model and method served as a foundation for a deeper understanding of the complex and dynamic nature of forecasting commodity prices. The transition from Naïve to VAR, and then to Auto ARIMA, reflected the progression from basic exploration to a more complex and specific expedition, mirroring the evolving landscape of time series analysis in finance.

1. **Results**

The journey through various stages of time series forecasting for commodity prices culminated in revealing and insightful results. Each model applied—Naïve, VAR, ARIMA, SARIMA, and Exponential Smoothing—offered distinct perspectives on the dataset's behavior.

**4.1 Insights from Naïve Forecasting:**

The initial phase with the Naïve model established baseline metrics for the commodities over the first quarter of 2023. For instance, gold exhibited an MAE of 108.71 and a MAPE of 5.56%, while palladium showed a significantly higher MAE of 401.53 and a MAPE of 30.78%. The other three precious metal prices forecast metrics fall in between Gold and Palladium metrics. These results underscored the need for more sophisticated models to capture the complexities of the market.

**4.2 VAR Model Performance:**

The application of the VAR model revealed its limitations in accurately forecasting commodity prices. While it provided a broader view of the interrelations within the data, the error metrics, particularly MAE and RMSE, were higher than desired across commodities. The result suggested that while VAR could capture interactions between multiple variables, it needed to be more precise for accurate price forecasting in this context.

**4.3 Refined Forecasts with Auto ARIMA and Advanced Models:**

A significant improvement in forecasting accuracy was observed with the implementation of Auto ARIMA and other advanced models. For example, the Exponential Smoothing model for gold showed a lower MAE of 51.67 and a reduced MAPE of 2.71%, indicating a more accurate forecast compared to the Naïve model. Similar patterns of improved accuracy were observed for other commodities like Silver, Copper, Platinum, and Palladium, with each model showing its strengths in different aspects of the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| Commodity | Best Model | Parameters | MAE | MSE | RMSE | MAPE |
| Gold | Exponential Smoothing | Trend: add, Seasonal: add, Seasonal Periods: 12 | 51.67 | 3,475.74 | 58.96 | 2.72% |
| Silver | ARIMA | Order: (0, 1, 1), Seasonal Order: (0, 0, 0, 12) | 1.35 | 2.53 | 1.59 | 6.13% |
| Copper | Exponential Smoothing | Trend: add, Seasonal: add, Seasonal Periods: 12 | 0.19 | 0.04 | 0.20 | 4.66% |
| Platinum | ARIMA | Order: (0, 1, 0), Seasonal Order: (0, 0, 0, 12) | 52.85 | 3,458.38 | 58.81 | 5.45% |
| Palladium | ARIMA | Order: (0, 1, 1), Seasonal Order: (0, 0, 0, 12) | 232.30 | 70,193.72 | 264.94 | 15.75% |

Table 2: Best Performing Models

As Table 2 above shows, some of the best-performing models outperform the Naïve model (Gold, Platinum, and palladium), and others underperform compared to the Naïve model. Furthermore, there is no universal best-performing model for precious metals.

**4.4 Comprehensive Model Evaluation:**

The evaluation metrics like MAE, MSE, RMSE, and MAPE for each model provided a comprehensive view of their performance. The results indicated that while no single model was universally superior, each had its advantages depending on the specific characteristics of the commodity being forecasted. The Exponential Smoothing model often emerged as a strong performer, particularly for commodities with more stable market trends.

1. **Conclusion**

The journey through the world of time series forecasting the commodity price has been enlightening and offers valuable insights into time series models for various precious metals. The application of various models, including Naïve, VAR, ARIMA, SARIMA, and Exponential Smoothing, has highlighted their varied performances in different products. In particular, the Exponential Smoothing and ARIMA models have shown promising results, while the VAR model has encountered challenges in capturing market volatility. These observations underscore the importance of improving the dataset and approaching each precious metal as a unique forecast domain.

The insights gleaned from this analysis pave the way for several potential improvements and directions in future explorations of commodity price forecasting. One key avenue is the integration of machine learning, which promises to develop more sophisticated models and a deeper understanding of the precious metals market. Enriching the data set with various features can offer a more holistic view of market dynamics, improving the predictive capabilities of the models. Furthermore, venturing into modern modeling techniques could open up new possibilities for more accurate forecasts.

In conclusion, this foray into "Precious Metal Stock Price Time Forecasting" has been a remarkable adventure, full of learnings and discoveries. As we close this chapter, we plan on another journey, equipped with new knowledge and strategies, ready to delve deeper into the fascinating world of time series analysis and financial forecasting in the future.

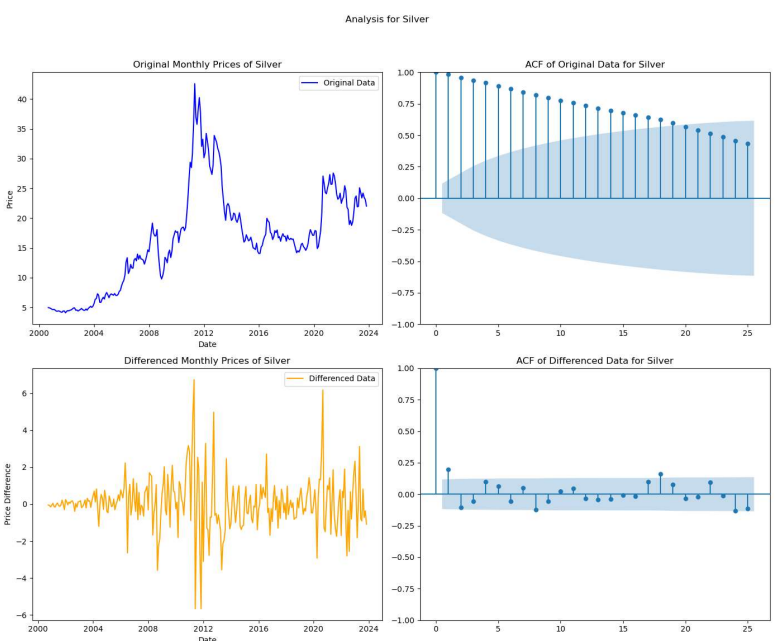
Appendix

1. Analysis of Gold Price

A close-up of a graph

Description automatically generated

1. Analysis of Silver Price



1. Analysis of Copper Price

A close-up of several graphs

Description automatically generated

1. Analysis of Platinum Price

A collage of graphs and diagrams

Description automatically generated

1. Analysis of Palladium Price

A group of graphs showing different types of data

Description automatically generated

1. Gold Price Forecast (Exponential Smoothing)

A graph showing a line

Description automatically generated

1. Silver Price Forecast (ARIMA)

A graph showing the growth of silver

Description automatically generated

1. Copper Price Forecast (Exponential Smoothing)

A graph showing the growth of copper

Description automatically generated

1. Platinum Price Forecast (Exponential Smoothing)

A graph showing a line

Description automatically generated

1. Palladium Price Forecast (Exponential Smoothing)

A graph showing a line

Description automatically generated

Work Cited

Farr, Cameran. "Stock Forecasting with ARIMA Models: A Comprehensive Guide." April 18, 2023. https://www.cameyeam.com/post/stock-forecasting-with-arima-models-a-comprehensive-guide. Accessed 20 Nov. 2023.

Harper, A. "A Hybrid Modelling Framework for Real-time Decision-support for Urgent and Emergency Healthcare." 2021, https://core.ac.uk/download/372986195.pdf.

Harris, Charles R., et al. "NumPy: Array Programming with Python." Nature, vol. 585, no. 7825, 2020, pp. 357–362.

Hunter, John D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, vol. 9, no. 3, 2007, pp. 90–95.

Hyndman, Rob J., and George Athanasopoulos. "Forecasting: Principles and Practice." OTexts, 2018.

McKinney, Wes. "pandas: a Foundational Python Library for Data Analysis and Statistics." Python for High-Performance and Scientific Computing, 2011.

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, 2011, pp. 2825-2830.

Seabold, Skipper, and Josef Perktold. "statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference, 2010, pp. 61-67.

Servera, Guillem. "Gold, Silver & Precious Metals Futures Daily Data." Kaggle, [www.kaggle.com/datasets/guillemservera/precious-metals-data](http://www.kaggle.com/datasets/guillemservera/precious-metals-data). Accessed 20 Nov. 2023.

Smith, Anthony, et al. "pmdarima: ARIMA Estimators for Python." pmdarima 1.8.2 documentation, 2020.

Waskom, Michael. "seaborn: Statistical Data Visualization." seaborn 0.11.2 documentation, 2021.