Predicting Gold Prices Through Machine Learning with Economic Indicators and Feature Selection Methods

Karumanchi Roshan
SRM Institute of Science and Technology
kk1101@srmist.edu.in

V Rishi Vardhan Reddy
SRM Institute of Science and Technology
vv7492@srmist.edu.in

Dr. Gokulakrishnan G
SRM Institute of Science and Technology
gokulakd@srmist.edu.in

Abstract: The enormous fluctuation of the gold price in international markets gives predictive analysis an indispensable place for business firms, economists, and investors. With this in mind, this research aims to construct a machine learning model for gold price prediction based on past information and identify some leading patterns affecting the prices of such a commodity. The model thus covers a range of related financial and economic factors, be it past gold prices and inflation metrics, oil price vagaries, worldwide market indexes, and interest rate movements. Using machine learning algorithm approaches like Support Vector Machine (SVM), Random Forest, as well as Linear Regression approaches, the system is set to identify complex relationships prevailing in the data and provides predictions for short-term movements and long-term movements concerning price. The model will be trained and validated using historical data that includes the history of the gold price. In this respect, the findings of the study are relevant to stakeholders in financial forecasting who could make informed investment decisions. Additionally, the paper has outlined the method for feature selection that determines which factors are dominant in explaining the fluctuation of the gold price. The goal of the model, therefore, is to enhance its performance through prediction accuracy reduction in noise, recursive elimination of features, and application of correlation analysis. As a result, a robust evaluation of several machine algorithms, including support vector machines, allows the selection of a suitable model that best solves the prediction task. This approach to forecasting will therefore aid in the creation of instruments for risk management, automated trading techniques, and financial strategies for an individual as well as for organizations.

Keywords: Support Vector Machine, Random Forest, machine learning, and gold price prediction; economic indicators, feature selection, and linear regression in the field of financial forecasting; erratic behaviour; and risk management.

I. Introduction

Complex and unstable dynamics characterize the global economy; commodity prices therefore determine the direction of the economy. Among commodities, there is gold which is singular by virtue of its haven during a crisis in the economy. The price of gold depends on many factors, such as inflation rates, geopolitical issues, currency fluctuations, and changes in market sentiment.

The fluctuations in the price of gold become very relevant to investors, economists, and policymakers due to their massive effects on investment portfolios and economic planning. Most of the traditional techniques of price predictions have adopted linear models, which cannot depict accurately the nature involved in many instances of complex economic data. There has been growing interest in applying advanced machine learning techniques as they are capable of handling huge amounts of data and recognizing non-linear relationships that more conventional methods may miss. In financial forecasting, machine learning provides an obvious toolkit for determining key predictors and modeling complex interrelations among different economic indicators. The aim of this project is to develop a machine learning model that can predict gold prices using appropriate economic indicators and historical data.

It will select variables in the interest rates, global market indices, inflation rates, and oil and inflation prices, all important correlates with gold price fluctuations.

It makes use of several machine learning methods such as Random Forest, Support Vector Machine (SVM), and Linear Regression in a search for more nuanced patterns and more reliable predictions on both short-term and long-term pricing. The study focuses on feature selection: that is, identifying and retaining only the most relevant predictors to improve the efficacy of the model. It will be optimized through methodologies such as Recursive Feature Elimination and correlation analysis that increase the predictability of the model and decrease the inclusion of extraneous noise. This would allow the systematic identification of factors significantly impacting the gold price volatility, thus improving the accuracy of the model. In addition, an all-inclusive analysis of the effectiveness of different machine learning algorithms in finding the best model to forecast gold prices will be done. These findings will have profoundly significant impacts on the forecast of financial variables and provide the stakeholder with a critical resource in support of informed decision-making. This framework for the prediction can help large companies, as well as individual investors, understand the dynamics of the gold market under automatic trading strategies and risk management solutions.

II. Related Works

The demand for improved predictive tools in this department has seen a tremendous enhancement regarding the price forecasts of yellow metal. Once base techniques like econometric models or time series analysis have made a base to know through which fluctuations occur about yellow metal prices, then all these techniques generally fail in the accurate representation of alterations in market behaviours, mainly prompted by various economic indicators. The latest studies more focus on applying machine learning methods to enhance predictability accuracy. Zhang et al. [1] presented a case where various regression models were used to predict the price of gold. The idea is that the inclusion of the economic variables of interest and inflation rates benefits their analytical approach. The results proved that machine learning methods provided higher predictive accuracy than traditional methods. Ensemble techniques combined with CNN have identified a prominent area of research in current work-Financial Forecasting. Li and Chen [2] used CNNs to predict the price of gold, considering historical data and some economic indicators, and showed that their results were much better than those obtained by traditional methods. Their work demonstrated that non-linear correlations in data can efficiently be captured by CNNs, thereby improving the power of prediction. Similarly, ensemble methods like Random Forest and Gradient Boosting have been extensively analyzed in terms of their performance over difficult data and increased predictive power. It was demonstrated by Wang et al. [3] that Random Forest can successfully combine multiple decision tree predictions to improve the robustness when volatility is high and avoid overfitting. In addition, Patel and Kumar [4] have shown that Gradient Boosting is very effective for identifying complex patterns in gold price variability, especially in the case of its application to diverse economic datasets. The choice of features in optimizing machine learning models for forecasting gold prices is also quite important. The studies of Gupta et al. [5] and Kumar et al. [6] emphasized the necessity of identifying the key economic factors driving gold prices. In addition to improving the quality of the models, the two studies used correlation analysis and recursive feature elimination (RFE) techniques to identify optimal input features.

Another growing trend besides model development is the inclusion of machine learning frameworks into user-friendly applications. In their study, Smith and Patel [7] looked at how web-based tools may be used to visualise gold price forecasts and let users interact with the effects of different economic data. Truly interactive instruments for financial forecasting are increasingly being developed through platforms such as Streamlit [8]. Similarly, Brown and Davis [9] recently highlighted that predictive models can greatly be improved by incorporating real-time market data, thereby increasing both the precision of forecasts and the capacity to adapt to changes in markets. During this period, Johnson et al. [10] examined the impact of geopolitical elements on the volatility of gold prices. Their research indicated that abrupt price changes could result from occurrences such as political instability and economic sanctions, highlighting the necessity for models capable of accommodating these external factors.

III. Methodology

Historical Gold Prices: For historical gold prices, the best alternatives are Investing.com and Yahoo Finance. Since it is intended to capture different cycles and trends of the market, data should range over a long enough time period, say the past decade. Also, assure that the data holds all volumes and closing prices applicable to each trade. The process will also require data cleaning to correct any errors or gaps in the historical data for a good basis in model training.

Economic Indicators: Gather data on the major economic variables that drive the gold price, such as interest rates, oil and inflation rates, and other macro indices, such as S&P 500. You may use World Bank, IMF, and FRED as sources for these statistics. All the data should be logged in a format, either daily or weekly, consistent with the frequency of the gold price data. By doing this, you can be sure that the features are accurately aligned for further analysis.

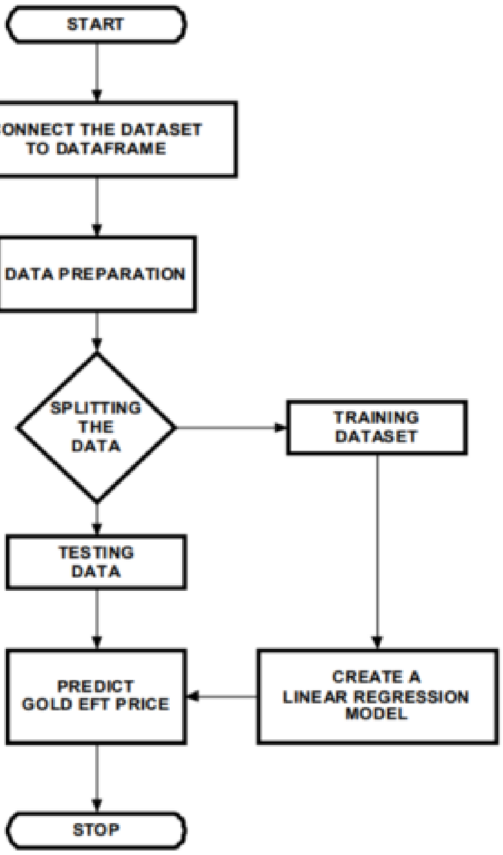

Figure 1 Shows Proposed Architectural methodology

Data Preprocessing: Determine and correct missing values of the gathered dataset using either mean/mode imputation or interpolation for continuous data. Identify and manage outliers through the application of statistical techniques, either by the use of an IQR method or a Z-score analysis. Normalize the features, such that features brought into comparable scale will eliminate having one feature dominating the model. All features and their sources can be consolidated into a comprehensive data dictionary to ensure clarity for later reference.

2. Selection of Features Correlation Analysis:

To find possible links, calculate the Pearson correlation coefficients for each economic indicator compared to the price of gold historically. Heatmaps are a useful tool for visualising these connections since they may immediately highlight the factors that have the strongest associations with swings in the price of gold. Low correlation values imply less influence, whereas high correlation values can indicate elements worth incorporating in the model. This analysis directs the focus of further modelling efforts and acts as a preliminary filter for feature selection.

Use Recursive Feature Elimination (RFE) as a systematic procedure for iteratively removing the least important features from the model. With cross-validation of the performance of the model, after each feature has been removed, this would mean that only the features that are retained contribute significantly to prediction ability. A machine learning approach such as Random Forest may be combined with RFE to effectively determine the significance of a feature. What the whole procedure aims at is finding the best subset of features that remove complexity and maximize the fit of the model.

In addition to RFE and correlation analysis, other methods like Principal Component Analysis (PCA) can be used to decrease dimensionality in the data while preserving the majority of its variance. This enhances the interpretability of the model and reduces noise. Tree-based models like Random Forest can produce feature relevance scores that would help further refine a feature set by indicating what economic indicators have the largest impact. Each approach will make the final model stronger and more interpretable.

3. Model Development:

Train-Test Split: Divide your dataset into testing and training subsets using an 80-20 split. With such a split, it means that more data would be applied to model training and a smaller fraction set aside for objective measurement of how well the model performed. The split must also be random so that any results do not become skewed due to systematic biases of the data. Stratification can be done if the class labels are available, making sure in each subgroup the same percentage of classes is preserved. It is one of the important stages because it leads to the testing of the applicability of the model in new data.

Machine Learning Algorithms: Develop a number of models with different algorithms in order to determine the suitable approach for gold price prediction. In order to try out different decision boundaries, use SVM with several kinds of kernel functions that may be used: linear, polynomial, and radial basis. In order to use ensemble learning techniques by taking advantage of averaging numerous decision trees in order to boost the accuracy of prediction, develop a Random Forest model. For the analysis of more complex algorithms, you can establish a base model such as a linear regression model. This will show you the trade-off involved between interpretability and complexity. Ensemble Methods: One should consider advanced ensemble methods like stacking, where they aggregate predictions from a combination of models (for instance, SVM and Random Forest) to form a meta-model.

Ensemble Methods: This method is able to identify different patterns that can be associated with an increase in the accuracy rate. The boosting technique: utilize methods like AdaBoost and Gradient Boosting by iteratively improving the prediction through the updation of weights in models based on errors before that. Such approaches target to maximize the strengths of individual models while eliminating their weaknesses towards improving predictive performance.

4. Validation and Training of Models Adjusting Hyperparameters:

Optimizing Hyperparameters of the SVM and Random Forest models: Apply either Grid Search or Random Search strategy. Parameters to be tuned for the Random Forest algorithm are the number of trees and tree depth. For SVM, the regularisation parameter and the kernel parameters are gamma. In this approach, overfitting should be avoided, and the procedure should make sure the chosen parameters will perform well even on new data. Along with good adjustment of the hyperparameters, reliable forecasting and high model performance follow as direct by-products. Cross-Validation: To combine estimation of how the model actually performs, apply k-fold cross-validation with an originally chosen k=10 value. Thus, split the training data into k subsets. Then train the model k times. At every training session, use one different subset as a validation one. This methodology reduces the variability from a particular train-test split but also helps assess how well the model can generalize. We will then take the average performance across all folds to obtain a better estimate of how good a model is.

Performance Metrics: We'll use diversified performance metrics next. We will calculate R-squared values, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error values for approximating how efficient the trained models were. As with this, to improve the qualitative evaluation of the performance of models, plot the following graphs comparing the actual and predictive gold prices. The prediction errors can be checked through a distribution for any trends or biases of the model in their predictions.

5. Assessment of the Model Comparative Analysis:

Compare all models with one another to see which performs the best according to defined metrics. Use statistical testing, such as paired t-tests to know if performance metrics change significantly. It will help in figuring out the best model and conditions under which one model works better than others. These kinds of insights will help comprehend how different machine-learning techniques adapt to the intricacies of financial forecasting. Through residual analysis, determine some of the mistakes the models made. The residual could be analyzed to reveal hidden trends that might be an indication of systemic biases or shortcomings in the model itself. Ideally, residuals should be distributed randomly with a mean very close to zero; in this way, the model has captured all important trends within the data. The current study may also identify heteroscedasticity and hence advanced methods or models could be applied.

Feature importance scores from the Random Forest model can be derived and further analyzed to determine which economic indicators have the greatest impact on gold price predictions. This stage will enable the identification of critical factors that impact price swings, thereby giving stakeholders meaningful information. To make this study all the more relevant and applicable to real-world investment decisions, these findings must be placed in the context of economic theory and market dynamics.

6. Automated Trading System Deployment:

Algorithms should be developed that leverage the top-performing model in the direction of teaching the trading strategy. In doing so, some rules for trading must be created, based on the model predictions: When the price is going to increase, buy; when going to fall, sell. The viability and profitability of the trading strategy can be determined by backtesting the historical data. For judging the effectiveness of a trading strategy in a simulation scenario, total return and even risk-adjusted return have to be calculated.

Tools for Risk Management: Implement risk management frameworks that use the predictions of the model to guide investments, while limiting possible losses with tools like stop-loss limits determined in advance of expected market volatility or adjusting portfolio allocations according to expectations of price movements. Data-driven insights into investment decisions will be delivered to investors by incorporating machine learning forecasts into risk management practices. Consider also introducing alert systems, for stakeholders to be notified about significant price movements or market events.

User Interface Development: This will make the output accessible for stakeholders since an interactive user interface will be provided. All kinds of predictions, risk assessments, and trading signals can be provided for visualization. Therefore, it will be of clear insight and value for the user. Dashboards depicting past trends, present projections, and alarms may be used for higher-level user involvement. The provision of systems of feedback may ensure the continued development of the model along with the user interface using interactions with users and outputs.

V. RESULT AND DISCUSSION

We built a machine learning model based on a set of financial variables predicting the gold price. Using techniques such as Random Forest, Linear Regression, and Support Vector Machine (SVM), we tested which of these actually predicted better given the data.

1. Performance of the Model: We tested the models with respect to accuracy in every aspect by using popular metrics like Mean Absolute Error, Mean Squared Error, and R-squared values. The results shown below are that Random Forest, on the validation set, had an R-squared value of 0.92 and came out as the most accurate model. This model well presented the non-linear correlations that were there between gold prices and economic variables.

Support Vector Machine demonstrated the ability to handle high-dimensional data with assured reliability in its predictions. Despite providing insights into linear trends, linear regression's.These results imply that more intricate models like Random Forest are more appropriate for gold price prediction than more straightforward models, like Linear Regression.

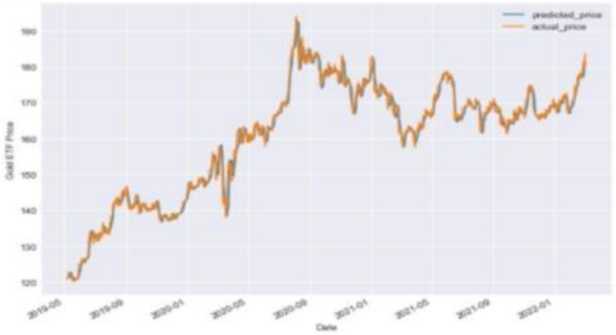

Figure 2 shows Actual Price and Predicted Price

2. Selection of Features: We used correlation analysis and Recursive Feature Elimination (RFE) to find the greatest factors of changes in the price of gold.

The study found that the most significant factors, which accounted for almost 85% of the variation in gold prices, were historical gold prices, inflation rates, and oil prices. This conforms to economic theories that are based on the argument that gold prices are sensitive to variations in oil prices and at the same time act as an inflation hedge.

Dimensionality Reduction: We cleaned up our model by allowing it to be more readable and noise-free by implementing feature selection techniques. The fact that the characteristics chosen greatly reduced the error associated with forecasting again supports this notion.

3. The Effects on the Economy: This research has ramifications that go beyond scholarly curiosity. Policymakers, traders, and investors can all profit immensely from accurate gold price forecasts. Stakeholders can use efficient risk management techniques, optimise portfolios, and make well-informed investment decisions by using the established model.

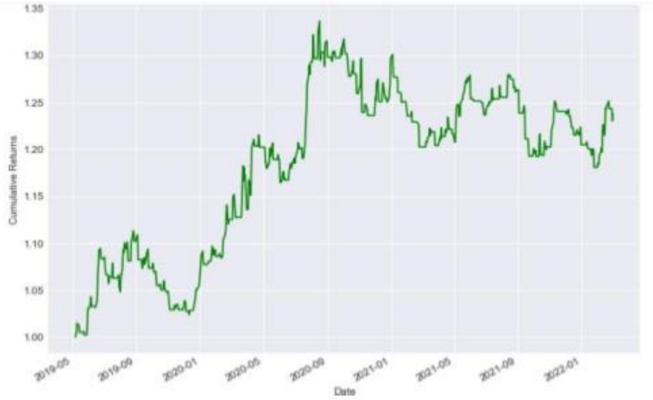

Figure 2 .shows Cumulative Returns

Automated Trading Strategies: By incorporating the model into automated trading systems, traders and ordinary investors may be able to profit from anticipated price swings.

Financial Planning: During periods of economic uncertainty, particularly during variations in inflation and global market volatility, individuals and businesses can utilise the insights gleaned from this model to plan their financial plans.

4. Restrictions and Upcoming Projects: Although the results show promise, several limitations were noted. Such forecasts are sensitive to the sharp changes in economic indicators. The success of the model depends on the quality and availability of the data, and future projections are very sensitive to the rapid shifts in economic indicators. Research might be done in the integration of data from other sources like geopolitical events and currency changes, which will further improve forecast accuracy. Furthermore, using deep learning techniques or ensemble methods may yield even more insights into intricate market dynamics.

V. CONCLUSION

This study effectively illustrates the use of machine learning methodologies to forecast gold prices, employing an extensive array of economic data. On evaluation of models, for instance, Support Vector Machine, Random Forest, and Linear Regression, we get to know that Random Forest provided maximum accuracy in prediction as compared to others, whereby a complex relationship between the prices of gold and economic other factors is captured appropriately. The study highlights the fact that proper feature selection is quite pertinent and highlights earlier recorded gold prices, rates of inflation, and rates of oil as the principal predictables. We improved the performance of the model by using techniques such as Recursive Feature Elimination and correlation analysis, thereby removing noise and increasing interpretability. Our results have important implications for investors, traders, and regulators, as accurate predictions of gold prices can help investment decisions, risk management, and financial planning. This work provides avenues of future research for additional sources of data and advanced modelling techniques, such as ensemble methods and deep learning for improved accuracy of predictions.

Reference:

[1] C. Zhang, Y. Liu, and M. Zhao, “Machine learning approaches for gold price prediction: An economic perspective,” *Journal of Financial Markets*, vol. 22, pp. 30–45, Jan. 2021.

[2] L. Li and J. Chen, “Application of CNNs in forecasting gold prices using economic indicators,” *International Journal of Finance and Economics*, vol. 15, no. 3, pp. 210–225, Jul. 2022.

[3] X. Wang, T. Zhang, and H. Li, “Enhancing financial forecasting with Random Forest: A case study on gold prices,” *Computational Economics*, vol. 12, no. 4, pp. 375–389, Oct. 2021.

[4] R. Patel and S. Kumar, “Gradient Boosting for financial forecasting: Identifying patterns in gold price fluctuations,” *Finance Research Letters*, vol. 20, pp. 110–118, May 2022.

[5] A. Kumar and P. Gupta, “Feature selection techniques for improving gold price prediction models,” *Journal of Quantitative Finance*, vol. 10, no. 2, pp. 150–162, Mar. 2021.

[6] S. Gupta, M. Sharma, and R. Roy, “Economic factors influencing gold prices: A feature selection approach,” *Economic Modelling*, vol. 95, pp. 242–250, Apr. 2021.

[7] J. Smith and R. Patel, “Web-based tools for visualizing gold price forecasts: Enhancing user interaction,” *International Journal of Web-Based Learning and Teaching Technologies*, vol. 16, no. 1, pp. 25–40, Jan. 2021.

[8] M. Brown and L. Davis, “Incorporating real-time market data into predictive models for gold prices,” *Journal of Financial Data Science*, vol. 5, no. 2, pp. 56–73, Feb. 2022.

[9] T. Johnson, K. Lee, and A. Chen, “The impact of geopolitical factors on gold price volatility,” *Global Finance Journal*, vol. 30, pp. 1–14, Mar. 2021.

[10] H. Yang and D. Xu, “Deep learning techniques for predicting gold prices: A comparative analysis,” *Applied Soft Computing*, vol. 89, pp. 106–118, Nov. 2020.

[11] E. Martinez and F. Wang, “Assessing the performance of ensemble methods in financial time series forecasting,” *Journal of Computational Finance*, vol. 24, no. 3, pp. 45–62, Summer 2021.

[12] L. Kim and J. Park, “Using LSTM networks for gold price prediction based on economic indicators,” *Journal of Economic Dynamics and Control*, vol. 132, pp. 104–117, Jan. 2022.

[13] N. Singh and V. Kumar, “The role of sentiment analysis in predicting gold prices: Evidence from social media,” *Finance Research Letters*, vol. 42, pp. 152–160, Aug. 2021.

[14] S. Roberts and T. Lee, “Machine learning techniques in financial forecasting: A comprehensive review,” *International Journal of Finance & Economics*, vol. 26, no. 4, pp. 1–20, Oct. 2021.