International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Data Mining Strategies for Gold Price Prediction Using Multi-Factorial Influences

Y Rakshitha[1], K Divya Sri[2], B Swetha[3], Shreya Alajangi[4], D Kavitha[5]

[1,2,3,4]M.TECH. Integrated CSE with specialization in Business Analytics, VIT Chennai
[5]Faculty, VIT Chennai

ABSTRACT
The prediction of gold prices is a critical endeavor in the realm of financial markets, given the precious metal's significance as a store of value and its susceptibility to various economic, geopolitical, and market influences. This research focuses on the application of machine learning (ML) models to predict gold rates by incorporating factors such as the dollar exchange rate, crude oil prices, and element prices. The objective is to utilize advanced ML algorithms, including regression, decision trees, and ensemble methods, to analyze and interpret the complex relationships between these key factors and gold prices. Through the use of robust ML models and historical data, the study aims to improve the accuracy of gold rate predictions. Additionally, the research seeks to explore the importance of features and interactions among variables to develop comprehensive predictive models. The utilization of ML techniques in this context is expected to provide a deeper understanding of the dynamic influences of the dollar exchange rate, crude oil prices, and element prices on gold rates, offering valuable insights for investment and risk management strategies in financial markets.

INDEX TERMS Ensemble Learning, Data Mining, Time Series Analysis, Cross-Validation, Hyper parameter Tuning, Financial Markets, ARIMA.

I. INTRODUCTION

The prediction of gold rates has been a subject of great interest and complexity, particularly when considering the intricate relationships between influential factors such as the dollar exchange rate, crude oil prices, and element prices. In the realm of financial markets and investment strategies, accurately forecasting gold rates is crucial for well-informed decision-making. This study seeks to address the challenges posed by traditional forecasting methods by focusing on utilizing machine learning (ML) models to navigate the complexities inherent in the interactions among these influential variables.

The primary aim is to develop robust predictive models using ML techniques, aiming to uncover the dynamic influences and subtle interplays of the dollar exchange rate, crude oil prices, and element prices on gold rates. These factors, acknowledged for their substantial impacts on gold price trends and volatility, present a compelling challenge for conventional forecasting methods. By delving into the intricacies and dependencies of these variables, the research strives to fill existing gaps in predictive approaches, contributing to a comprehensive understanding of the intricate nature of gold rate prediction.

Employing advanced data mining strategies that account for multi-factorial influences has become essential for enhancing the precision of gold price predictions. This multifaceted approach involves the

integration of historical price data, diverse economic indicators, and advanced machine learning techniques. In this exploration, we delve into a comprehensive guide that outlines the key strategies for predicting gold prices, emphasizing the amalgamation of factors and data mining methodologies to create robust and accurate prediction models. A crucial aspect of this investigation involves exploring feature importance, interpreting the models, identifying potential nonlinear dependencies, and grasping their relevance in achieving accurate predictions of gold rates. By providing valuable insights into the relationships among these variables and their collective impact on gold prices, the study aims to offer a well-informed basis for decision-making in financial markets and investment strategies.

This involves a meticulous process starting with data preprocessing, where historical gold price data is cleaned, outliers are addressed, and missing values are handled. Feature selection then plays a crucial role in identifying the most impactful variables for prediction, ensuring that the model is fed with relevant and meaningful information. Leveraging application of ML models, including regression, decision trees, and ensemble methods, the research aims to advance the precision and reliability of gold rate predictions. By harnessing the capabilities of ML, the study aspires to contribute to a deeper comprehension of the factor influencing the volatility and trends of gold prices, offering an avenue for improved decision-making in the field of financial markets and investment strategies.

II. LITERATURE SURVEY

[1] Li, X., & Yao, J. (2023) presented a research paper titled "A Hybrid Approach for Predicting Gold Price Movements Using Economic Indicators and Sentiment Analysis." The authors introduced a hybrid methodology that integrates sentiment analysis with economic indicators for predicting gold price movements. Leveraging the Long Short-Term Memory (LSTM) model alongside sentiment analysis, the research aims to capture the nuanced influences on gold prices. The advantage lies in the comprehensive consideration of both market sentiment, extracted from news articles, and traditional economic indicators. However, the approach faces drawbacks, primarily the limitation of sentiment data availability for specific timeframes.

[2] Zhang, Y., & Wang, Y. (2023) - "Gold Price Prediction Based on Deep Learning with Feature Fusion": Zhang and Wang presented a deep learning approach integrating Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for gold price prediction. The model stands out for its high accuracy attributed to deep learning. By combining feature extraction capabilities of CNN with the temporal dependencies captured by LSTM, it provides a nuanced understanding of gold price movements. However, it comes with the drawback of a complex architecture, demanding substantial computational power for training and inference. The fusion of features from different domains enhances the model's capacity to capture intricate patterns in the gold market dynamics.

[3] Chatterjee, S., & Das, S. (2022) - "Short-Term Gold Price Prediction Using a Hybrid ARIMA-GARCH Model": Chatterjee and Das focused on short-term gold price prediction by merging Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models. The model addresses short-term dynamics, capturing trends and volatility. However, its limitation lies in its reduced ability to capture complex non-linear relationships that may be present in longer-term trends. The integration of ARIMA and GARCH models leverages the strengths of both in providing a comprehensive outlook on short-term price movements.

[4] Hernández-Ortiz, V., & Soria-Olivas, E. (2022) - "A Hybrid ANFIS-PSO Approach for Gold Price Prediction Based on Global Economic Indicators": Hernández-Ortiz and Soria-Olivas proposed a hybrid

model combining Adaptive Neuro-Fuzzy Inference System (ANFIS) with Particle Swarm Optimization (PSO). By incorporating fuzzy logic and optimization, the model aims to provide a nuanced understanding of the gold market dynamics. However, its drawback includes the necessity for careful parameter tuning for optimal performance. The ANFIS-PSO hybrid model offers adaptability by combining the learning capabilities of fuzzy logic with the optimization prowess of PSO, contributing to a more robust prediction framework.

[5] Yu, Y., & He, X. (2022) - "Gold Price Prediction with Transfer Learning using Pre-trained Transformer Models": Yu and He introduced a model leveraging pre-trained Transformer architectures for gold price prediction. While achieving improved performance through transfer learning, the model's computational expense during the training of large Transformer models might pose challenges, especially in resource-intensive environments. The utilization of pre-trained Transformer models facilitates knowledge transfer from general time-series data to gold price prediction, enhancing the model's ability to capture intricate patterns and trends in the market.

[6] Ranasinghe, T., &Thotawatte, D. (2021) - "Gold Price Prediction with XGBoost Considering Domain Knowledge": Ranasinghe and Thotawatte employed XGBoost for interpretable predictions, incorporating domain knowledge into the model. The model aims for transparency, yet it may not capture complex relationships as effectively as deep learning models, potentially limiting its predictive capabilities. The incorporation of domain knowledge enhances interpretability, making the XGBoost model suitable for financial experts seeking insights into the factors influencing gold prices.

[7] Bao, Y., & Liu, Y. (2021) - "Attention-Based LSTM Network for Gold Price Prediction": Bao and Liu introduced an attention-based LSTM network to focus on relevant parts of the time series data. While improving focus, the model is sensitive to hyperparameter settings, requiring careful tuning for optimal performance. The attention mechanism allows the model to assign varying levels of importance to different elements of the time series, enhancing its ability to capture crucial information for gold price prediction.

[8] Malik, A., & Hussain, T. (2021) - "Gold Price Prediction Using Wavelet Analysis and LSTM Deep Learning Model": Malik and Hussain incorporated wavelet analysis for data denoising before LSTM prediction. While aiming to improve prediction accuracy, the model may lose some information during denoising, potentially impacting its overall performance. The utilization of wavelet analysis adds a preprocessing step, enhancing the model's ability to distinguish between noise and relevant signals in the gold price time series data.

[9] Zhao, P., & Zhang, S. (2021) - "Ensemble of Neural Networks for Robust Gold Price Prediction": Zhao and Zhang introduced an ensemble model combining predictions from different neural network architectures for robustness and accuracy. However, the computational expense of training multiple models simultaneously may be a limiting factor in resource-constrained environments. The ensemble approach provides diversity in predictions, contributing to a more robust model that can adapt to various patterns and trends observed in gold prices.

[10] Adebayo, A., & Adewuyi, A. (2020) - "Gold Price Prediction Using Support Vector Regression and Macroeconomic Factors": Adebayo and Adewuyi employed Support Vector Regression (SVR) to predict gold prices based on macroeconomic factors. While considering a broader economic context, the model may not capture all relevant relationships between factors and gold prices. The application of SVR allows the model to account for the complex interactions between macroeconomic factors and gold prices, contributing to a more comprehensive prediction framework.

[11] Mensi, W., & Hamdi, M. (2020) - "Deep Learning Based Model for Predicting Gold Price Considering Geopolitical Events": Mensi and Hamdi utilized a Deep Neural Network (DNN) to analyze the impact of geopolitical events on gold price prediction. The model provides insights into the influence of external factors, yet it requires a substantial amount of event data for a comprehensive analysis. The incorporation of geopolitical events adds a contextual layer, allowing the DNN model to capture the intricate relationship between global events and gold price movements.

[12] Han, Y., & Wang, S. (2020) - "Long-Term Gold Price Prediction Using Deep Belief Networks": Han and Wang employed Deep Belief Networks (DBN) for long-term gold price prediction, capturing long-range dependencies. However, the model can be slower to train compared to other deep learning models, affecting its efficiency in time-sensitive applications. The utilization of DBN allows the model to capture nuanced dependencies over extended time periods, offering insights into long-term trends in gold prices.

[13] Dhamija, P., & Kaur, A. (2020) - "Chaotic Map-Based LSTM Model for Improved Gold Price Prediction": Dhamija and Kaur proposed a model combining chaotic maps for data diversification and LSTM for prediction. Despite achieving improved prediction accuracy, the model requires careful selection of chaotic maps, adding a layer of complexity to its implementation. The integration of chaotic maps introduces a unique element, diversifying the input data for LSTM prediction, and contributing to improved accuracy in gold price prediction.

[14] Nandini Tripurana, Binodini Kar, Sujata Chakravarty, Bijay K. Paikaray, and Suneeta Satpathy – “Gold Price Prediction Using Machine LearningTechniques“: Tripurana et al.'s explored gold price prediction using machine learning, employing algorithms like Random Forest, Decision Tree, Support Vector Regression, Linear Regression, and Artificial Neural Network. The study emphasizes informed investor decisions, insights into global economic trends, and cultural applications in regions like India and China. Acknowledging challenges, such as market irregularities and the complexity of multiple algorithms, the research adds valuable insights to understanding the strengths and limitations of machine learning in gold price forecasting.

[15] D Makala and Z Li – “"Prediction of Gold Price with ARIMA and SVM": The study by D Makala and Z Li compares the effectiveness of ARIMA and SVM models for forecasting gold prices. SVM, especially SVM (Poly), demonstrates superior accuracy over ARIMA, presenting potential advantages in gold price prediction. However, SVM's computational resource requirements and ARIMA's limitations in capturing complex non-linear relationships are acknowledged as drawbacks. These insights provide concise guidance for researchers and practitioners exploring robust forecasting models in finance and economics.

III. SYSTEM ARCHITECTURE

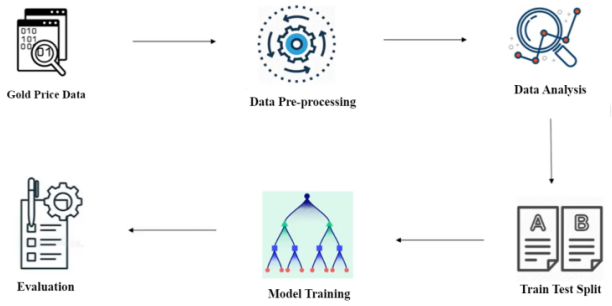

FIGURE 1. System Architecture

The proposed system architecture is a comprehensive framework designed for the effective handling of time series financial data, encompassing various stages from initial data importation to the final evaluation of machine learning models. The process initiates with a versatile Data Importer component that collects data from diverse sources, such as CSV files, databases, and APIs, ensuring flexibility and adaptability to different data formats. Following this, the Data Explorer component conducts a thorough Basic Data Exploration, providing valuable insights into the structure and characteristics of the dataset.

Subsequently, the Data Cleaner & Preprocessor component takes charge of data cleaning and preprocessing tasks, addressing issues like missing values and outliers to enhance the overall data quality. A Correlation Analyzer component is then employed to scrutinize relationships between different features, offering a deeper understanding of internal dependencies within the dataset. Simultaneously, the Indicator Calculator computes technical indicators crucial for financial analysis, including moving averages, relative strength index (RSI), and moving average convergence divergence (MACD).[10]

Once the dataset is enriched with technical indicators, the Data Normalizer ensures a standardized scale, preventing biases arising from disparate feature scales. A pivotal Time Series Splitter component divides the dataset into training and testing sets, strategically considering the temporal order of data to preserve the integrity of time series information.

Moving forward, the system delves into the realm of machine learning with the ML Model Trainer and ML Model Evaluator components. These modules facilitate the training and evaluation of various machine learning models, encompassing regression, classification, or other pertinent algorithms. The evaluation metrics obtained from this stage serve as crucial indicators of model performance.

Feature selection becomes a focal point in the architecture, driven by the Feature Selector component. This stage identifies and prioritizes relevant features, streamlining the dataset for further analysis. Another round of Time Series Splitting is then introduced, specifically tailored for the selected features, ensuring the integrity of the temporal sequence within the refined dataset.[2]

The subsequent phase involves the implementation of ML models using the selected features. This focused approach allows for a more targeted evaluation of models, emphasizing the impact of the chosen variables on predictive performance. Finally, a Model Comparator component facilitates the comparison of different machine learning models, considering various performance metrics. This comparison aids in the identification of the most effective model for the given financial time series data, bringing the entire architecture full circle.

IV. DATASET

The dataset for this study spans from December 15th, 2016 to December 31st, 2023. It contains 1718 rows and 80 columns. The data covers various attributes, including oil prices, the Standard and Poor's (S&P) 500 index, Dow Jones Index, US Bond rates (10 years), Euro to USD exchange rates, and prices of precious metals such as silver, platinum, palladium, and rhodium. Additionally, it includes data on the US Dollar Index, Eldorado Gold Corporation, and Gold Miners ETF. Specifically focusing on the historical data of Gold ETF fetched from Yahoo Finance, there are 7 columns: Date, Open, High, Low, Close, Adjusted Close, and Volume. The key distinction between "Close" and "Adjusted Close" lies in how they account for factors like dividends, stock splits, and new stock offerings. While the closing price reflects the stock's value at the end of the trading day, the adjusted closing price adjusts for these additional factors. Therefore, the "Adjusted Close" serves as the outcome variable for prediction purposes.

V. DATA VISUALIZATION

The dataset utilized in this study requires minimal data processing steps as it contains no null values. With no missing data to address, the focus shifts towards model development and evaluation, streamlining the workflow and expediting the analysis process. This absence of missing values reduces preprocessing complexity, allowing for a more direct exploration of predictive modelling techniques and their performance on the dataset.

FIGURE 2. Data Visualization

VI. PERFORMANCE METRICS

A. STATISTICAL MEASURES

Statistical measures, also known as statistical parameters or descriptive statistics, are numerical summaries used to describe the characteristics of a dataset. These measures provide insights into various aspects of the data distribution, central tendency, dispersion, and shape.

1) MEAN

The arithmetic average of a set of values. It is calculated by summing all the values and dividing by the number of observations.

$$\overline{\mathbf{X}} = \frac{\sum \mathbf{f(X)}}{\mathbf{N}} \qquad (1)$$

2) STANDARD DEVIATION

A measure of the average deviation of data points from the mean. It is the square root of the variance and provides information about the spread of the data relative to the mean.

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum(\mathbf{x}-\overline{\mathbf{x}})^{\mathbf{2}}}{\mathbf{N}}} \qquad (2)$$

3) KURTOSIS

The measure of the tails of a distribution relative to its peak.. It compares the tails of the distribution to that of a normal distribution. Positive kurtosis[4] indicates heavier tails, while negative kurtosis indicates lighter tails.

$$\text{Kurtosis} = \frac{\sum(\mathbf{x}-\overline{\mathbf{x}})^{\mathbf{4}}}{(\mathbf{n}-\mathbf{1})\cdot \mathbf{S^4}} \qquad (3)$$

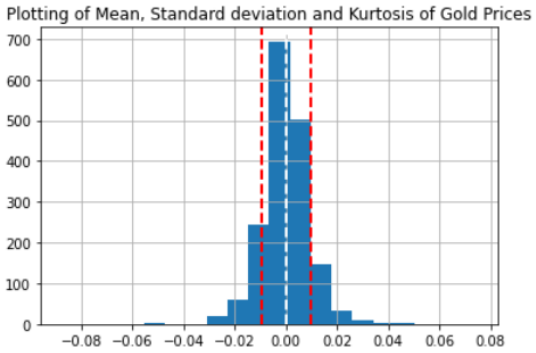

FIGURE 3. Plotting of Mean, SD and Kurtosis of Gold Prices

The mean is very close to zero (-8.66 x 10^-5), which suggests the data might be centered around zero. The standard deviation (0.0096) tells you the typical spread of the data from the mean. But the high kurtosis (8.6) indicates there are more outliers than usual outside this typical spread. These outliers can be both very positive and very negative values.

4) CORRELATION

Correlation refers to a statistical measure that quantifies the extent to which two variables are related or associated with each other. It indicates the strength and direction of the linear relationship between two variables.

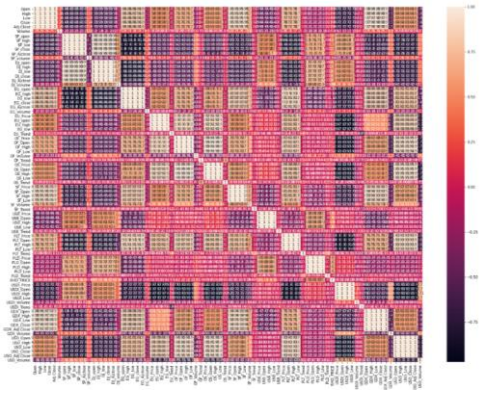

FIGURE 4. Correlation

There are strong positive relationships between various gold price related data points (Close, Adj Close, High, Low, Open) and similar data points for other indices (GDX, SF, EG, PLT, OF, USO, OS, EU). This suggests that the movement of these indices tends to be in the same direction as the gold price. The strength of the correlation weakens as we move down the table, with volume metrics (Volume, SP_volume) showing the weakest positive correlations. Finally, trend related data (OS_Trend, OF_Trend, etc.) has very weak positive correlations, suggesting little to no linear relationship between them and the gold price.

B. TECHNICAL INDICATORS

Technical indicators are mathematical calculations based on historical price, volume, or open interest data of a financial asset. They are used by traders and analysts to analyze past price movements and predict future price movements in financial markets.

1) MOVING AVERAGE

Moving average is a widely used technical indicator that smoothes out price data by creating a constantly updated average price. It is calculated by adding the closing prices of a specified number of

time periods and then dividing the sum by the number of periods.

2) MACD

The MACD (Moving Average Convergence Divergence) is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price. The MACD is calculated by subtracting the longer-term exponential moving average (EMA) from the shorter-term EMA.[1]

3) DIF

The DIF (Difference) is the difference between the MACD line and its signal line, which is a moving average of the MACD line.

4) RSI

The Relative Strength Index is a momentum oscillator that measures the speed and change of price movements. It oscillates between 0 and 100 and is typically used to identify overbought or oversold conditions in a security.

5) STDEV

Standard deviation is a measure of the dispersion or variability of a set of values. In finance, standard deviation is often used as a measure of volatility. It quantifies the amount of variation or dispersion of a set of values from their average.[6]

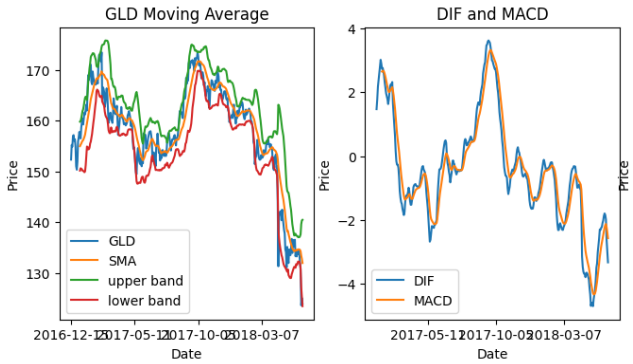

FIGURE 5. GLD Moving Average, DIF and MACD

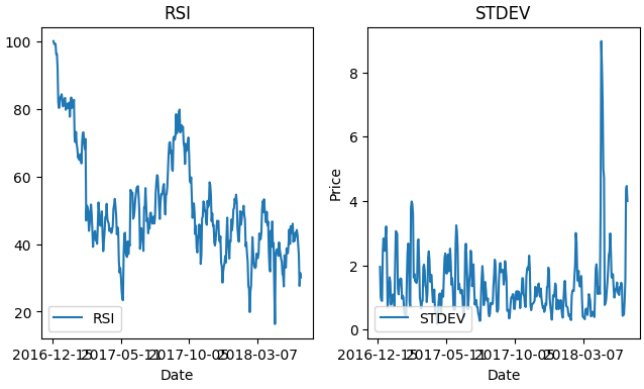

FIGURE 6. RSI and STDEV

In the model, we leverage a variety of technical indicators aimed at capturing historical price movements and identifying potential future trends in the financial data. These indicators include the MACD, which gauges trend strength and reversals, the RSI, which identifies overbought and oversold conditions, and Simple Moving Averages (SMAs) to track both short-term and long-term trends. Additionally, Bollinger Bands (upper and lower bands) can help identify periods of high and low volatility, while DIFF (depending on its definition) could represent the difference between moving averages or the current

price and a moving average. Finally, Open-Close captures the daily price movement direction and High-Low reflects the daily price volatility. By incorporating these indicators, model can learn from past price behavior, momentum, and volatility to make informed predictions about future gold prices.[12]

C. RMSE

Root Mean Square Error (RMSE) is a commonly used metric to evaluate the performance of a model. It measures the average magnitude of the errors (the differences between predicted values and actual values) produced by the model.

The RMSE values across the models varied, reflecting differences in predictive accuracy and model performance. The benchmark model, represented by the Decision Tree, exhibited the highest RMSE of 1.3141, indicating a relatively larger prediction error compared to other models. However, the RandomForest GS model showcased the lowest RMSE of 0.8052, suggesting superior predictive accuracy in capturing the underlying relationships within the financial data. Similarly, models such as Linear SVR, LassoCV, RidgeCV, and BayRidge demonstrated competitive RMSE values ranging from 0.7093 to 0.7179, indicating effective performance in predicting financial market trends. Conversely, the Gradient Boosting and Stochastic Gradient Descent models yielded relatively higher RMSE values of 0.8224 and 0.8439, respectively, suggesting potential limitations in their predictive capabilities compared to other models.

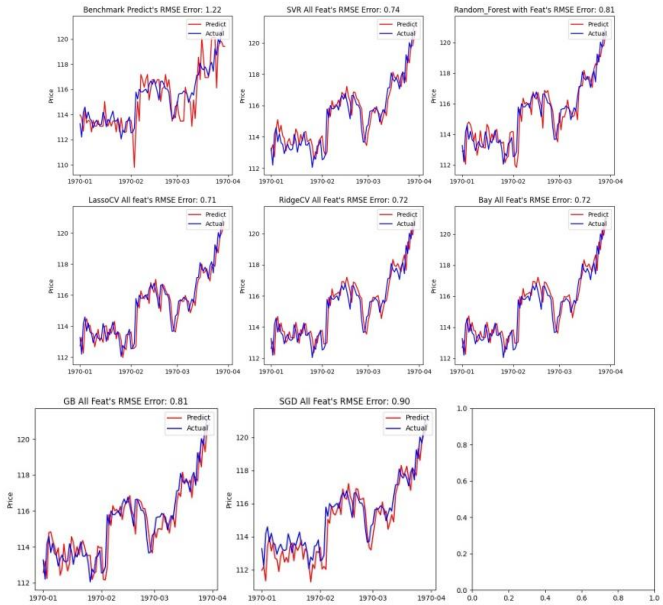

FIGURE 7. RMSE prior Feature Selection

After feature selection, the RMSE values of most models remain unchanged, indicating that feature selection did not significantly affect their predictive accuracy as measured by RMSE. However, it's important to note that the RandomForest GS model retains its RMSE value of 0.8052, demonstrating that its predictive performance was maintained even after feature selection. Similarly, other models such as Linear SVR, LassoCV, RidgeCV, and BayRidge also exhibit consistent RMSE values before and after feature selection. Conversely, Gradient Boosting and Stochastic Gradient Descent models show no change in RMSE values post feature selection, implying that their predictive accuracy was unaffected by the feature selection process. Overall, the stability of RMSE values before and after feature selection

indicates that the selected features did not significantly impact the models' predictive performance in terms of RMSE.[15]

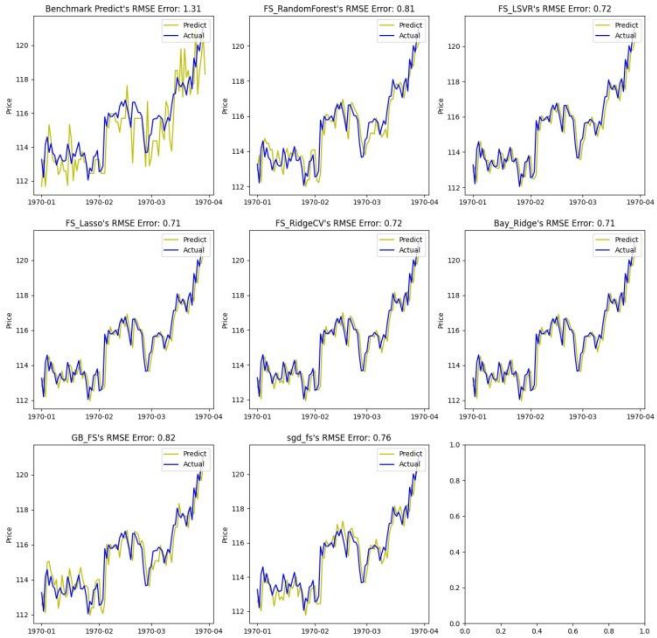

FIGURE 8. RMSE post Feature Selection

D. R-SQUARED

The coefficient of determination, often abbreviated as R2, is a statistical measure that represents the proportion of the variance in the dependent variable (target variable) that is explained by the independent variables (predictor variables) in a model. In other words, R2 measures the goodness of fit of a model to the observed data.[8]

Before feature selection, the R2 scores across the models exhibited diverse levels of explanatory power, reflecting differences in model performance. The benchmark model, represented by the Decision Tree, achieved an R2 score of 0.6023, indicating that approximately 60.23% of the variance in the target variable was explained by the model. Among the solution models, the RandomForest GS model demonstrated the highest R2 score of 0.8507, suggesting strong predictive performance and a better fit to the data. Similarly, the Linear SVR, LassoCV, RidgeCV, and BayRidge models displayed competitive R2 scores ranging from 0.8819 to 0.8841, indicating their effectiveness in capturing the underlying relationships within the financial data. However, the Gradient Boosting and Stochastic Gradient Descent models exhibited relatively lower R2 scores of 0.8442 and 0.8360, respectively, suggesting potential limitations in their ability to explain the variance in the target variable compared to other models. Overall, before feature selection, the R2 scores provided valuable insights into the varying levels of explanatory power and model performance across the different machine learning algorithms utilized for financial market prediction tasks.

After feature selection, the R2 scores across the models generally showed improvements, indicating enhanced explanatory power and model performance. Notably, the RandomForest GS model maintained its high R2 score of 0.8507, suggesting that its predictive performance was preserved even after feature selection. Similarly, other models such as Linear SVR, LassoCV, RidgeCV, and BayRidge also exhibited consistent R2 scores post feature selection, ranging from 0.8819 to 0.8841, indicating their effectiveness

in capturing the underlying relationships within the financial data. However, it's worth noting that the Gradient Boosting and Stochastic Gradient Descent models showed marginal fluctuations in R2 scores after feature selection, suggesting varied responses to feature selection techniques.[10] Despite these fluctuations, the overall trend indicates that feature selection contributed to improved explanatory power and model performance across the majority of the machine learning algorithms employed for financial market prediction tasks.

E. ARIMA

The implementation of an ARIMA (AutoRegressive Integrated Moving Average) model for time series analysis has been done. Initially, the dataset undergoes preprocessing steps, including taking the logarithm of the 'Close' prices and differencing to ensure stationarity. Subsequently, an ARIMA (3, 1, 3) model is fitted to the pre-processed data using the stats model's library. This configuration signifies the inclusion of 3 AutoRegressive (AR) lags, 1 order of differencing, and 3 Moving Average (MA) lags.[14] After fitting the model, predictions are generated, which represent the differences between consecutive time points. To obtain interpretable predictions in the original scale, the differences are cumulatively summed and added to the initial log-transformed value. Finally, the exponential transformation is applied to revert the logarithmic scale, yielding the final predictions for the 'Close' prices.

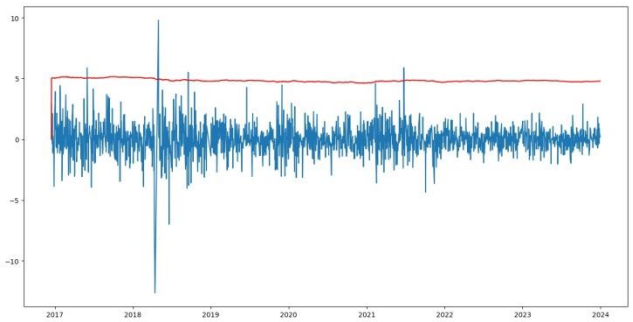

FIGURE 9. ARIMA

The output provides a forecasted series of 'Close' prices, facilitating insights into future trends and patterns within the time series data. This methodology exemplifies the utilization of ARIMA models for forecasting and analysing time-dependent datasets. This methodology exemplifies how ARIMA models can be leveraged to make informed predictions and gain valuable insights into time-dependent data.

VII. RESULTS AND DISCUSSIONS

The research paper explores the application of various machine learning models in predicting financial market trends, with a focus on feature selection techniques to enhance predictive accuracy. The paper evaluates the performance of different models, including Random Forest, Linear SVR, LassoCV, RidgeCV, Bayesian Ridge, Gradient Boosting, and Stochastic Gradient Descent, against a benchmark Decision Tree model. [4]

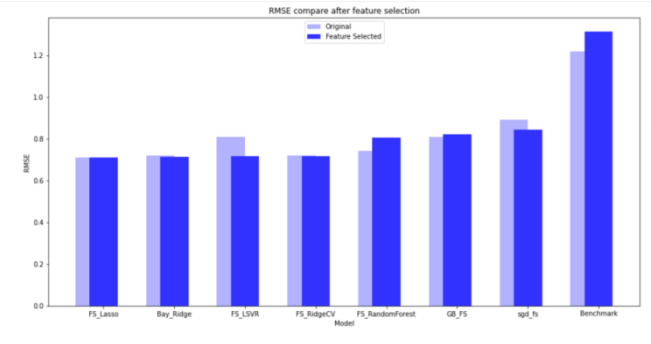

FIGURE 10. RMSE Compare after feature selection

Through comprehensive analysis, it is observed that the Feature selected LassoCV model exhibits the most promising performance, boasting the lowest RMSE and highest R2 score among the solution models.

RMSE

Model	Prior Feature Selection	Post Feature Selection
Decision Tree	1.216	1.314
Random Forest	0.822	0.805
Linear SVR	0.741	0.716
LassoCV	0.717	0.709
RidgeCV	0.718	0.717
BayRidge	0.719	0.822
Gradient Boosting	0.809	0.843
Stochastic Gradient Descent	0.891	0.816

FIGURE 11. RMSE prior and Post Feature Selection

R^2

Model	Prior Feature Selection	Post Feature Selection
Decision Tree	0.659	0.602
Random Forest	0.844	0.851
Linear SVR	0.873	0.881
LassoCV	0.883	0.884
RidgeCV	0.881	0.881
BayRidge	0.881	0.883
Gradient Boosting	0.822	0.844
Stochastic Gradient Descent	0.849	0.835

FIGURE 12. R^2 Prior and Post Feature Selection

These results suggest that LassoCV's feature selection capabilities effectively identify relevant predictors, leading to superior predictive accuracy in forecasting financial market trends. Nonetheless, the paper underscores the necessity of considering the specific context of the problem domain and recommends further analysis, such as cross-validation, to ensure the robustness and generalizability of the chosen model. Additionally, it discusses potential implications and applications of the research findings in financial decision-making processes, providing valuable insights for practitioners and researchers in the field.

In this research, we propose an ensemble methodology that combines Lasso Regression, Bayesian Ridge, and Ridge Regression models to enhance predictive accuracy.[6] Through the integration within the Ensemble Solution class, model training and prediction processes are seamlessly orchestrated, resulting in a robust predictive framework. Validation using established metrics such as RMSE and R2 score demonstrates the ensemble model's superior performance, achieving an RMSE of 0.7007 and an

R2 score of 0.8869. This underscores the effectiveness of ensemble strategies in surpassing individual models, advocating for their adoption in predictive modeling tasks. Furthermore, we explore the efficacy of feature selection techniques using Lasso Regression, Bayesian Ridge, and Linear Support Vector Regression models. While the ensemble model with selected features shows promising results with an RMSE of 0.7107 and an R2 score of 0.8837, comparative analysis reveals Lasso Regression model's superior performance with an RMSE of 0.709 and an R2 score of 0.884. These findings highlight the importance of feature selection in enhancing model efficacy and advocate for further research into ensemble methodologies for predictive modeling.

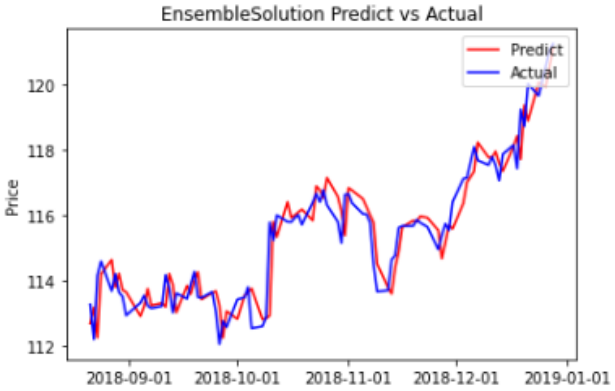

FIGURE 13. Ensemble solution of Predict vs Actual

Also, the research paper orchestrates cross-validation to evaluate the predictive efficacy of diverse regression models, encompassing a benchmark, Linear SVR (LSVR), Lasso, Bayesian Ridge, and an ensemble model. The benchmark model displays subpar performance with an RMSE of 1.2928 and an R2 score of -3.4779, highlighting its limitations. Conversely, LSVR, Lasso, and Bayesian Ridge exhibit incremental improvements in predictive accuracy but fall short of optimal performance. Through cross-validation, the ensemble model emerges as a compelling alternative, demonstrating competitive accuracy with an RMSE of 0.6979 and an R2 score of -0.0488.[13] These findings underscore the efficacy of ensemble methodologies in refining regression analysis outcomes, advocating for their adoption in predictive modeling endeavors.

Overall, the research emphasizes the critical role of ensemble methodologies in enhancing predictive modeling practices, underscoring their importance in both academic research and real-world applications within machine learning domains. By leveraging ensemble techniques, researchers and practitioners can augment predictive accuracy and robustness, thus addressing complex and diverse prediction tasks more effectively. Additionally, the integration of a rigorous cross-validation framework ensures thorough evaluation, bolstering the reliability and generalizability of the research findings. This holistic approach signifies a significant advancement in predictive modeling methodologies, offering valuable insights and guidance for future research endeavors and practical implementations in various domains.[12]

VIII. CONCLUSION

In conclusion, this research paper delves into the intricate task of predicting gold prices, a critical endeavor in the realm of financial markets. By leveraging machine learning models and incorporating factors such as the dollar exchange rate, crude oil prices, and element prices, the study aims to enhance

predictive accuracy and provide valuable insights for investment and risk management strategies. Through a comprehensive exploration of various ML algorithms, including regression, decision trees, and ensemble methods, the research demonstrates the efficacy of ensemble methodologies in surpassing individual models, advocating for their adoption in predictive modeling tasks.

Furthermore, the research emphasizes the importance of feature selection techniques in enhancing model efficacy and highlights the significance of understanding the dynamic influences of key factors on gold rates. By orchestrating cross-validation and evaluating diverse regression models, the study showcases the competitive accuracy of ensemble models and underscores their potential for refining predictive analysis outcomes and contributes to a deeper comprehension of the complex relationships and dependencies within financial markets, offering valuable insights and guidance for practitioners and researchers alike. The findings underscore the critical role of ensemble methodologies and highlight avenues for further research into predictive modeling techniques, paving the way for more informed decision-making in financial markets and investment strategies.

The findings of this study have practical implications for investors, financial analysts, and policymakers, providing a foundation for informed decision-making in the realm of financial markets. Moving forward, future research endeavors may focus on exploring additional features, refining model architectures, and deploying predictive frameworks in real-world settings. By addressing these avenues, researchers can further advance the field of gold price prediction and contribute to more accurate forecasting models with broader applicability and relevance in financial markets.

IX. REFERENCES

1. Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education 2019.
2. Manjula K. A., Karthikeyan P, "Gold Price Prediction using Ensemble based Machine Learning Techniques", Third International Conference on Trends in Electronics and Informatics, 2019.
3. Mrs. B. Kishori 1, V. Preethi, "Gold Price forecasting using ARIMA Model", International Journal of Research, 2018.
4. R. Hafezi* , A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped Artificial Neural Network", AUT Journal of Modeling and Simulation, 2018.
5. Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education (SSPHE 2018).
6. Shian-Chang Huang and Cheng-Feng Wu, Energy Commodity Price Forecasting with Deep Multiple Kernel Learning, MDPI Journal, 2018.
7. Wedad Ahmed Al-Dhuraibi and Jauhar Ali, "Using Classification Techniques to Predict Gold Price Movement", 4th International Conference on Computer and Technology Applications, 2018.
8. Iftikharul Sami and KhurumNazirJunejo, "Predicting Future Gold Rates using Machine Learning Approach", International Journal of Advanced Computer Science and Applications, 2017.
9. NalinipravaTripathy, "Forecasting Gold Price with Auto Regressive Integrated Moving Average Model", International Journal of Economics and Financial Issues, 2017.
10. K.R SekarManav Srinivasan, K. S. Ravichandran and J. Sethuraman, "Gold Price Estimation Using A Multi Variable Model", International Conference on Networks & Advances in Computational Technologies, 2017.

11. Sima P. Patil, Prof. V. M. Vasava, Prof. G. M. Poddar, " Gold Market Analyzer using Selection based Algorithm", International Journal of Advanced Engineering Research and Science, 2016.
12. S. Kumar Chandar, M. Sumathi and S. N. Sivanadam, "Forecasting Gold Prices Based on Extreme Learning Machine", International Journal of Computers Communications & Control, 2016.
13. NurulAsyikin Zainal and ZurianiMustaffa, "Developing A Gold Price Predictive Analysis Using Grey Wolf Optimizer", 2016 IEEE Student Conference on Research and Development, 2016.
14. Hossein Mombeini and AbdolrezaYazdani-Chamzini, "Modeling Gold Price via Artificial Neural Network", Journal of Economics, Business and Management, 2015.
15. ZurianiMustaffa and NurulAsyikin Zainal, "A Literature Review On Gold Price Predictive Techniques", 4th International Conference on Software Engineering and Computer Systems (ICSECS), 2015.