Title

Gold Price Prediction Using Machine Learning

Project Report

Submitted by

Htet Myet Zaw
mmdt2024.064@gmail.com

Table of Contents

Abstract

This project focuses on building models to predict gold price using machine learning techniques based on various economic indicators such as CPI, oil prices, USD index, and stock market indices. Using Ridge Regression and ARIMA methods, we set about investigating and understanding the basic nature of price changes of gold. The results show promising predictive power and highlight the potential of combining economic theory with data-driven approaches for financial forecasting.

Project Description

The aim of the project is to predict gold prices using a combination of economic indicators. Gold is a worldwide important asset. It represents safety in uncertain times, a symbol of stability when economies tremble. Thus, I was drawn to the challenge of predicting gold prices. It was not just a technical exercise but a method to explore how world gold prices depend on economic indicators.

I chose this problem because of the volatility of gold prices and their complex dependence on economic dynamics. With machine learning and time series analysis to forecast his price, I wanted to build a bridge between numbers and the world we live in- to take the chaos of global economics and find patterns which make sense.

Data Collection

I gathered data from some of the most trusted public sources.

- Consumer Price Index (CPI) from FRED (Federal Reserve Economic Data)
- Brent Crude Oil Prices from FRED
- USD Index from FRED
- S&P 500 Index from Yahoo Finance
- Gold Prices from World Gold Council, offering historical prices across multiple currencies

These indicators were chosen because they are often associated with changes in the gold price. For example, CPI shows inflation levels, oil prices can affect economic conditions, the USD index reflects the strength of the U.S. dollar, and the S&P 500 represents the overall performance of the stock market. All the data were resampled weekly and combined into one single dataset. Furthermore, lag values, moving averages, and volatility were also generated as extra features for the model so it can better associate itself with trends.

Modeling Approach

Two core modeling techniques were used:

1. Ridge Regression (from scikit-learn):

    I chose this model due to its ability to handle multicollinearity between economic variables. Before training, numerical variables were standardized in scaling. Some of the feature engineering methods considered were 4-week moving average, 12-week moving average, lagged values from lag1 and lag2, and measures of volatility.

2. ARIMA Models for Time Series Forecasting (via statsmodels):

    ARIMA models were fit separately for each feature and for gold price with AIC-based parameter selection. The stationarity of the data was tested using an Augmented Dickey-Fuller test. This enabled multi-step forecasting up to 12 months ahead, supporting both short-term and long-term planning.

Tools and Libraries

1. Pandas, NumPy: Used for data manipulation and numerical computations.
2. Scikit Learn:
    a. Ridge Regression for modeling.
    b. RobustScaler for feature scaling.
    c. TimeSeriesSplit for time series cross-validation.
    d. RMSE, MAE, and R^2 for calculating metrics
3. Statsmodels: Used for ARIMA modeling and diagnosing time series, e.g., indicating the Augmented Dickey-Fuller test and autocorrelation plots.
4. Matplotlib, Seaborn: data visualization and plotting results.
5. Streamlit: An interactive web app to visualize and interact with predictions in real time.

Workflow Process

To complete this project, I followed a step-by-step process to make sure everything was organized and logical. It started with understanding the problem and breaking it down into manageable stages.

☐ **Data Collection**

I searched for reliable sources like FRED, EIA, Yahoo Finance, and the World Gold Council to gather historical data. I made sure to include different economic indicators such as CPI, crude oil prices, the USD index, and the S&P 500 index. These were chosen because they often affect the price of gold.

☐ **Data Preprocessing**

This part was quite challenging, because the data had different time formats (some daily, some monthly). All datasets were resampled on a weekly basis. Missing values were forward-filled to maintain time ordering and prevent discontinuity. All numerical features were scaled through standardization. The lag values and moving averages were created to improve the model's time awareness and care was taken not to cause leakage among the features.

☐ **Feature Engineering**

I added lag features, moving averages, and volatility to give the models more context about recent trends. This helped the models understand how past values could affect the current week's price.

☐ **Modeling Phase**

I used Ridge Regression to handle the economic variables, especially because some of them were correlated. I also used ARIMA for pure time-series forecasting, which was useful for understanding long-term trends. I tested the stationarity of the data using the Augmented Dickey-Fuller test and carefully selected model parameters based on AIC values.

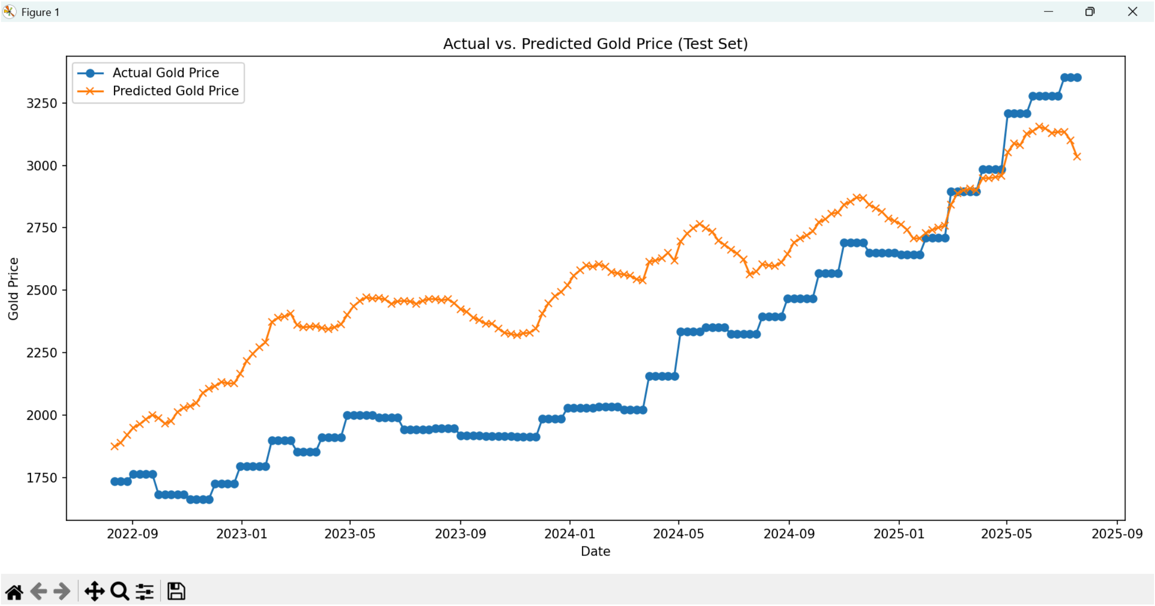
Finally, I built a Streamlit web app to visualize the predictions. It was exciting to see everything come to life in an interactive way. Users could see how the gold price changed based on different economic indicators, and I could even test it with new data.

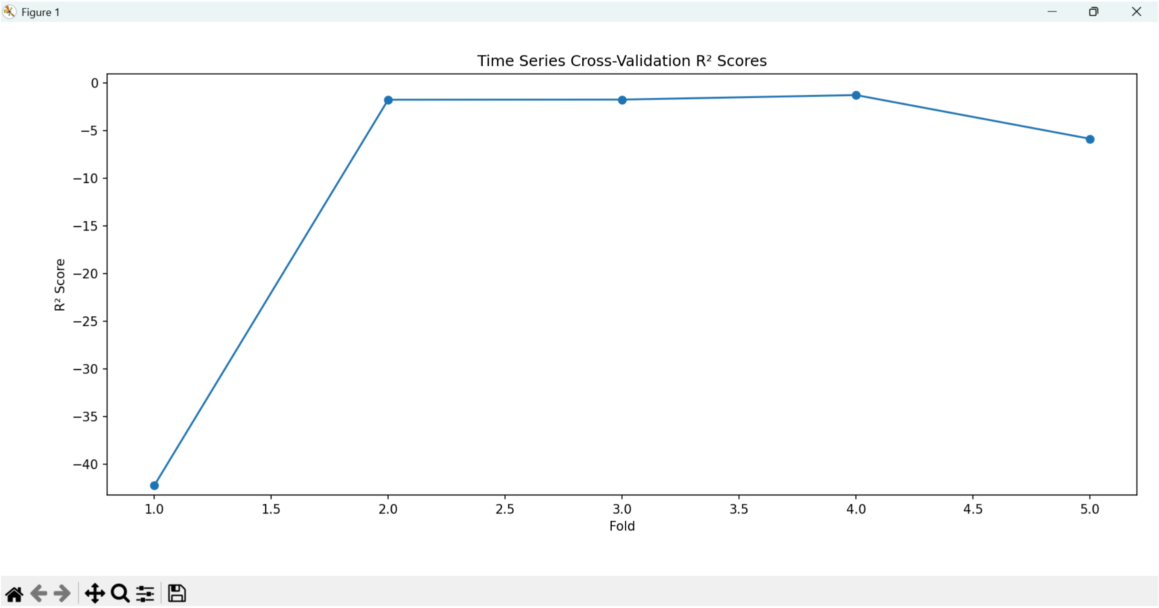
Evaluation

Here is the performance of the Ridge Regression model.

- **R² Score:** 77.05%
- **Mean Absolute Percentage Error (MAPE):** 5.57%
- **Root Mean Squared Error (RMSE):** ~$85

These results indicate that the model generalizes well enough on unseen data with fairly low prediction errors. A cross-validation was used with a 5-fold time series split to check if the model behaves robustly under any other time period.





Reflection

This project helped me understand how economic indicators can be used in machine learning to make meaningful predictions. One of the major difficulties was in working with different data granularities: CPI being monthly while others were daily and it was a big challenge to align them properly. Making sure everything matched weekly and didn’t introduce data leakage took a lot of careful processing.

From this project, I deepened my understanding of feature engineering, how to use lag features in time series analysis and the interpretability of regression coefficients in the context of economic data. Furthermore, I had the opportunity to gather experience in developing a Streamlit app, which allowed me to visualize and interact with predictions in real-time.

In the future, I want to extend this project by applying more advanced models like the XGBoost or LSTM to grasp more profound trends. I also plan to connect the system to real-time data sources using APIs, so the app can predict current gold and dollar price changes more accurately. I want to try putting this into practice for my own country's currency and economy, where price changes can be very impactful.

Overall, this was a valuable learning experience that combined technical skills with real-world data, and it gave me more confidence to work on financial prediction problems in the future.