Forecasting Gold Prices Using Economic Indicators and Historical Data: A Hybrid Approach Combining Random Forest, PCA, and Linear Regression

Hayden Virtue, Helen Hao, Iris Liu, Victor Zhao, Weicheng Wang

Duke University - Fuqua School of Business

DECISION 520Q: Data Science for Business

Final Team Project

1. INTRODUCTION

1.1 Executive Summary

Our project explores the potential of predicting gold prices using a data-driven approach that leverages historical price data and various economic indicators. Accurate gold price forecasts are crucial for investors, institutions, and policymakers, making it vital for us to identify key predictors and evaluate model performance for strategic decision-making. We used Principal Component Analysis (PCA) and Random Forest for feature selection, followed by a linear regression model, focusing on recent price movements, stock market indices, and exchange rates. Using data from March 1st, 2022, to August 7th, 2024, we trained and tested our model, which resulted in an out-of-sample R^2 of 0.954, MSE of 439.19, and RMSE of 20.96 USD/OZ, reflecting strong predictive accuracy. Although our model includes complex terms like "log(A*B)" that limit straightforward literal interpretation, it offers a reliable tool for forecasting short-term gold price trends and supporting investment decisions. Future enhancements could include integrating real-time data sources and advanced time series analysis to improve accuracy during periods of market volatility.

1.2 Business Problem and Understanding

Gold is often considered a secure investment during economic uncertainty. Still, its effectiveness as a hedge is complex and depends on various economic factors like inflation, interest rates, and market indices. The challenge lies in understanding when and how gold functions as a safe-haven asset. This uncertainty makes decision-making more difficult for investors, government institutions, and policymakers.

This project addresses this challenge by developing a predictive model that explores the relationships between gold prices and key economic indicators, such as the yields on 2-year US Treasury Bonds, S&P 500 Index, and crude oil prices. Using historical data and a supervised learning approach, the model offers insights into gold's potential as an investment under different economic conditions. By offering a data-driven understanding of gold's behavior, this project helps to guide strategic, informed decisions in gold investment.

2. LITERATURE REVIEW

Several studies have successfully constructed models to predict asset prices using historical dynamics of gold prices and some economic indicators. For instance, Zheng et al. (2024) used historical stock price data to predict future stock trends for Apple, Samsung, and GE. Similarly, gold as a crucial asset, might also be predictable using its historical prices. In addition, Shafiee and Topal (2010) leveraged crude oil prices and inflation to predict gold prices based on their relations, indicating that some economic indicators might also be predictable for gold prices. Hence, this study aims to explore gold price prediction with both historical gold price data and economic indicators.

3. DATA DESCRIPTION

The selected dataset contains historical daily gold prices in the past two years, from March 1st, 2022 to August 7th, 2024. The unit of gold price in this study is USD/oz. We identify ***Price.Today***, the gold price on August 7th, 2024, is also the target variable for the following models. This dataset also contains gold price dynamics, including gold prices two days and one day before today, price changes in the past 10 days, 20 days, 50 days, and 200 days, as well as the standard deviation of gold prices in the last 10 days. Historical data and economic indicators could serve as features in the following models. Finally, this dataset includes the gold price tomorrow, which is

helpful to validate our prediction when we predict the gold price tomorrow using the model in this study.

To predict gold price and validate it with real price tomorrow in our dataset, we split the dataset into two segments by date, with the ***first_half*** as training data and ***second_half*** as testing data for evaluating out-of-sample metrics. In addition, in later models, we adjust the dataset by omitting NA values if necessary.

4. DATA ANALYSIS APPROACHES

4.1 Exploratory Data Analysis (EDA) - Correlation Matrix Heatmap

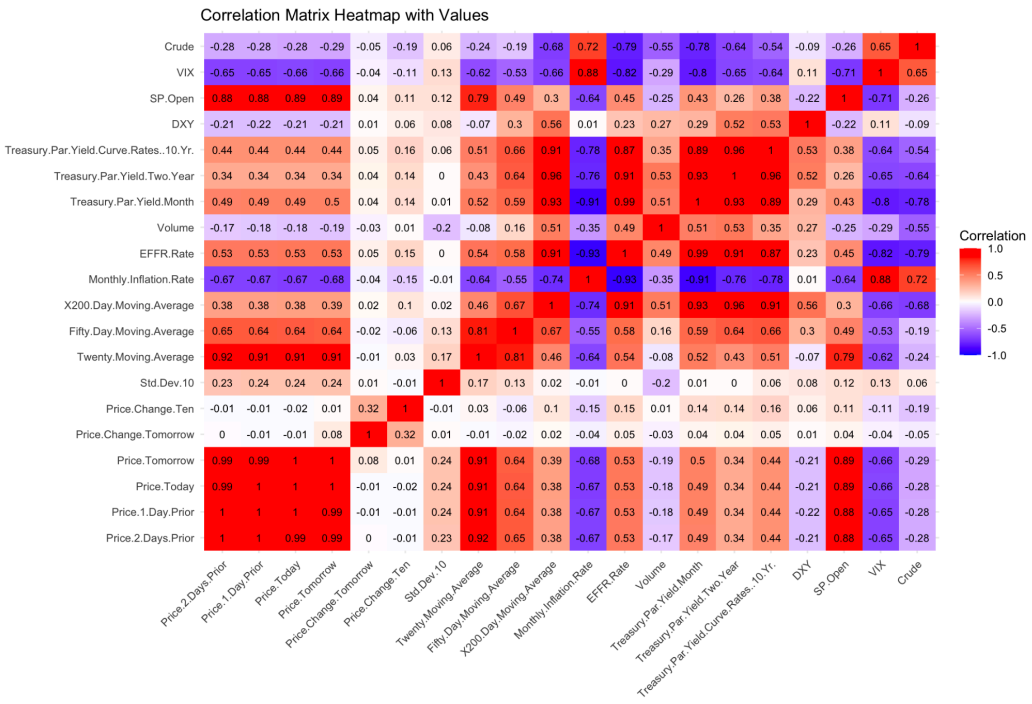

Fig. 1: Correlation Matrix Heatmap

We generated a correlation heatmap (see Fig. 1) to understand the relationships between ***Price.Today*** and various economic indicators, and to identify key predictors for our modeling

process. The heatmap reveals that ***Price.Today*** has very strong positive correlations with ***Price.1.Day.Prior*** and ***Price.2.Days.Prior*** (both around 1.00), highlighting high autocorrelation in gold prices. We also found a moderately strong positive correlation with ***SP.Open*** (0.89) and ***Twenty.Moving.Average*** (0.91), suggesting that market trends and moving averages play a significant role. Conversely, ***Price.Today*** has moderate negative correlations with ***Monthly.Inflation.Rate*** (-0.67) and ***VIX*** (-0.65), indicating that higher inflation or market volatility tends to align with lower gold prices. This analysis provides us with a preliminary choice of predictors for the gold price model.

4.2 Principal Component Analysis (PCA) for Independent Variables Selection

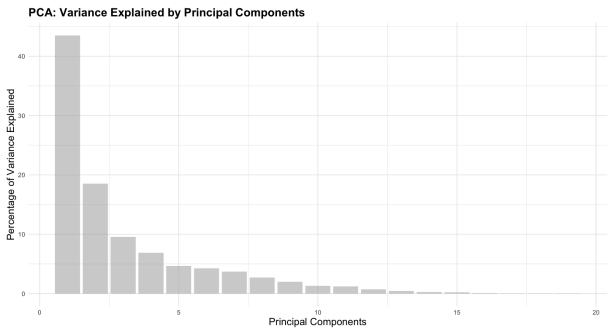

Fig. 2: PCA Scree Plot

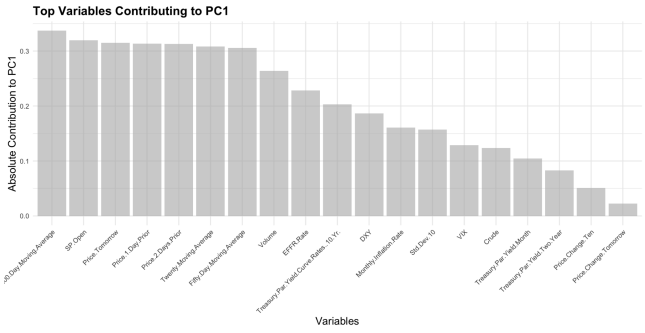

Fig 3: Top Variables Contribution to PC1

We used PCA to identify the most influential predictors for ***Price.Today***, it focuses on variables that capture the most variance for a more efficient model. The scree plot (Fig. 2) shows that PC1 explains over 40% of the variance, which emphasizes the importance of a few key variables.

Key Variables from PCA Analysis (Fig. 3): The bar plot highlights ***Price.1.Day.Prior***, ***SP.Open***, and ***Twenty.Moving.Average*** as top contributors to PC1 guides our independent variable selection:

1) ***Price.1.Day.Prior***: Its robust contribution to PC1 shows its importance in capturing daily price trends.

2) ***SP.Open***: It reflects stock market influences on gold prices, which supports its inclusion.

3) ***Twenty.Moving.Average***: This captures longer-term trends, ensuring our model considers broader market dynamics.

We used PCA to 1) Identify Key Variables: We focused on variables with the highest loadings and predictive power and 2) Screen Variables: PCA guided us in prioritizing variables before further testing.

5. MODELING APPROACH AND EVALUATION

5.1 Random Forest Model

We first choose the random forest model to distinguish feature importance in addition to the key variables identified in the PCA analysis. Intuitively, a bar plot is constructed to rank variables in the sequence of importance where ***IncNodePurity*** (Increased Node Purity) is extracted from each feature, a metric that measures the change inhomogeneity of groups created by trees in a Random Forest model. A higher value indicates more importance.

Model Formula: Our (feature selecting) random forest model is:

Price.Today ~ Price.2.Days.Prior + Price.1.Day.Prior +Price.Change.Ten + Std.Dev.10 + Twenty.Moving.Average + Fifty.Day.Moving.Average + Monthly.Inflation.Rate + Crude + X200.Day.Moving.Average + EFFR.Rate + Volume + Treasury.Par.Yield.Month + Treasury.Par.Yield.Two.Year + `Treasury.Par.Yield.Curve.Rates..10.Yr.` + DXY + SP.Open + VIX

Feature Importance in Random Forest Model

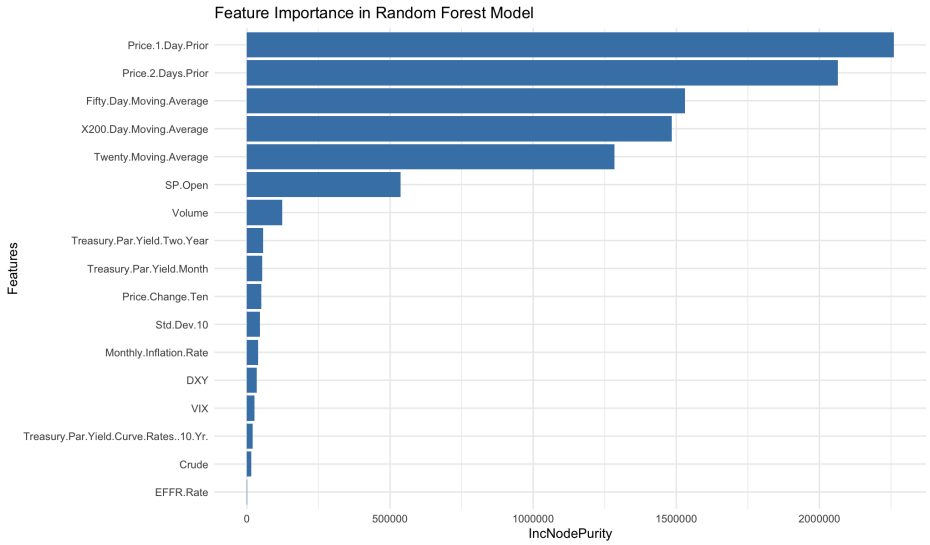

Fig. 4: Feature Importance Bar Plot

We can conclude that indicators ***Price.1.Day.Prior, Price.2.Days.Prior, Fifty.Day.Moving.Average, X200.Day.Moving.Average, Twenty.Moving.Average,*** and ***SP.Open*** are more inhomogenous and more influential in improving the model's accuracy. The result from feature importance also recalls our PCA analysis where indicators ***Price.2.Days.Prior, Twenty.Moving.Average,*** and ***SP.Open*** are proven to be the key indicators, suggesting that the above influential key variables can be selected for our prediction model.

5.2 Linear Regression Model

We used *linear regression* as our final best model because it demonstrated a high in-sample (0.990) and out-of-sample R^2 (0.954), indicating a strong predictive accuracy on both the training and testing data. This suggests that our model captures the key relationships without sacrificing generalization to new data. To avoid overfitting, we carefully selected a limited number of X

variables based on PCA, random forest analysis, and statistical significance, making sure that we included only the most impactful predictors rather than overloading the model with too many features.

We used a linear regression model with ***log(Price.Today)*** as the dependent variable to stabilize variance and normalize skewed financial data.

In terms of our final selection of predictors (i.e. independent variables) was guided by the top contributors from PCA, feature importance from random forest analysis, and iterative trials focusing on statistical significance (p-values) and the overall standard error of regression. Variables like ***Price.1.Day.Prior***, ***SP.Open***, ***Twenty.Moving.Average***, and ***DXY*** were consistently identified as significant contributors. For instance, ***Price.1.Day.Prior*** and ***SP.Open*** showed high importance in both PCA and random forest results, which indicates their strong influence on gold price trends. ***Twenty.Moving.Average*** was also selected for its ability to capture longer-term price trends, while ***DXY*** was included due to its role in representing the US dollar's influence on gold.

Model Formula: Our best-performing model is:

*log(Price.Today) ~ log(Price.1.Day.Prior * DXY) + Twenty.Moving.Average * Crude + SP.Open + Treasury.Par.Yield.Two.Year + DXY*

This model includes interaction terms (***log(Price.1.Day.Prior * DXY)*** and ***Twenty.Moving.Average * Crude***) to capture combined effects that could influence gold prices. Specifically, the interaction between ***Price.1.Day.Prior*** and ***DXY*** help to model how past gold prices respond to fluctuations in the US dollar. Similarly, ***Twenty.Moving.Average*** and ***Crude*** together account for how energy prices influence longer-term gold price movements.

5.3 Model Evaluation

To evaluate the model, we calculated out-of-sample (OOS) metrics using the second half of our dataset to test predictive accuracy. We focused on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), as well as R-squared (R^2):

OOS MSE: The mean squared difference between actual and our predicted prices is 439.19.

OOS RMSE: Our RMSE was 20.96 USD/OZ (same as our dependent variable unit), which indicates the typical deviation between actual and predicted prices.

OOS R-squared: Represents the proportion of variance in ***Price.Today*** explained by our model on unseen data, with an R^2 of 0.954, which reflects strong predictive power.

5.4 Model Results Analysis

Our linear regression model with ***log(Price.Today)*** as the target demonstrated strong predictive power, achieving an out-of-sample R^2 of 0.954, MSE of 439.19, and RMSE of 20.96. These metrics indicate a high level of accuracy in predicting unseen data.

The most influential predictors in our linear regression model include ***Price.1.Day.Prior***, ***SP.Open***, ***Twenty.Moving.Average***, and ***DXY***, as identified through PCA and feature importance from the random forest. ***Price.1.Day.Prior*** was particularly critical, reflecting the persistence of price trends, while ***SP.Open*** and ***DXY*** captured broader market influences and currency dynamics.

The model's robust performance as analyzed above suggests that gold prices could be effectively predicted using recent price movements, market indicators, and economic variables.

6. DECISION AND RECOMMENDATIONS

The linear regression model, using indicators identified through PCA and random forest's feature importance, is statistically effective in forecasting gold prices. We recommend using this model for short-term price forecasting.

We also need to acknowledge that our predictive model depends largely on historical price and economic data, which may not adequately reflect sudden shifts or unforeseen events that influence gold prices, such as sudden geopolitical conflict. As a result, the model's accuracy could decrease in times of significant market volatility or economic uncertainty. Another limitation is that even though our model demonstrates strong predictive accuracy, with a Mean Squared Error (MSE) of 439.19, a Root Mean Squared Error (RMSE) of 20.96 USD/OZ, and an out-of-sample R^2 of 0.954, it includes terms like "log(A*B)" that are challenging to interpret in straightforward, literal terms. This complexity can limit the ability to derive clear insights about the relationship between predictors and the dependent variable, despite the high accuracy.

Future work should focus on enhancing the prediction model by incorporating additional data sources, such as real-time market indicators and news sentiment analysis, to better capture sudden shifts and unforeseen events. Additionally, utilizing longer historical datasets could also improve the model's ability to identify long-term patterns in gold prices. Implementing advanced time series analysis methods, such as ARIMA or GARCH (Setyowibowo et al., 2021), may further refine predictions during periods of market volatility.

7. CONCLUSION

This project aimed to develop a predictive model for gold prices using a comprehensive data-driven approach that integrates historical price data and key economic indicators. Our analysis revealed that recent price movements, along with influential variables such as the yields

on 2-year US Treasury Bonds, the S&P 500 Index, and crude oil prices, significantly impact gold price forecasting.

The final linear regression model demonstrated strong performance, achieving a Mean Squared Error (MSE) of 439.19, a Root Mean Squared Error (RMSE) of 20.96 USD/OZ, and an out-of-sample R^2 of 0.954. These results suggest that our model can effectively forecast short-term gold price trends, providing a reliable tool for investors, institutions, and policymakers in their decision-making processes.

The implications of this study extend to various stakeholders. Investors may use the model to help optimize their gold investments, government institutions can utilize the model to manage national gold reserves, and policymakers can develop better economic plans. However, it is crucial to understand that the model's accuracy may be limited by abrupt market shifts due to its reliance on historical data.

Moving forward, to address any potential gaps in the model's forecasting power, future research could improve the model's accuracy by incorporating real-time market data and applying advanced analytical methods, such as ARIMA or GARCH models. Overall, this project provides valuable information on the dynamics of the gold price and provides a data-driven tool to stakeholders for informed decision-making about gold trading and investment.

REFERENCES

Gold Price & Relevant Metrics. Kaggle. https://www.kaggle.com/datasets/cvergnolle/gold-price-and-relevant-metrics

Setyowibowo, S., As'ad, M., Sujito, S., & Farida, E. (2022). Forecasting of Daily Gold Price using ARIMA-GARCH Hybrid Model. *Jurnal Ekonomi Pembangunan*, *19*(2), 257–270. doi.org/10.29259/jep.v19i2.13903

Shafiee, S., & Topal, E. (2010). An overview of global gold market and gold price forecasting. *Resources Policy*, *35*(3), 178–189. doi.org/10.1016/j.resourpol.2010.05.004

Zheng, J., Xin, D., Cheng, Q., Tian, M., & Yang, L. (2024). The Random Forest Model for analyzing and Forecasting the US Stock Market under the background of smart finance. In *Atlantis Highlights in Computer Sciences/Atlantis highlights in computer sciences* (pp. 82–90). doi.org/10.2991/978-94-6463-419-8_11