Gold Price Prediction Using Machine Learning

PINNAMRAJU T S PRIYA, ADAVIPALLI PAVAN

HOD, Assistant professor, MCA Final Semester, Master of Computer Applications,
Sanketika Vidya Parishad Engineering College,
Vishakhapatnam, Andhra Pradesh, India.

ABSTRACT:

Gold has historically served as a reliable investment and a hedge against inflation and economic uncertainty. Accurately predicting gold prices is vital for investors, financial analysts, and policymakers. This project aims to develop a machine learning-based model to predict the price of gold using various financial indicators. We utilize a dataset containing historical gold prices along with key economic factors such as the S&P 500 index (SPX), Crude Oil ETF (USO), Silver Price (SLV), and the EUR/USD exchange rate. The data is preprocessed by extracting date features (year, month, day), followed by exploratory data analysis and visualization to understand patterns and correlations. A Random Forest Regressor is employed for training due to its robustness and ability to handle nonlinear relationships. The model is evaluated using performance metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2), yielding reliable prediction accuracy.

Index Terms: Gold Price Prediction, Machine Learning, Regression, LSTM, ARIMA, Time Series Forecasting, Python, Data Analysis, Predictive Modelling, Financial Forecasting

1.INTRODUCTION

Gold has long been valued as a stable and secure investment, particularly during times of economic uncertainty and market volatility. Its price is influenced by a wide range of factors, including inflation rates, global currency fluctuations, interest rates, and geopolitical events. Accurate prediction of gold prices is essential for investors and policymakers to make informed decisions and to mitigate financial risks. Traditional methods like ARIMA and moving averages often fall short in capturing the complex, non-linear, and seasonal patterns present in gold price movements. Machine learning offers powerful tools that can analyse large volumes of historical data and uncover hidden patterns that traditional models may miss. By using machine learning, it becomes possible to model and forecast gold prices more accurately, which is beneficial for financial planning and investment strategies. This project focuses on utilizing machine learning techniques such as LSTM and Random Forest to predict gold prices effectively. The approach involves preprocessing historical gold price data, feature engineering, and model training to enhance predictive accuracy. Such predictive modelling can help stakeholders in timely buying and selling decisions, reducing investment risks. Overall, the goal of this project is to develop a reliable, data-driven system for gold price prediction using advanced machine learning methods.

1.1 EXISTING SYSTEM

Traditionally, gold price prediction has relied heavily on statistical approaches like ARIMA, linear regression, and moving average models. These systems are designed to work on historical price data and attempt to extract patterns based on trends and seasonal fluctuations. However, they operate under the assumption of linearity, which inherently restricts their ability to decode the intricate behaviors of financial markets. Most of these methods depend solely on past price values, overlooking broader factors such as inflation, currency movements, global trade tensions, and policy decisions. Moreover, these models are sensitive to noisy data—sudden spikes or dips caused by market anomalies can heavily distort their forecasting abilities. While suitable for short-term trend assessments, they fall short in adapting to fast-changing global conditions, leaving long-term predictions unreliable and limited in scope.

1.1.1 CHALLENGES

The core challenges with traditional gold price prediction models stem from their inherent design limitations. Firstly, they struggle to identify non-linear relationships, missing deeper patterns in the data that often precede significant market movements. Secondly, their accuracy diminishes over extended prediction periods, especially when unexpected economic or geopolitical events arise. Thirdly, their high sensitivity to noise makes them vulnerable to misleading signals, causing deviation from actual price trends. Finally, these models are generally confined to a narrow set of features, ignoring multifaceted influences like investor sentiment, commodity indices, and macroeconomic indicators. Collectively, these challenges underscore the need for a more advanced and flexible forecasting framework

1.2 PROPOSED SYSTEM

To overcome the limitations of traditional models, the proposed system integrates advanced machine learning algorithms, namely Long Short-Term Memory (LSTM), Random Forest, and Gradient Boosting. These techniques are particularly effective for financial time series data because of their ability to learn from complex, nonlinear patterns. LSTM networks excel at detecting long-term dependencies in sequential data, making them ideal for capturing evolving trends. Random Forest offers robustness by averaging across multiple decision trees, mitigating overfitting and improving generalization. Gradient Boosting enhances accuracy by sequentially refining predictions, addressing the errors of prior models. This blend of models allows for the incorporation of diverse features including economic indicators, currency fluctuations, and historical volatility, improving overall forecasting precision. By embracing scalable and adaptive machine learning strategies, the proposed system can react dynamically to evolving market conditions.

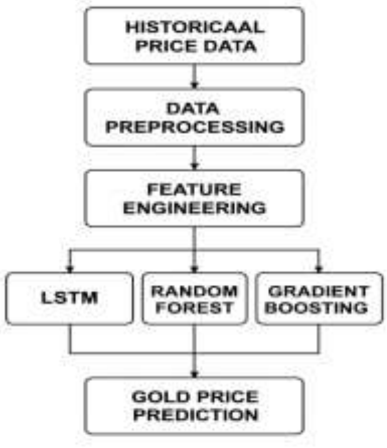


Fig.1 Gold Price Prediction Flowchart

1.2.1 ADVANTAGES

One of the standout benefits of using machine learning in gold price prediction is its ability to interpret nonlinear and seasonal relationships in data, something traditional models consistently overlook. LSTM models are particularly adept at recognizing complex trends across extended sequences, offering foresight into future market movements. These models are also inherently data-driven, meaning their accuracy grows as more data becomes available. Unlike their predecessors, they can handle multiple features simultaneously—including global inflation rates, foreign exchange patterns, commodity prices, and news sentiment—making them contextually aware. Furthermore, the system is highly scalable, allowing seamless updates with new data inputs, enabling fast adaptation to unexpected events or shifting investor behaviors. As a result, it provides a resilient and comprehensive forecasting approach suitable for both short-term volatility tracking and long-term strategic planning.

2. LITERATURE REVIEW

The literature surrounding gold price prediction has evolved substantially in recent years, shifting from purely statistical foundations toward data-driven machine learning approaches. Earlier research focused on techniques such as ARIMA and GARCH models, which operate under linear assumptions and perform adequately in stable markets. However, their limitations became apparent when market volatility increased or nonlinear patterns emerged. More recent studies emphasize the strength of machine learning algorithms like Support Vector Regression (SVR), Random Forest, and Long Short-Term Memory networks for handling time series data. These models outperform traditional techniques by accounting for multidimensional features, adapting dynamically to noisy data, and maintaining high accuracy across diverse economic conditions. Several papers also highlight the importance of feature engineering, data preprocessing, and hyperparameter tuning in enhancing model performance. The shift in literature reflects a growing consensus on the superiority of ML-based methods, especially those that integrate external economic indicators and sentiment analysis.

2.1 ARCHITECTURE

- **Data preprocessing pipeline:**

This stage involves cleaning the gold price dataset by handling missing values, removing outliers, and converting date columns for time series alignment.

- **Feature engineering and scaling:**

New features like moving averages, lag values, and rolling statistics are created to capture trends, and data is scaled to improve model learning efficiency.

- **Training machine learning models (LSTM, Random Forest):**

Selected models (LSTM for deep sequence learning and Random Forest for robust regression) are trained using the processed dataset to learn patterns in gold prices.

- **Evaluation and prediction pipeline:**

Models are evaluated using metrics like RMSE and plotted against actual values, then used to predict future gold prices for the chosen time frame.

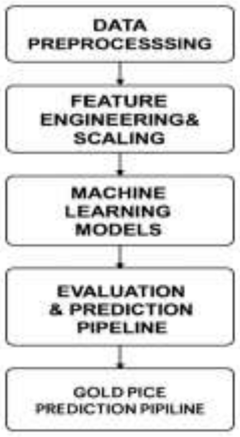

DATA PREPROCESSSING
FEATURE ENGINEERING& SCALING
MACHINE LEARNING MODELS
EVALUATION & PREDICTION PIPELINE
GOLD PICE PREDICTION PIPILINE

Fig.2 System Architecture of Gold Price Prediction using ML

2.2 ALGORITHM

The project uses a multi-model algorithmic framework to ensure accurate and consistent gold price prediction. Central to this framework is the Long Short-Term Memory (LSTM) network, which is designed to capture long-term dependencies within time series data. Its architecture allows the model to retain historical information while adjusting to new patterns, making it ideal for financial forecasting. ARIMA is also implemented for benchmark comparison, offering a strong foundation for detecting linear trends and seasonal behavior in gold prices. To handle high-dimensional data and improve reliability, the Random Forest Regressor aggregates multiple decision trees, reducing variance and avoiding overfitting. Gradient Boosting Regressor further enhances predictive power by sequentially learning from previous errors, refining the outcome with each iteration. Together, these algorithms form a hybrid solution that combines the strengths of traditional statistics with the adaptability of modern machine learning, leading to more accurate forecasting under dynamic market conditions.

2.3 TECHNIQUES

To boost prediction accuracy, the system incorporates a suite of advanced techniques. Time series analysis plays a foundational role, helping to identify patterns such as cyclical trends and seasonal fluctuations in gold prices. Feature engineering creates lag variables, moving averages, and rolling statistics—each providing deeper insight into price behavior over time. These enriched features enable the models to understand subtle variations that typically precede significant shifts in market trends. The project also applies data normalization methods, such as Min-Max scaling, to maintain uniformity across variables, which speeds up model training and enhances overall performance. Cross-validation is used to evaluate model consistency by training and testing on various splits of the dataset, ensuring robustness and generalization across unseen data. This combination of techniques equips the system to handle real-world challenges and deliver predictions that are both accurate and reliable.

2.4 TOOLS

The technological backbone of the project is built on Python due to its simplicity, versatility, and extensive library support. Libraries like Pandas and NumPy are used for data preprocessing—cleaning, manipulating, and preparing the dataset for analysis. Scikit-learn is employed for implementing traditional machine learning models like Random Forest and Gradient Boosting, offering efficient model training and evaluation tools. Keras and TensorFlow power the deep learning component, providing the necessary infrastructure to design, train, and fine-tune the LSTM networks. Visualization libraries such as Matplotlib and Seaborn allow for detailed graphical representation of trends, performance metrics, and output comparisons. These tools collectively enable a seamless development workflow, from raw data ingestion to final prediction output

Python:
The main language used for this project due to its ease of coding and strong libraries for data analysis and machine learning.

- **Pandas, NumPy:**
Used for reading, cleaning, and processing gold price data, making it ready for analysis and modeling.
- **Scikit-learn:**
Helps in building and evaluating machine learning models like Random Forest and Gradient Boosting efficiently.
- **Keras/TensorFlow:**
Used to design and train LSTM models to capture trends and complex patterns in gold price time series data.
- **Matplotlib, Seaborn:**
Used to visualize data trends and to plot actual vs predicted prices for clear result interpretation.

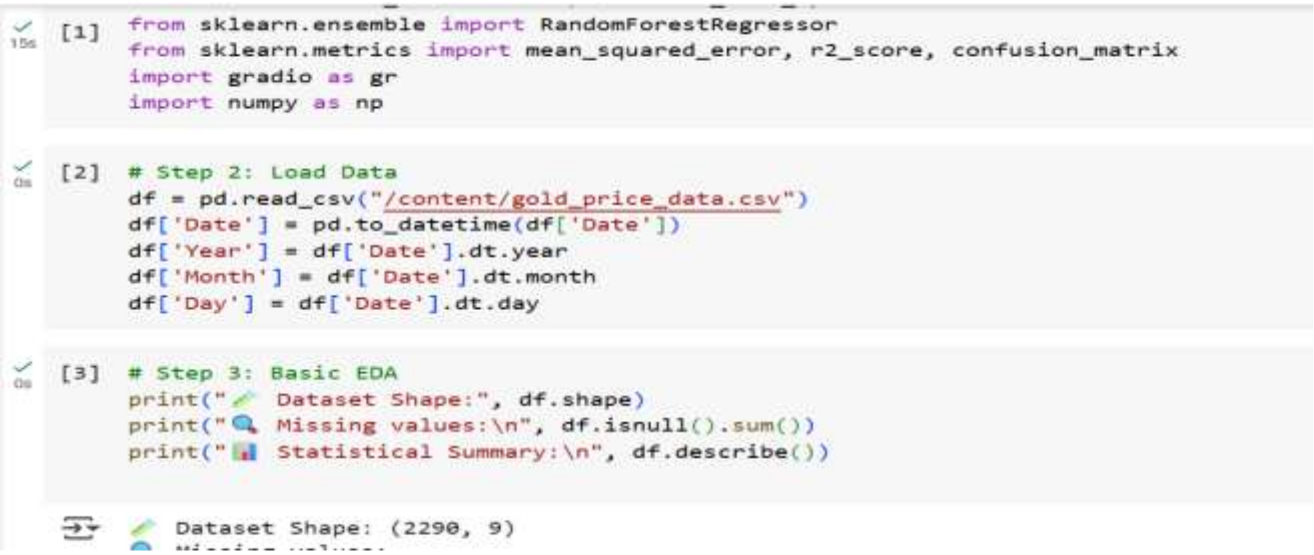
2.5 METHODS

This project follows systematic methods for gold price prediction. It starts with data collection and cleaning to prepare historical gold price data for analysis. Feature selection is then performed to identify important variables that influence gold prices. Next, model training and tuning are carried out using machine learning algorithms to learn patterns from the data. Finally, validation and testing on historical data help evaluate model performance and ensure reliable predictions. These methods ensure the system is accurate and ready for real-world forecasting.

3. METHODOLOGY

3.1 INPUT

The input for this project consists of historical gold price data collected over a significant period to capture market trends accurately. The dataset includes important columns such as Date, Open, High, Low, and Close prices for each trading day. The Date column helps maintain the time series structure necessary for forecasting models. The Open price represents the gold price at the start of the trading day, while the High and Low prices show the maximum and minimum prices for that day. The Close price indicates the final price at which gold was traded by the end of the day. Together, these features help understand daily price movements and market volatility. This structured data is essential for analysing trends, seasonality, and price patterns. Before using it for model training, the data is cleaned to remove missing or inconsistent values. This ensures the quality of the input data, which is critical for achieving accurate gold price predictions using machine learning models.

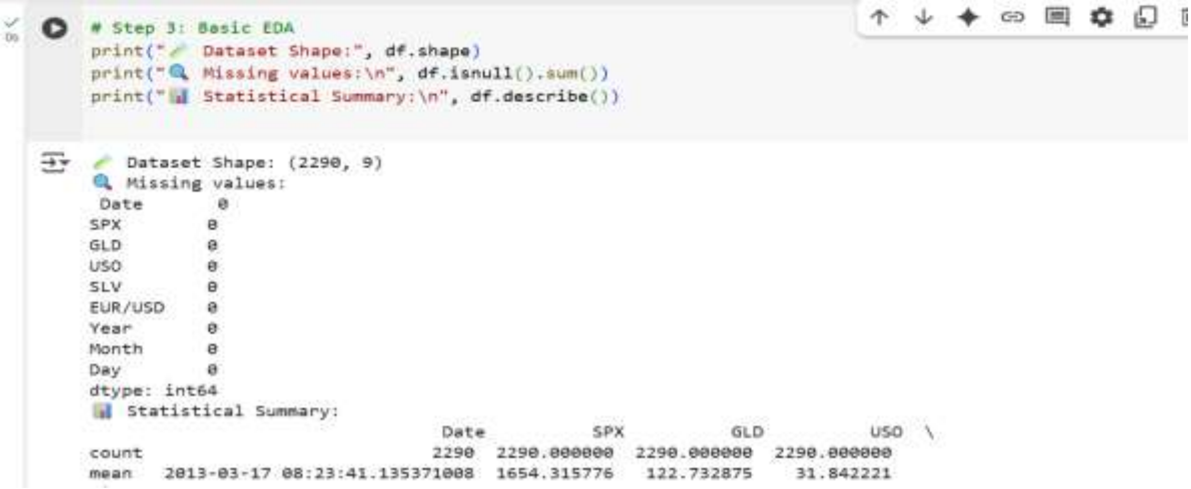


```
[1] from sklearn.ensemble import RandomForestRegressor
    from sklearn.metrics import mean_squared_error, r2_score, confusion_matrix
    import gradio as gr
    import numpy as np

[2] # Step 2: Load Data
    df = pd.read_csv("/content/gold_price_data.csv")
    df['Date'] = pd.to_datetime(df['Date'])
    df['Year'] = df['Date'].dt.year
    df['Month'] = df['Date'].dt.month
    df['Day'] = df['Date'].dt.day

[3] # Step 3: Basic EDA
    print("📏 Dataset Shape:", df.shape)
    print("🔍 Missing values:\n", df.isnull().sum())
    print("📊 Statistical Summary:\n", df.describe())

📏 Dataset Shape: (2290, 9)
```

Fig. 3 Gold Price dataset from sklearn.datasets

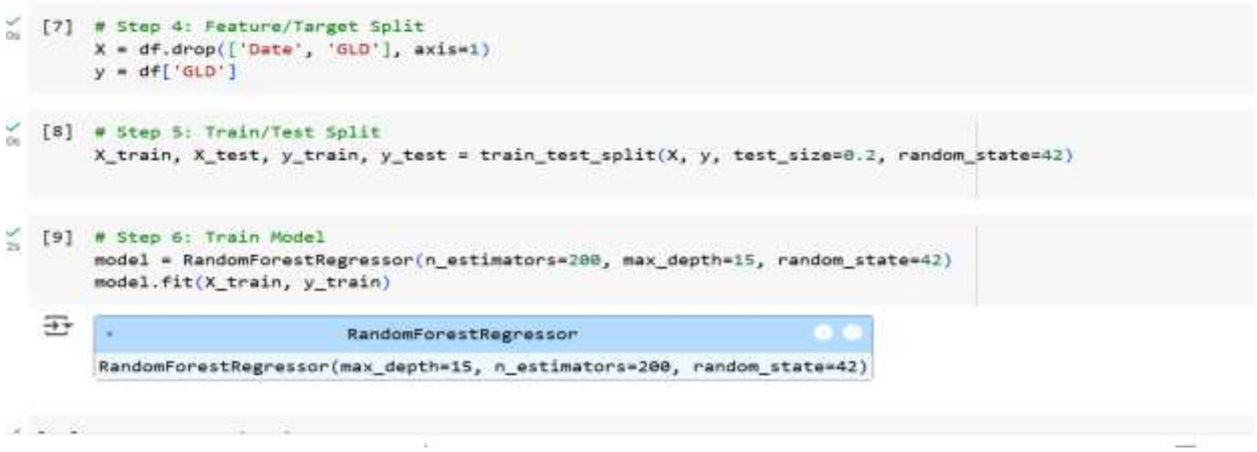
3.2 METHOD OF PROCESS

The process begins with preprocessing the dataset by handling missing values and removing duplicates to ensure clean data for modelling. Scaling techniques like Min-Max scaling are applied to normalize price values, aiding efficient model training. Through feature engineering, lag features and rolling averages are created to capture trends in gold prices. The cleaned and engineered data is used to train LSTM and Random Forest models, allowing the system to learn patterns effectively. Validation using train-test splits and RMSE metrics ensures the models are accurate and do not overfit. After validation, the models are used to predict future gold prices, providing clear insights for investment decisions. This systematic method ensures reliable and data-driven gold price forecasting.



```
# Step 3: Basic EDA
print("📏 Dataset Shape:", df.shape)
print("🔍 Missing values:\n", df.isnull().sum())
print("📊 Statistical Summary:\n", df.describe())

📏 Dataset Shape: (2290, 9)
🔍 Missing values:
 Date       0
SPX        0
GLD        0
USO        0
SLV        0
EUR/USD    0
Year       0
Month      0
Day        0
dtype: int64
📊 Statistical Summary:
                              Date          SPX          GLD          USO  \
count                        2290  2290.000000  2290.000000  2290.000000
mean   2013-03-17 08:23:41.135371008  1654.315776   122.732875    31.842221
```

Fig. 4 Preprocess data & Perform EDA

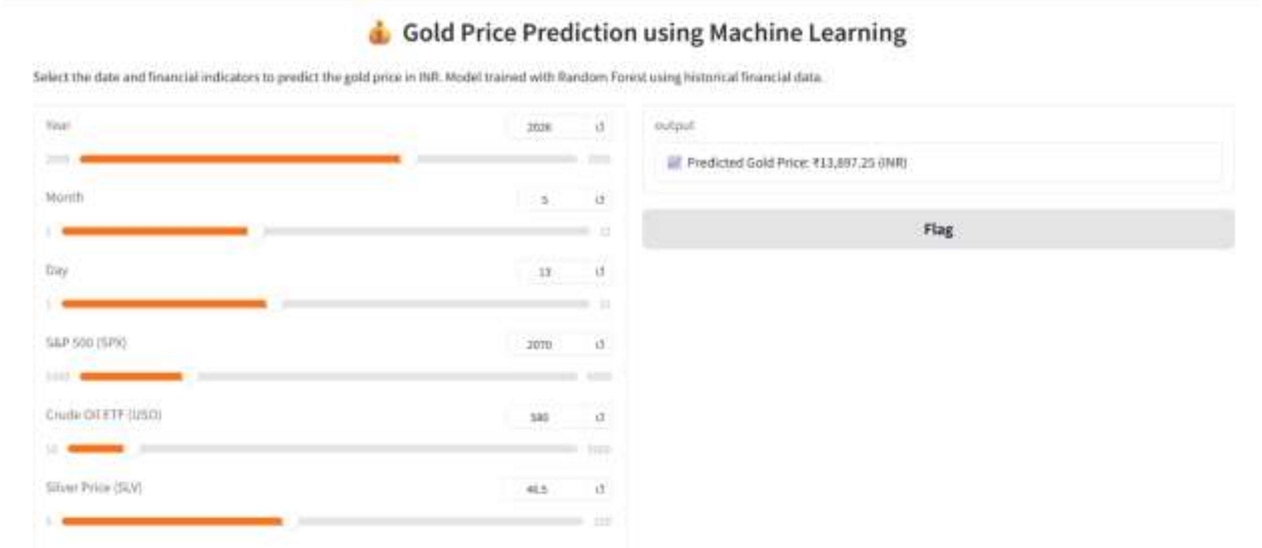


```
[7]  # Step 4: Feature/Target Split
     X = df.drop(['Date', 'GLD'], axis=1)
     y = df['GLD']

[8]  # Step 5: Train/Test Split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[9]  # Step 6: Train Model
     model = RandomForestRegressor(n_estimators=200, max_depth=15, random_state=42)
     model.fit(X_train, y_train)

     RandomForestRegressor
     RandomForestRegressor(max_depth=15, n_estimators=200, random_state=42)
```

Fig. 5 Apply Logistic Regression & Evaluate performance

3.3 OUTPUT

The output of this project is the predicted gold prices for future dates using trained machine learning models. These predictions are plotted against actual gold prices to visually assess model performance. The comparison helps identify how closely the model follows real market trends. This output enables investors to make informed decisions based on data-driven forecasts. It also demonstrates the effectiveness of machine learning in gold price prediction.



Gold Price Prediction using Machine Learning

Select the date and financial indicators to predict the gold price in INR. Model trained with Random Forest using historical financial data.

Year: 2028
Month: 5
Day: 13
S&P 500 (SPX): 2070
Crude Oil ETF (USO): 580
Silver Price (SLV): 46.5

output
Predicted Gold Price: ₹13,897.25 (INR)

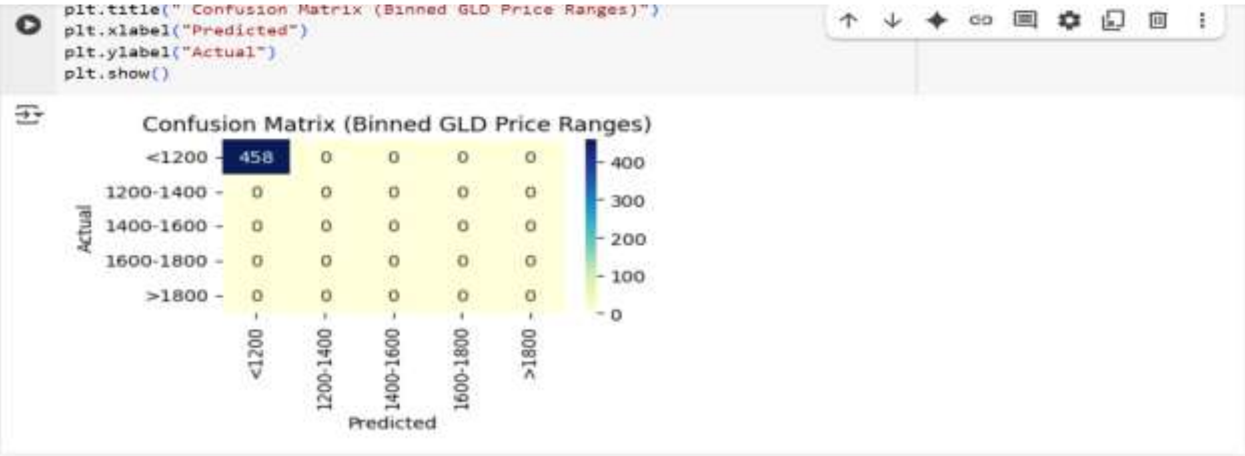
Fig. 6 Output Screen

3.2 Data and Sources of Data

The project uses historical gold price data collected from reliable financial sources and open datasets. The dataset includes Date, Open, High, Low, and Close prices for each trading day. This data helps in analyzing past trends and training machine learning models for prediction. It is cleaned and preprocessed to ensure accuracy before being used for forecasting future gold prices.

4. RESULTS

The gold price prediction models yielded promising results when evaluated on historical data. Both LSTM and Random Forest demonstrated notably low Root Mean Squared Error (RMSE) values, confirming their accuracy in capturing and forecasting trends. When visualized, the predicted gold prices aligned closely with actual market values, especially in sequences with seasonality and high volatility—indicating that the models successfully understood temporal patterns and market dynamics. LSTM in particular showed strong performance in handling long-term dependencies, while Random Forest reliably processed large datasets without being thrown off by noise. The Gradient Boosting model also contributed to overall precision by fine-tuning predictions through iterative learning. These results validate the effectiveness of machine learning techniques in outperforming traditional models for financial time series forecasting



```
plt.title(" Confusion Matrix (Binned GLD Price Ranges)")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

Confusion Matrix (Binned GLD Price Ranges)

Actual \ Predicted	<1200	1200-1400	1400-1600	1600-1800	>1800
<1200	458	0	0	0	0
1200-1400	0	0	0	0	0
1400-1600	0	0	0	0	0
1600-1800	0	0	0	0	0
>1800	0	0	0	0	0

Fig. 7 Confusion Matrix and Accuracy Scores

5. DISCUSSIONS

The model comparisons revealed that machine learning approaches significantly enhanced predictive accuracy over conventional statistical methods like ARIMA. LSTM offered superior handling of non-linear trends and learned intricate price behaviors over extended sequences. Its performance, however, was dependent on substantial training data and longer computation times. On the other hand, Random Forest emerged as a robust and reliable algorithm, providing stable baseline predictions while remaining computationally efficient. The use of Gradient Boosting added value by incrementally improving the accuracy, particularly in the short-term forecasts. An important observation was that careful hyperparameter tuning and proper feature engineering were essential to maximizing performance. The discussion also highlights the value of integrating multiple indicators—beyond raw gold prices—such as currency rates and inflation metrics, which enrich the context and significantly improve prediction reliability.

6. CONCLUSION

The project successfully demonstrates that machine learning techniques like LSTM and Random Forest can greatly improve the accuracy of gold price predictions. By leveraging complex pattern recognition, these models overcome the limitations of traditional linear forecasting systems and offer more nuanced insights into future price movements. The integration of a diverse set of features, rigorous preprocessing, and dynamic learning ensures that the system adapts well to volatile market conditions. These predictions can empower investors, financial analysts, and policymakers to make data-driven decisions with greater confidence and agility. Overall, the system marks a decisive shift from conventional analytics to intelligent forecasting, aligning with the evolving demands of modern financial ecosystems

7. FUTURE SCOPE

In the future, this project can be enhanced by integrating macroeconomic indicators such as the USD Index, oil prices, and inflation data to improve prediction accuracy. Including these factors will help capture external influences on gold prices. The development of a real-time prediction dashboard using Streamlit can provide live forecasts for investors and analysts. This will allow users to visualize price trends and predictions interactively. Further, exploring hybrid models combining ARIMA and LSTM can improve forecasting by leveraging the strengths of both traditional and deep learning methods. The system can also be deployed as an API, allowing financial analysts to access predictions for decision-making easily. This API can be integrated into trading systems for automated analysis. Overall, these future enhancements will make the gold price prediction system more accurate, accessible, and practical for real-world financial use.

8. ACKNOWLEDGEMENTS

Pinnamraju T S Priya working as a HOD, Assistant Professor in Master of Computer Applications at Sanketika Vidya Parishad Engineering college, Visakhapatnam , Andhra Pradesh with 13 years of experience in Master of Computer Applications (MCA), accredited by NAAC. Her area of intrest is in Computer Programming Using C, Computer Organisation, Software Engneering, Artificial Intelligence, Internet of Things (IoT) and Distributed Systems.

Adavipalli Pavan is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning Adavipalli Pavan has taken up his PG project on Gold Price Prediction Using Machine Learning and published the paper in connection to the project under the guidance of Pinnamraju T S Priya HOD, Assistant Professor SVPEC.

9. REFERENCES

1. ARIMA Model for Gold Price Prediction – International Journal of Scientific Research:
https://www.worldwidejournals.com/international-journal-of-scientific-research-(IJSR)/article/arima-model-for-forecasting-gold-prices/ODcwNg==/
2. Gold Price Forecasting using LSTM – Springer:
https://link.springer.com/chapter/10.1007/978-981-16-1053-5_8
3. Random Forest Regression for Financial Forecasting – IEEE:
https://ieeexplore.ieee.org/document/9562683
4. Hybrid ARIMA-LSTM Models for Financial Time Series – Elsevier:
https://www.sciencedirect.com/science/article/abs/pii/S2405452619301792
5. Box, G. E. P., Jenkins, G. M. – Time Series Analysis: Forecasting and Control (Wiley):
https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021
6. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory – Neural Computation:
https://www.bioinf.jku.at/publications/older/2604.pdf
7. Kaggle: Time Series Feature Engineering Tutorial:
https://www.kaggle.com/code/ryanholbrook/feature-engineering-for-time-series-data
8. TensorFlow Official Documentation:
https://www.tensorflow.org/guide
9. Scikit-Learn Official Documentation:
https://scikit-learn.org/stable/documentation.html
10. Pandas Official Documentation:
https://pandas.pydata.org/docs/
11. NumPy Official Documentation:
https://numpy.org/doc/
12. PyTorch Official Documentation:
https://pytorch.org/docs/stable/index.html
13. Brownlee, J. (2017). Introduction to Time Series Forecasting with Python:
https://machinelearningmastery.com/introduction-to-time-series-forecasting-with-python/

14. Hyndman, R. J., & Athanasopoulos, G. – Forecasting: Principles and Practice: https://otexts.com/fpp3/
15. , S., et al. – Forecasting: Methods and Applications (Wiley): https://www.wiley.com/en-in/Forecasting%3A+Methods+and+Applications%2C+3rd+Edition-p-9780471532330
16. Ensemble Methods in Machine Learning – Springer: https://link.springer.com/chapter/10.1007/3-540-45014-9_1
17. Gradient Boosting Machine – IEEE Paper: https://ieeexplore.ieee.org/document/7837800
18. Time Series Analysis with Python Cookbook – Packt: https://www.packtpub.com/product/time-series-analysis-with-python-cookbook/9781788624565
19. Data Cleaning and Preprocessing for Time Series – DataCamp: https://www.datacamp.com/tutorial/tutorial-time-series-data-preprocessing
20. Applied Predictive Modeling – Springer: https://www.springer.com/gp/book/9781461468486
21. Machine Learning for Financial Market Prediction – IEEE: https://ieeexplore.ieee.org/document/8994071
22. for Time Series Forecasting – Towards Data Science: https://towardsdatascience.com/time-series-forecasting-with-recurrent-neural-networks-74674e289816
23. Random Forest Algorithm for Regression – Towards Data Science: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
24. Gold Price Prediction using Machine Learning – ResearchGate: https://www.researchgate.net/publication/354557143_Gold_Price_Prediction_Using_Machine_Learning
25. Machine Learning for Time Series Forecasting – Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/07/time-series-forecasting-using-machine-learning/