A Model to Predict Daily Gold Price Using Machine Learning-Based Predictive Analysis Approach

A dissertation submitted for the Degree of Master of Business Analytics

R.W.U.S Rajapakse

University of Colombo School of Computing

2024

Declaration

Name of the student: R.W.U.S Rajapakse
Registration number: 2020/BA/030
Name of the Degree Programme: Masters of Business Analytics
Project/Thesis title: A Model to Predict Daily Gold Price Using Machine Learning-Based Predictive Analysis Approach

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.

3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.

4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.

5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.

6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	Date (DD/MM/YYYY)
Rodani	24/09/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	Supervisor 2	Supervisor 3
Name	Mr. Viraj Welgama		
Signature			
Date	24/09/2024		

I would like to dedicate this thesis to my family.

Acknowledgment

I would like to express my sincere gratitude to Mr. Viraj Welgama, my internal supervisor, for his invaluable guidance, assistance, and unwavering support throughout the entire research process. His extensive knowledge and wealth of experience served as a constant source of motivation.

I extend my heartfelt thanks to all the senior lecturers, lecturers, and instructors at UCSC for their guidance and the knowledge they generously shared with me.

I am particularly grateful to my family and friends for their constant love, support, and encouragement throughout the duration of this research.

Special thanks and appreciation are also due to my colleagues and individuals who willingly contributed their expertise to help me articulate my thoughts effectively.

Abstract

This thesis presents a comprehensive analysis of daily gold price prediction in the Sri Lankan market, utilizing a range of exogenous variables to enhance forecasting accuracy. The study explores the impact of key economic indicators, including Brent crude oil prices, USD to LKR exchange rates, CNY to LKR exchange rates, silver prices, S&P 20 index, Colombo Consumer Price Index (CCPI), and gold reserves on the volatility and trends of gold prices in Sri Lanka.

Two powerful machine learning algorithms, XGBoost and Random Forest, were employed to predict daily gold prices. The study investigates the predictive performance of these algorithms and compares their effectiveness in capturing the intricate dynamics of the Sri Lankan gold market. The models were trained and tested on a dataset spanning from January 2014 to September 2022 to ensure robustness and reliability in the results.

The findings reveal that XGBoost outperformed Random Forest in terms of predictive accuracy and model performance. The superiority of XGBoost suggests its efficacy in handling complex relationships and nonlinear patterns within the dataset, thereby providing more accurate and reliable predictions of daily gold prices in the Sri Lankan market.

Furthermore, the inclusion of diverse exogenous variables allows for a more holistic understanding of the factors influencing gold prices. The study contributes valuable insights to the financial community, policymakers, and investors.

Keywords: Gold Price, XGBoost, Radom Forest, Machine Learning

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADF – Augmented Dickey-Fuller

AI – Artificial Intelligence

ANN – Artificial Neural Network

CCPI – Colombo Consumer Price Index

CV – Cross Validation

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

ML – Machine Learning

MSE – Mean Squared Error

R^2 – Correlation Coefficient

RMSE – Root Mean Squared Error

S & P 20 – Standard and Poor's Sri Lanka 20

Chapter 1
Introduction

1.1 Problem Definition

The history of gold can be traced back thousands of years, and it has consistently held significant economic, cultural, and societal importance. Throughout human civilization, gold has been utilized as a medium of exchange, a store of value, and a symbol of wealth and power. From ancient civilizations like Egypt and Rome to modern economies, gold's allure has remained undiminished[1].

Over time, gold's value has fluctuated due to various reasons including economic conditions, geopolitical events, supply and demand dynamics, and changes in investor sentiment [2]. Traditionally, gold prices were influenced by the gold standard, which pegged currency values to specific amounts of gold. However, with the abandonment of the gold standard in the early 20th century, gold prices have become subject to market forces and speculative trading [3].

As financial markets became more complex and interconnected, the need for accurate and timely gold price predictions intensified. Investors, traders, financial institutions, and governments seek to understand and anticipate gold price movements to make informed decisions about asset allocation, risk management, and hedging strategies.

In the 21st century, the emergence of machine learning (ML) as a financial forecasting tool has opened new possibilities for understanding and predicting gold price movements. Advanced ML algorithms can analyze vast amounts of data, including historical price patterns, economic indicators, and sentiment analysis, to offer more accurate and timely predictions. This is particularly valuable for investors, who can use these insights to make informed decisions about when to buy or sell gold, or whether to include it in their portfolio for diversification and risk management. ML algorithms, such as decision trees, random forests, support vector machines, and neural networks, have shown promising results in predicting asset prices. A gold price prediction system using ML could significantly benefit market participants by providing insights into potential price trends and risks associated with gold investments [4].

1.2 Motivation

I. Predicting gold prices accurately can have practical implications in various industries, including finance, investment, jewelry, and mining.

II. Gold price data is usually readily available and accessible, making it easier to conduct research and build predictive models.

III. Increase employability with experience and exposure to machine learning projects.

IV. Interest in learning different machine learning techniques

V. Interest in the "Gold" field.

1.3 Research Aims and Objectives

1.3.1 Aims

I. To forecast future daily gold prices in Sri Lanka

II. To enable investors, traders, and other stakeholders to make informed decisions about buying and selling gold.

III. Identifying patterns and trends in gold price movements, which can be used to develop trading strategies or inform investment decisions.

1.3.2 Objectives

I. Identifying factors affecting the gold prices

II. Identifying patterns in the gold price movements

III. Identifying outliers in the gold prices

IV. Test and compare XGBoost and Random Forest Regressor algorithms with a set of variables and identify the best-suited ML model and the variables to predict gold prices.

V. Evaluate the output of XGBoost and Random Forest Regressor using different measures such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MSE), Mean Absolute Percentage Error (MAPE) and coefficient of determination (R^2).

1.4 Scope

This project primarily focuses on developing a model to predict the daily gold price in Sri Lanka. Given the vastness of this domain and the time limitation of one year, this project is confined to the Sri Lankan setting. The price prediction will be in a daily timeline based on the gold prices starting from 1st week of January 2014 to 4th week of September 2022. Part of the data will be used to train the model.

The price prediction would take into consideration any correlation of variables with gold price. This may include and is not strictly limited to silver price, CCIP, USD to LKR exchange rate, Brent crude oil price, and S&P Sri Lanka 20 index.

This project doesn't expect to explore traditional forecasting methods such as Box and Jenkins and exponential smoothing. Instead, this project will be using the ML model XGBoost and Random Forest Regressor.

As for the data visualization, this aspect falls outside of the scope of this project.

1.5 Contribution and Novelty

Most of the projects to predict the gold price in Sri Lanka are not ML based. There are only a very limited number of machine learning projects available to predict the gold price in the Sri Lankan context. To the best of our knowledge, this is the first one in Sri Lanka to use XGBoost to predict gold price in Sri Lankan context.

Moreover, the majority of existing gold price forecasting models operate on a monthly or quarterly basis, whereas this project stands out by predicting the daily gold price which is more useful for the buyers and sellers than the quarterly or monthly values.

This project incorporates central bank gold reserves as an additional variable to the initial list, aiming to analyze potential impacts associated with this inclusion.

1.6 Organization of the Study

This thesis is organized as follows.

Chapter 1 – Introduces the background, and problem definition with the novelty of the study, scope, motivation, aims, and objectives of the research.

Chapter 2 – Carries out a review of literature based on past studies related to forecasting gold prices using ML models in order to identify the factors that affect the gold price and the ML models that predicts the gold price well.

Chapter 3 – Describes the methodological background of the research study.

Chapter 4 – Gives a description on sources of data, method of data collection, variables and the formation of the data sets.

Chapter 5 – Provides the analysis and results of the study.

Chapter 6 – Discussion and conclusion about the results and suggestions for future work.

Chapter 2

Literature Review

2.1 Background

During the last few years, due to the emergence of machine learning techniques for predictions in different fields, the prediction of gold prices also has gained popularity in the industrial and scientific community.

Multiple research journals were analyzed prior to the start of the project to understand what types of methods were used to predict gold prices in different countries and what factors affect the gold price. This section only focuses on ML methods.

2.2 Machine Learning Models to Predict Gold Prices

A gold price prediction model using XGBoost in combination with the SHAP technique was presented by Jabeur, Salma, and Viviani. The ML model's output is interpreted using Shapely additive explanations (SHAP) to determine the main factors influencing gold prices. They examined six machine learning models in their project, namely, eXtreme Gradient Boosting, Random Forest, Light Gradient Boosting Machine, CatBoost method, and linear regression. They used the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) to evaluate the performance of the model. The results showed that XGBoost performed better than all the other models, producing the lowest RMSE (34.9), MSE, and MAE as well as the highest R^2 value (0.994), indicating its ability to predict gold prices accurately. Their predictor variables included the inflation rate of the US, USD/CNY exchange rate, USD/EUR exchange rate, silver price, crude oil price, S&P 500 index, and iron price. Among these, silver price emerged as the most important feature for gold price prediction, displaying a positive correlation with gold prices. Additionally, their findings revealed that gold prices have a positive correlation with crude oil, inflation, and the USD/CNY exchange rate while exhibiting a negative relationship with iron prices and the S&P 500 index. Monthly data was employed to predict monthly gold prices [5].

Restricted Boltzmann machines (RBM) were used for pre-training in Zhang and Bicong's deep belief network (DBN) model, which was then fine-tuned by a supervised back-propagation (BP) layer. They constructed a multi-layered neural network with the ideal 5-10-1 network topology. As input variables, they employed the Federal funds rate, the Dow Jones Index, the crude oil price, the U.S. CPI, and the effective exchange rate. They found that the price of gold is positively correlated with the U.S. Consumer Price Index (CPI), the price of crude oil, and the Dow Jones Index. The gold price was negatively correlated with the Federal funds rate and the effective exchange rate. This study was carried out with the same objective as the previously described one, which was to forecast the monthly volatility of gold prices. They compared the DBN model's results with those from a hyper model of GA-BPNN, a standard NN model of BP, and ARIMA in order to evaluate the accuracy of the model. According to the comparison, the suggested DBN model outperforms the other models, having the lowest MAE, RMSE, and MAPE [6].

Liu and Li used random forest to create a machine-learning model. As initial input variables, they used the US dollar index (USDX), crude oil price (COP), Dow Jones Industrial Average (DJIA), US CPI (USCPI), US ten-year bond futures prices (US10BFP), Hang Seng Index (HIS), and Standard & Poor's 500 Index (S&P500) and measured the performance of the model. They selected the Standard & Poor's 500 Index (S&P500) and the Dow Jones Industrial Average (DJIA) as the final model input variables using a stepwise backward variable selection procedure. The Standard & Poor's 500 Index (S&P500) and the Dow Jones Industrial Average (DJIA) displayed prediction performances of 0.9962 and 0.9999, respectively. Another goal of this model's construction was to predict the monthly gold price [7].

Makala and Z Li compared the outcomes of their use of ARIMA and SVM models for gold price prediction. They compared SVM (Linear), SVM (Poly), Arima (2, 1, 2) (2, 1, 2, 12), and SVM (RBF). The outcomes demonstrated that SVM (Poly) outperforms both the Arima model and other SVM models. SAV resulted in an RMSE of 0.028 and MAPE of 2.5 and 36.18 and 2897 for ARIMA respectively. They used gold price data for this study research [8].

In order to predict gold prices, Sami and Khurum contrasted artificial neural network (ANN) models with linear regression. Their main goal was to forecast the price of gold by utilizing the most extensive list of attributes. The input variables they used were the S&P 500 Index, US Bond Rates, Silver Spot Rates, Platinum Spot Rates, Palladium Spot Rates, Rhodium Spot Rates, Oil Spot Prices, NYSE Index, EGO Index, SLW Index, AU Index, ABX, BVN, China Interest Rate, USA Interest Rate, UK Interest Rate, and Russia Interest Rate. The unique aspect of this study is that, among these eighteen factors, some have never been used before, such as performance metrics for China, India, and Russia, which are the top three buyers of gold. Correlation analysis was used to determine which factors had the greatest impact on the price of gold. They discovered an interesting finding, the stock price of Silver Wheaton Corporation (SLW), the largest precious metals streaming company in the world, has the highest correlation with gold rates, not the performance of the US economy (or any other major economy), nor the price of other precious metals. Although ANN outperformed linear regression, the difference was not statistically significant [9].

Mustafa Yurtsever employed various indicators including the effective federal funds rate, consumer price index, S&P 500 stock market index, crude oil price, and effective exchange rate to forecast monthly gold prices. His investigation indicated that all the mentioned factors exert an influence on gold prices. The predictions were based on monthly data, utilizing three different models: LSTM, Bi-LSTM, and GRU, for time series forecasting of gold prices. Model performance was assessed using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The outcomes demonstrated that LSTM outperformed the other models, giving the most favorable results with a 3.48 MAPE, 61,728 RMSE, and 48.85 MAE values [10].

In their research on forecasting gold prices, Manjula and Karthikeyan incorporated variables such as stock market performance, crude oil prices, rupee-dollar exchange rates, inflation, and interest rates. They employed three machine-learning algorithms—gradient-boosting regression, random forest regression, and linear regression. Their analysis, based on monthly gold price data spanning from January 2000 to December 2018, revealed that all three models consistently achieved R2 values exceeding 95%, indicating a notably high level of accuracy in predicting future monthly gold prices [11].

Livieris and Pintelas employed sophisticated deep learning methods to forecast both the price and movement of gold. Their model consisted of two main components. The first layer is the convolutional and pooling layer, which performs complex mathematical operations to develop features of the input data. The second component integrated LSTM and dense layers, leveraging the generated features to make predictions. This model demonstrated the capability to forecast not only the next day's gold price but also the direction of its movement (increase or decrease). The research utilized daily gold price data in USD spanning from January 2014 to April 2018. The models exhibited an accuracy level of 55% [12].

Sandeep Patalay proposed a system based on the M5P model tree machine learning algorithm to predict gold prices. He used S&P 500 and crude oil prices as input variables. The results showed an accuracy of 85% [13].

Zakaria *et al*. proposed a novel model to forecast gold prices using a recent meta-heuristic method named whale optimization algorithm (WOA) as a trainer to learn the multilayer perceptron neural network (NN). They used the Indian rupee INR, South African rand (ZAR), and Chinese yuan (RMB) exchange rates, inflation rates of China and the US, crude oil prices, copper prices, silver prices, and iron prices as input variables to the model. They used a correlation matrix to identify the relationship between gold price and input variables. The output of the correlation matrix showed that copper price, iron price, silver price, oil price, South African exchange rate, and Indian exchange rate have a positive linear relationship with the gold price. Among them, the silver price has the strongest positive relationship. China's inflation, China's exchange rate, and the US inflation rate had weak linear relationships with gold prices. They compared their proposed model's (WAO-NN) results against five other models, namely classic NN, particle swarm optimization for NN (PSO–NN), genetic algorithm for NN (GA–NN), and grey wolf optimization for NN (GWO–NN) and ARIMA models. They used the RMSE, MSE, STD, and R^2 to evaluate each model. The results depicted that the WOA–NN model provides the highest out of sample R^2 0.9989 coupled with the lowest MSE, RMSE, and STD of 0.00047, 0.02131, and 0.00340, respectively. Overall, each model's R^2 value was above 95% [14].

2.3 Machine Learning Models to Predict Prices of Other Metals

Dylan, Herlina and Firdaniza conducted a study to forecast silver prices using XGBoost model. They used gold prices, platinum prices, and the dollar-to-euro exchange rate as predictor variables. Their dataset comprised a daily timeframe. They have analyzed the performance of three models, namely, XGBoost, Random Forest, and CatBoost. The results showed that XGBoost has the highest performance with MAPE of 5.98% and an RMSE of 1.6998, concluding that XGBoost is a good model to predict silver prices [15].

An ML model was suggested using Support Vector Regression to predict copper price by Gabriel, Carrasco, Christian, *et al*. Their predictions were for the 5, 10, 15, 20 and 30 days intervals from the copper closing price at the London Metal Exchange. The error of 2.2% in predicting copper prices for the intervals 5 and 10 days depicted the good performance of the model [16].

Christian and Marian conducted a study to predict the returns of four precious metals (gold, silver, platinum and palladium) using Random Forest. They evaluated the study in two criteria, univariate and multivariate forecast evaluation criteria. They utilized weekly returns from various financial indicators including the S&P-500 index, the GSCI commodity-price index, the Chicago Board Options Exchange Volatility Index (VIX), the term spread (the difference between the 10 year treasury constant maturity rate and the 3 month treasury constant maturity rate), the corporate bond (CB) spread (the difference between Moody's Baa rate and Moody’s Aaa rate), as well as weekly returns of the dollar against the yen, euro, British pound, Canadian dollar, Australian dollar, and an aggregate trade-weighted exchange-rate index (TWEXB) as predictor variables. Their findings concluded that multivariate forecasts using Random Forest on precious metals are much accurate than univariate forecasts [17].

2.4 Summary of Related Work

According to the above literature, below are the most successful ML algorithms in forecasting gold prices

- XGBoost
- DBN
- Random Forest
- SVM
- ANN
- LSTM

The most common variables found from the literature that affects the gold prices are below.

- Crude Oil Price
- Silver Price
- Exchange Rates
- Inflation
- S&P 500

This project uses XGBoost and Random Forest Regressor for forecasting daily gold price in Sri Lanka as the literature shows that XGBoost and Random Forests show higher performance.

In this study, the below variables are selected as predictor variables based on the insights from the existing literature.

- Sri Lankan Inflation (Used CCPI)
- LKR/CNY Exchange Rate
- LKR/USD Exchange Rate
- Silver Price (LKR)
- Crude Oil Price (USD)
- S&P 20

In addition to the variables drawn from the existing literature, this study incorporates central bank gold reserves to the initial variable list intending to explore and assess the potential impact of these reserves on the gold prices in Sri Lanka.

Chapter 3
Methodology and Solution Design

This section focuses on explaining the systematic methods followed to meet the objectives mentioned in the previous section. The first part explains the steps followed in the process of building the ML model and other analyses. The second part elaborates on the theory and methodology behind time series, machine learning techniques, and regression tasks.

3.1 Data Collection

Based on the literature presented in section 2, the independent variables considered to impact gold prices are the USD/LKR exchange rate, CNY/LKR exchange rate, S&P 20 value, silver price, CCPI, and the per-barrel price of Brent crude oil. Additionally, new variable, gold reserves was introduced in this study. To ensure data accuracy, all information, including the dependent variable, was meticulously gathered from the sources mentioned in section 4. The data were collected through separate CSV files downloaded from their respective websites, for the period from January 2014 to September 2022.

The data were predominantly available in a daily format, with the exception of the CCPI and gold reserves held by the central bank. The CCPI is provided on a monthly basis, and for analytical purposes, it was presumed that the CCPI value remained constant throughout the entire month. Similarly, as gold reserves tend to change infrequently, it was assumed that the quarterly figure was representative of the entire quarter.

3.2 Data Pre-processing

Data preprocessing involves converting the data into a format that may be used to perform data mining, machine learning, and other data science operations more quickly and efficiently.

The CCPI serves as a valuable indicator, presenting the monthly average change in the prices of goods. In order to maintain consistency and ensure the accuracy of our machine learning model, it is essential to have all variables in daily format when feeding data. Therefore, for the monthly figure of CCPI, we assume it to remain constant throughout the entire month to align with our daily data requirements.

The new addition to the variables is the gold reserves in the central bank, which are available in quarterly figures. Given that gold reserves typically don't change on a daily basis and the project requires daily data, we assume that the quarterly figure remains constant throughout all the days of that quarter.

Firstly, the data in the separate CSV files were collated into one master CSV file based on the date. There were fifteen missing values in the S&P 20 value column, one missing value in silver price column and four missing values in the Brent crude oil price column. The missing values were filled using the forward fill method [18],[19].

A sample of the dataset after pre-processing is given in table 1

Table 1: Sample of the Dataset After Preprocessing

Date	Gold Price	USD LKR Exchange rate	CNY LKR Exchange rate	Brent Crude oil price USD	Silver Price	CCIP	S&P 20	Gold Reserves
1/1/2014	157671.43	130.75	21.60	110.92	2548.54	104.20	3263.87	894.28
1/2/2014	158636.78	130.73	21.61	107.78	2614.96	104.20	3285.39	894.28
1/3/2014	160947.64	130.70	21.60	106.89	2634.54	104.20	3294.82	894.28
1/6/2014	162422.89	130.75	21.60	106.73	2636.19	104.20	3281.24	894.28
1/7/2014	162412.06	130.70	21.60	107.35	2596.19	104.20	3300.45	894.28
1/8/2014	160678.07	130.85	21.62	107.15	2555.57	104.20	3350.31	894.28
1/9/2014	160417.09	130.70	21.59	106.39	2555.04	104.20	3362.39	894.28
1/10/2014	160949.31	130.70	21.60	107.25	2633.63	104.20	3384.51	894.28
1/13/2014	163626.96	130.70	21.63	106.75	2667.79	104.20	3388.07	894.28
1/16/2014	162141.33	130.68	21.58	107.09	2627.90	104.20	3427.05	894.28
1/17/2014	162518.52	130.70	21.60	106.48	2653.01	104.20	3407.94	894.28
1/20/2014	164318.56	130.74	21.60	106.35	2657.30	104.20	3425.71	894.28
1/21/2014	163813.21	130.80	21.62	106.73	2599.19	104.20	3428.16	894.28
1/22/2014	162489.97	130.80	21.62	108.27	2591.40	104.20	3458.26	894.28

3.3Feature Engineering

Checking for stationarity before feature engineering is important because the XGBoost algorithm is not constructed to identify the quirks patterns in time series data [20] The time series graphs of each variable were plotted to identify the non-stationary series. The time series graphs for gold price, USD LKR exchange rate and CNY LKR exchange rate are shown below.

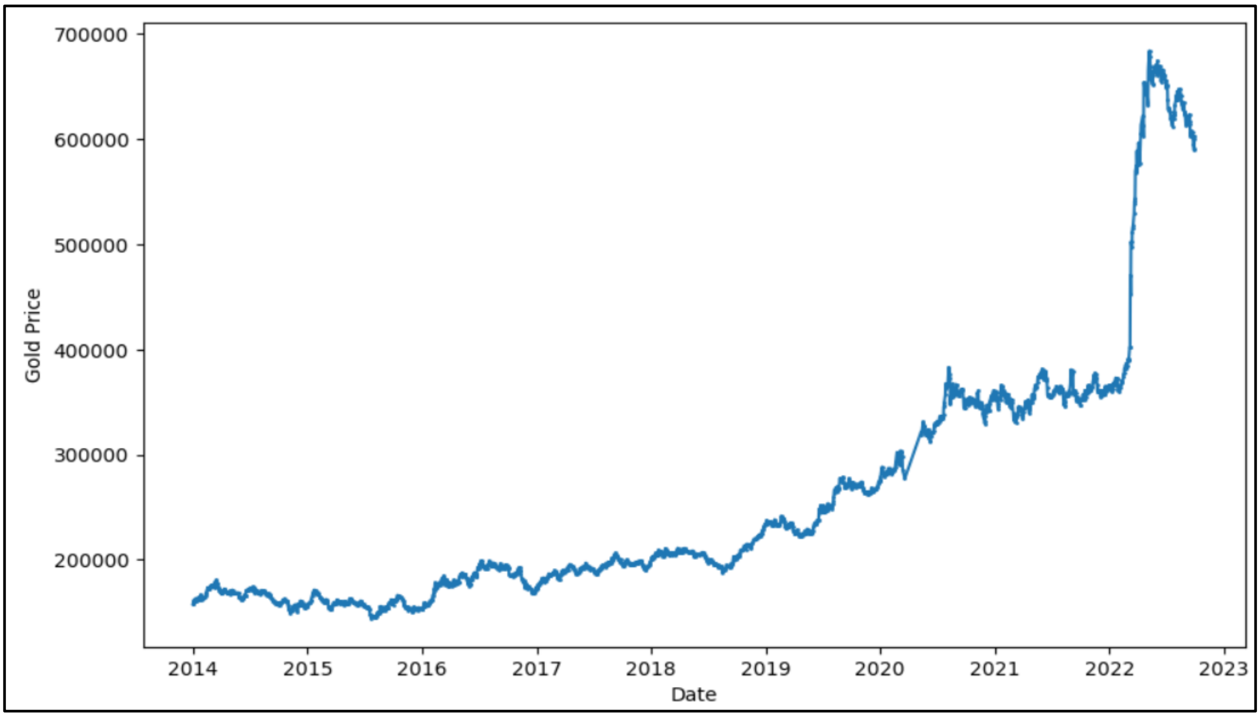


Figure 1: Gold Price Time Series Graph

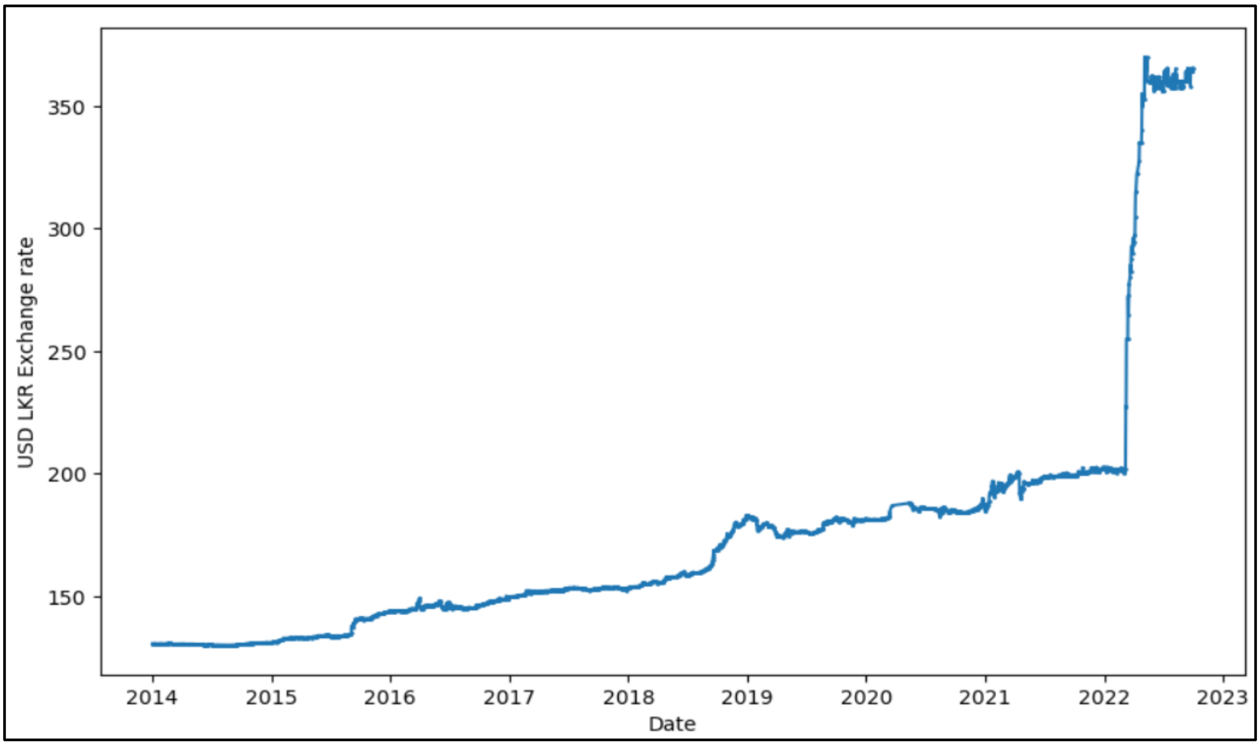


Figure 2: USD LKR Exchange Rate Time Series Graph

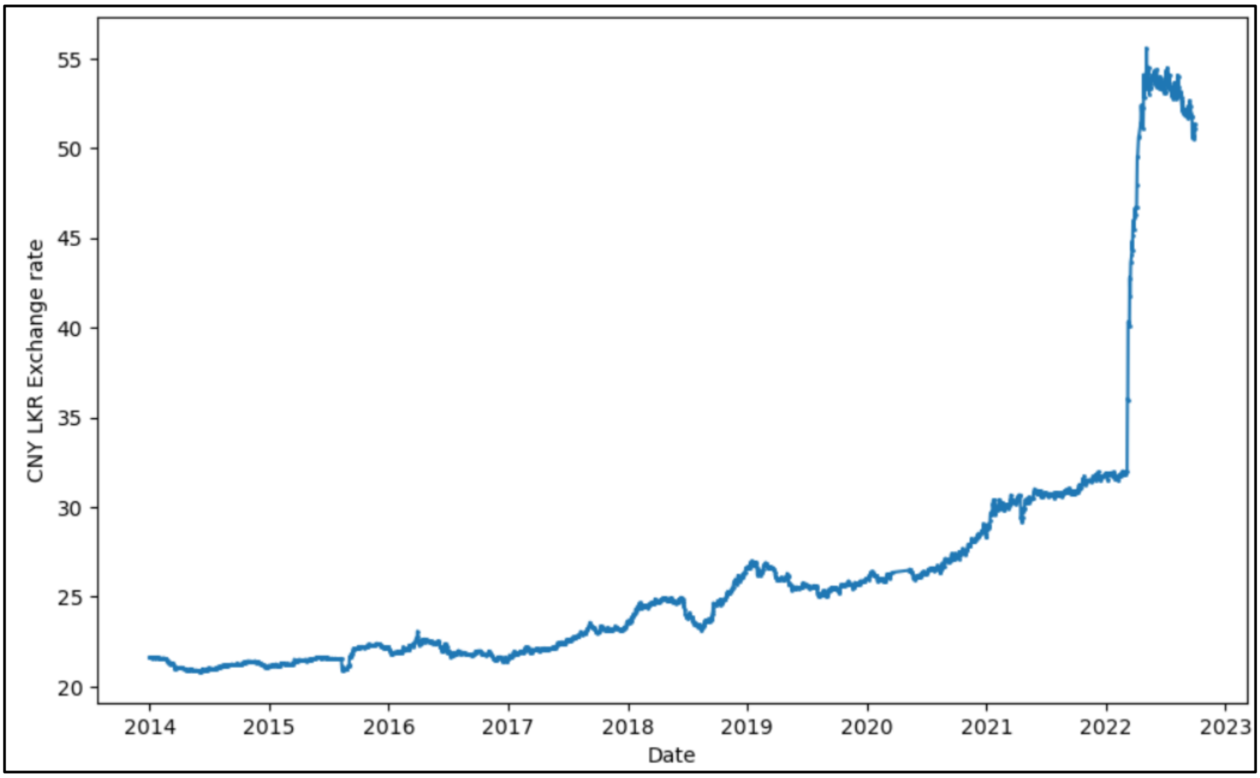


Figure 3: CNY LKR Exchange Rate Time Series Graph

It is evident from the above graphs that none of the series are stationary. Further, the Augmented Dickey-Fuller test was carried out to validate the findings from the time series graphs. Below are the test statistics for gold price and USD LKR exchange rate respectively.

```
ADF Statistic: 0.0961479419615481
p-value: 0.96579711001765
Critical Values: {'1%': -3.4335423467495283, '5%': -2.862950226171245,
'10%': -2.5675207408950835}
ADF Statistic: 0.28414503347721964
p-value: 0.9766052616123226
Critical Values: {'1%': -3.4335423467495283, '5%': -2.862950226171245,
'10%': -2.5675207408950835}
```

Given that all variables exhibited non-stationary behavior, it was deemed necessary to address this issue by creating first-differenced variables for each variable in the analysis. This adjustment was implemented to enhance the stationarity of the data and facilitate a more robust analytical framework. The Augmented Dickey-Fuller test showed that the series were

Subsequently, lag variables were introduced for all differenced data. Prior to the creation of these lag variables, Cross-Correlation Function (CCF) plots were generated between the dependent and independent variables, aiming to pinpoint the significant lags. Additionally, Partial Auto Correlation Function (PACF) plots were employed to identify the significant lags associated with the dependent variable. This comprehensive approach was adopted to ensure the identification and incorporation of relevant time lags in the subsequent analysis. The plots were created for the pre-whitened data. The below figure shows the CCF plot and the PACF plot.

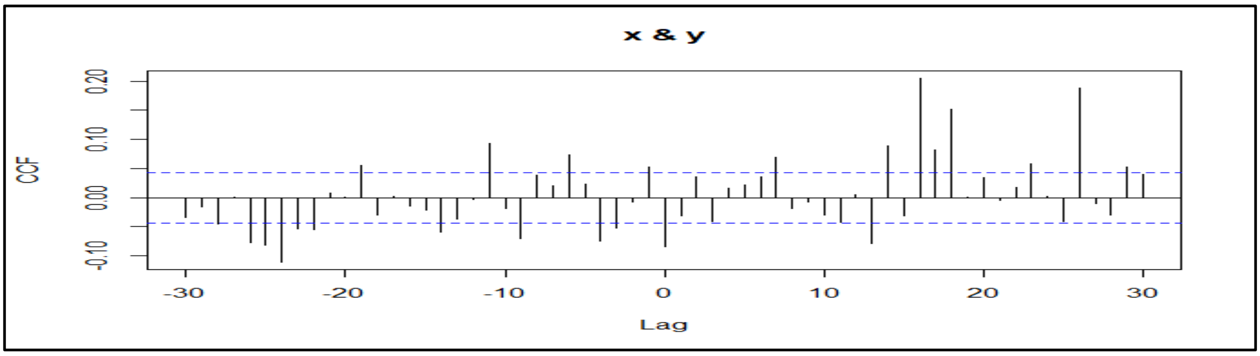


Figure 4: CCPI CCF Plot

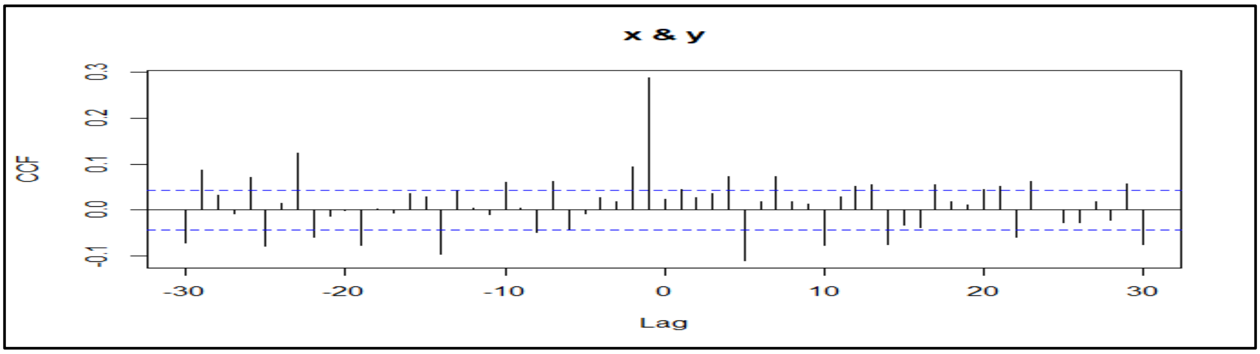


Figure 5: CNY LKR Exchange Rate CCF Plot

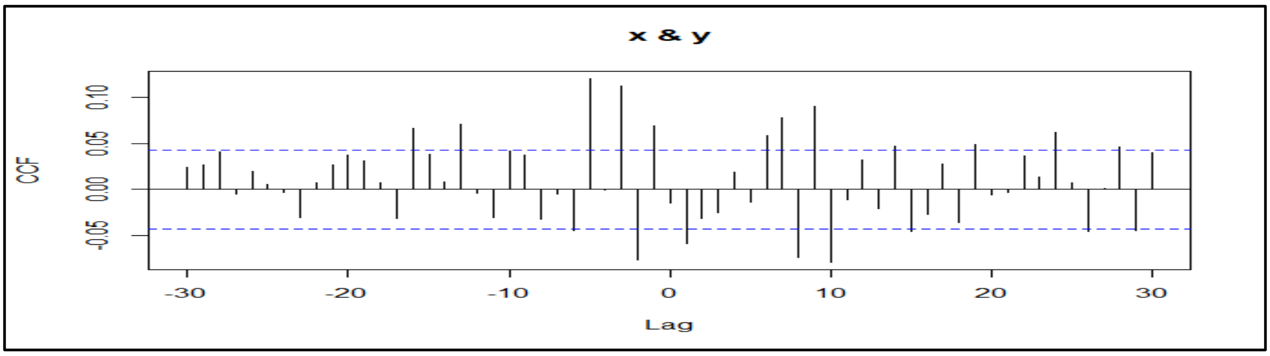


Figure 6: Crude Oil Price CCF Plot

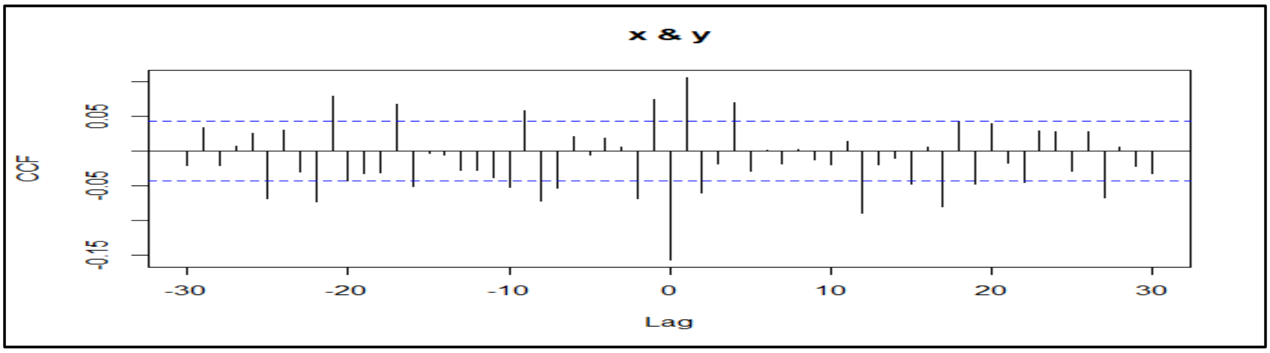


Figure 7: S&P 20 Index CCF Plot

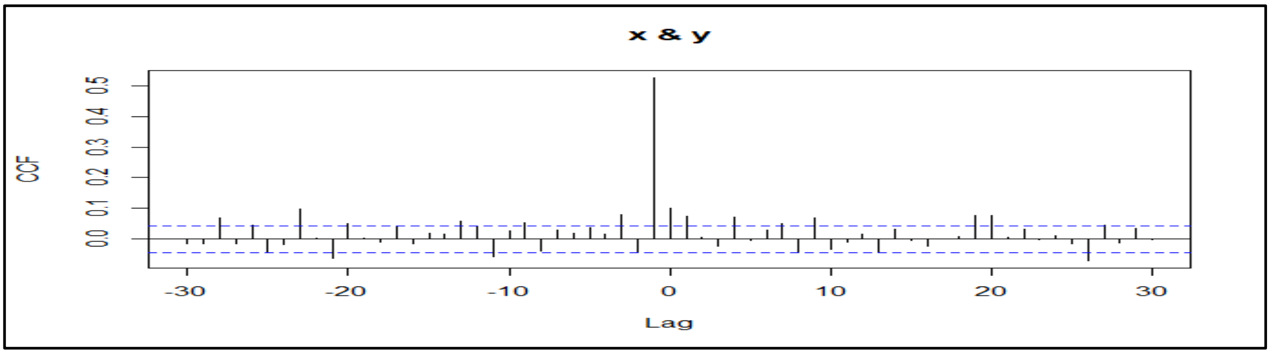


Figure 8: Silver Price CCF Plot

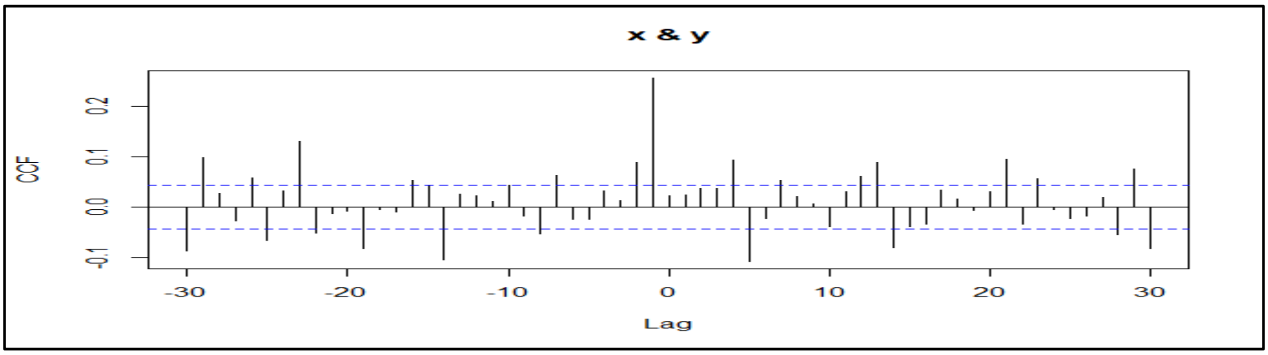


Figure 9: USD LKR Exchange Rate CCF Plot

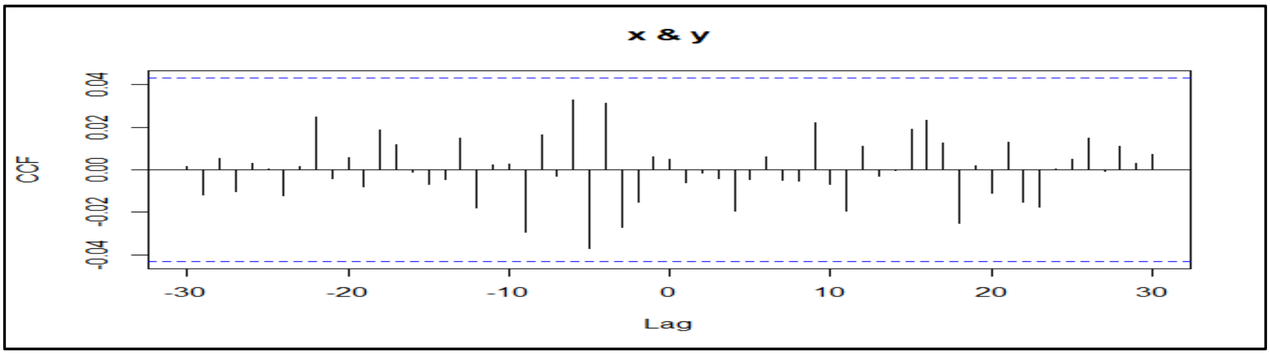


Figure 10: Gold Reserves CCF Plot

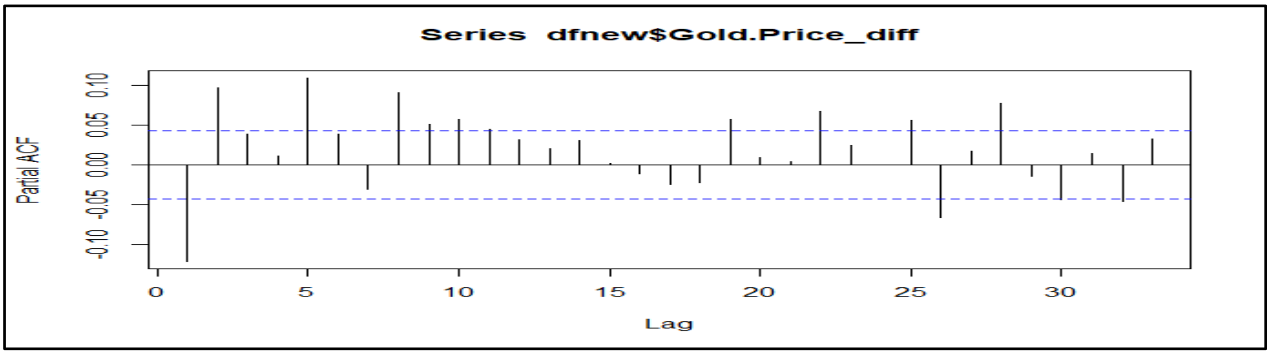


Figure 11: Gold Price PACF Plot

Based on the above CCF plots and PACF plot, the below lags of each variable were selected.

Table 2: Significant Lags of each Variable Selected for the Model

Variable	**Lags**
CCIP	1,3,4
CNY LKR Exchange rate	1,2,7,14
Crude Oil Price	1,2,3
S&P 20	1,2,7,21
Silver Price	1
USD LKR Exchange rate	1,2,7,14
Gold Price	1,2

As the CCF plot for gold reserves revealed the absence of significant lags, gold reserves were excluded from further analysis.

Following the identification of significant lags in both the dependent and independent variables, the subsequent analysis and model training involved utilizing the differenced data along with the identified lagged variables.

3.4 Check for Multicollinearity

Following the finalization of the dataset, which involved differencing and selecting significant lags, an examination of multicollinearity among predictor variables was conducted using a correlation matrix.

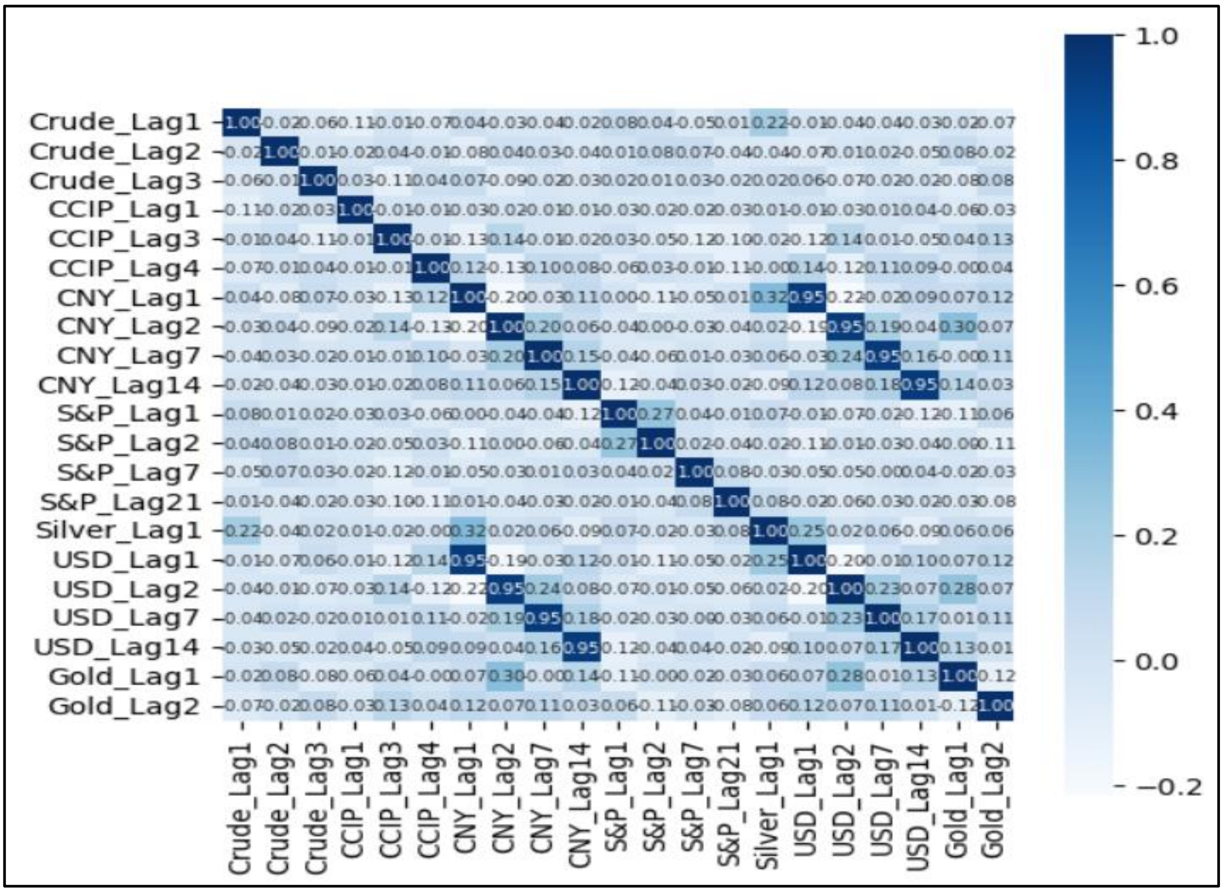


Figure 12: Correlation Matrix of the Predictor Variables

The correlation matrix suggested a multicollinearity between USD LKR exchange rate and CNY LKR exchange rate, hence regularization parameters in XGBoost alpha (L1 regularization) and lambda (L2 regularization) were tuned to minimize the multicollinearity and to prevent overfitting and improve the generalization of the model.

3.5 Training and Test Data Split

The data for all the variables for the period of January 2014 to September 2022 were used for this study, of which the first 80% of data were used to train the model, and the rest of the 20% was used to test the model.

3.6 Systematic Approach

3.6.1 Machine Learning

Machine Learning falls under the category of artificial intelligence (AI) and is dedicated to developing computer algorithms capable of self-improvement through experiential learning and data utilization. In simpler terms, ML empowers computers to acquire knowledge from data and autonomously make decisions or predictions without explicit programming.

At its essence, ML revolves around the creation and deployment of algorithms to support these decision-making processes and predictions. These algorithms are engineered to enhance their performance over time, progressively gaining accuracy and effectiveness as they process more data.

In contrast to traditional programming where computers adhere to predefined instructions for task execution, machine learning involves providing the computer with a dataset of examples and a task to accomplish, allowing the computer to determine how to achieve the task based on the provided examples [21].

There are three types of ML models.

I. Supervised Machine Learning
II. Unsupervised Machine Learning
III. Reinforcement Machine Learning

Supervised Machine Learning:

This kind of ML involves training the algorithm on a dataset with labels. By using the labeled training data, it gains the ability to map input features to targets. The algorithm learns to generalize from known data to make predictions on new, unseen data in supervised learning, where it is given input features alongside the matching output labels.

Two primary categories of supervised learning:

Regression:

An algorithm learns to predict continuous values based on input data through regression, a sort of supervised learning. Regression analysis uses continuous data, such as stock and home prices, as the output labels. In ML, there are various regression methods such as Random Forest, Polynomial, Ridge, Decision Tree, Support Vector, and Linear regression, among others.

Classification:

Classification is a sort of supervised learning in which the algorithm learns to categorize or classify incoming data based on input features. In classification, the output labels are discrete values. Binary classification methods yield an output that can belong to one of two potential classes, while multiclass algorithms yield an output that can belong to multiple classes. Various ML classification algorithms include Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Decision Trees.

Unsupervised Machine Learning:

In unsupervised learning, an algorithm picks up the ability to spot patterns in data without needing to be explicitly trained on labeled instances. Finding the data's underlying structure or distribution is the aim of unsupervised learning.

Reinforcement Learning:

By taking actions and getting rewarded or penalized for them, an agent learns how to interact with its environment through a sort of ML called reinforcement learning. In reinforcement learning, the aim is to build a strategy that maps states to actions in a way that optimizes the total expected reward as time progresses.

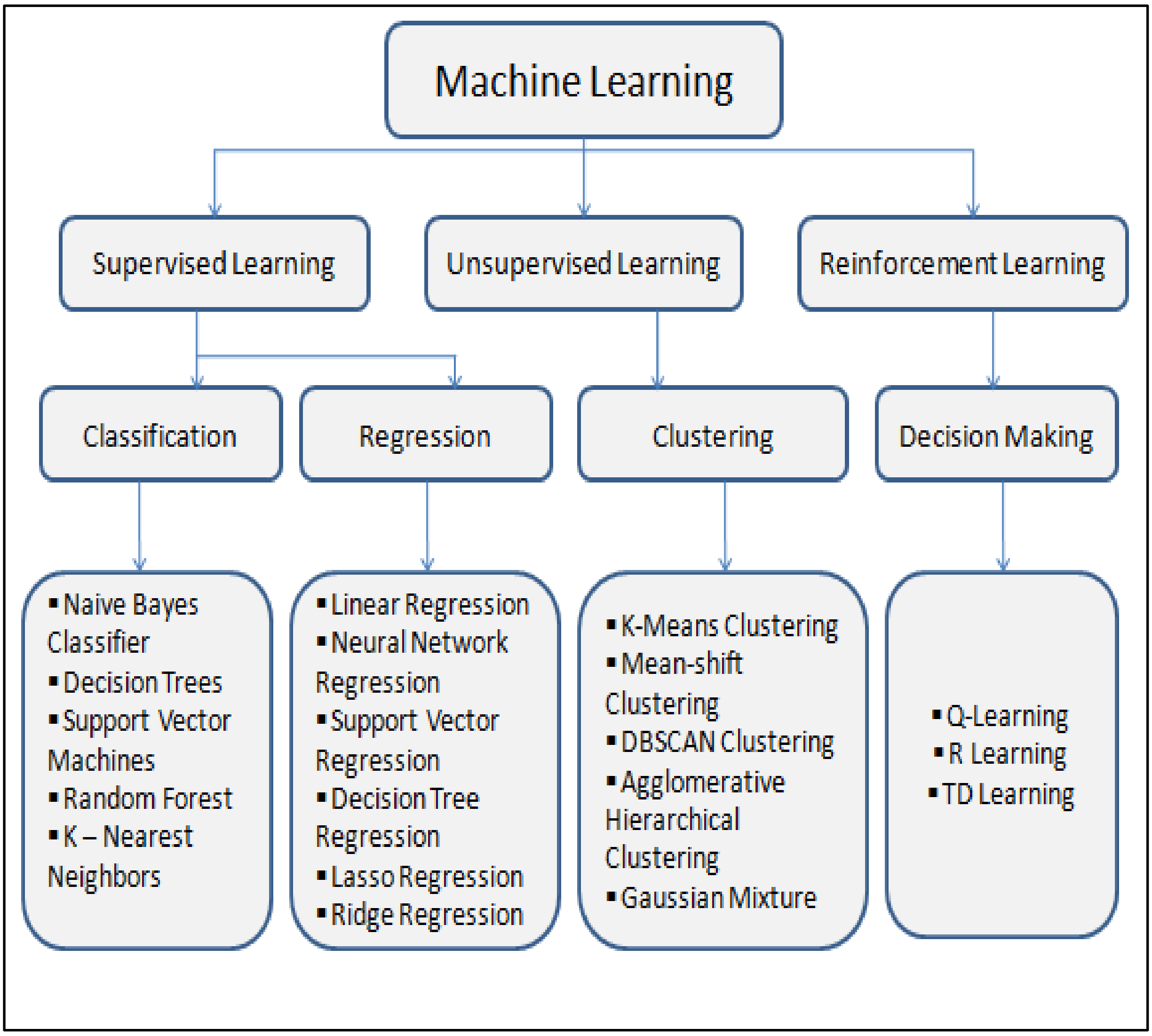


Figure 13: Types of Machine Learning

The research study outlined in this thesis addresses regression. Its primary objective is to identify the complex relationship between the daily gold price and various independent variables, ultimately enabling the prediction of future gold prices.

Based on the literature review in Chapter 2, XGBoost and Random Forest were chosen for this study.

3.6.2 XGBoost

XGBoost originated as a research project within the Distributed (Deep) Machine Learning Community (DMLC) group, led by Tianqi Chen. XGBoost, a highly optimized distributed gradient boosting toolkit which enables the efficient and scalable training of ML models. By combining the predictions of several weak models, this ensemble learning technique generates more robust predictions. Because of its capacity to manage large datasets and its ability to achieve state-of-the-art performance in numerous ML tasks, including regression and classification, XGBoost, which stands for "Extreme Gradient Boosting," has emerged as one of the most liked and extensively applied machine learning algorithms.

In this approach, decision trees are generated sequentially. One crucial component of XGBoost is weights. Each independent variable is assigned a weight before being fed into the decision tree to make predictions about the outcome. The second decision tree gives greater weight to the factors that the first decision tree mis predicted. [22].

These are the procedures used in XGBoost,

I. Assign equal weights to each observation in the chosen training set.
II. After that, use equally weighted data to apply the first base learner.
III. A higher weight will be given to that specific observation if the algorithm results in any prediction errors.
IV. On the weighted observations, apply the second base leaner.
V. Until a greater accuracy is attained or the number of base learning algorithms is exhausted, this process will be repeated.

The following figure illustrates how a strong learning algorithm is produced by combining weak learners.

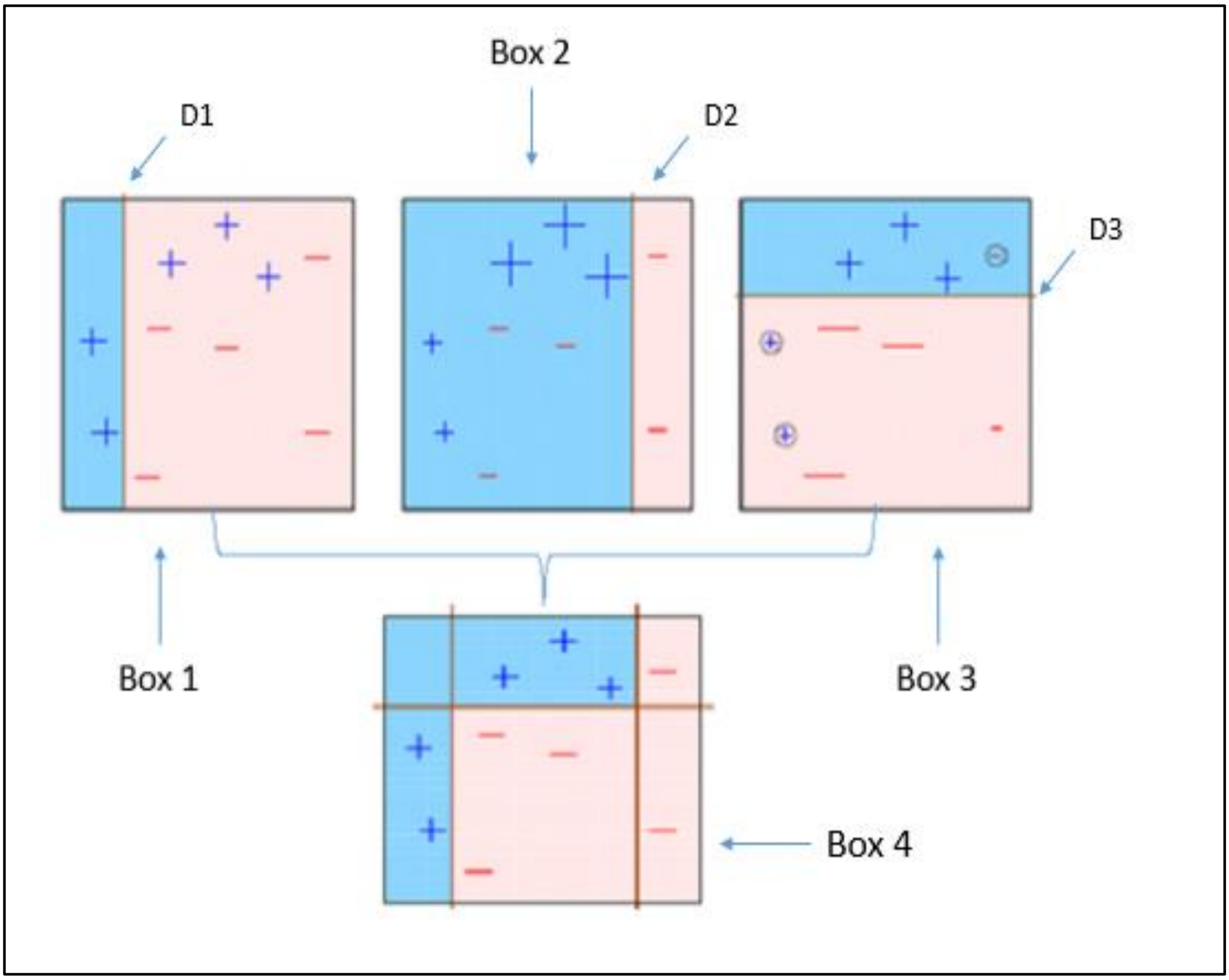


Figure 14: How to Create a Powerful Learner in XGBoost

Box 4 (Combination of Weak Learners) accurately classifies the observations, but Box 1, Box 2, and Box 3 (Weak Learners) incorrectly classify the same. This is seen in the image. Regularization is a powerful tool that XGBoost uses to punish complex models and prevent overfitting. One of the special features of XGBoost is its ability to handle various data sparsity patterns. This allows the data layout to be reused by successive iterations without having to be computed afresh, in contrast to other algorithms. XGBoost was designed to make the best possible use of the available hardware.

The below parameters of XGBoost were tuned in this study.

Parameters

I. max_depth – Maximum depth of a tree

II. subsample - Subsample ratio of the training instances

III. colsample_bytree – a family of parameters for subsampling of columns

IV. objective - Specify the learning task and the corresponding learning objective

V. n_estimators – Number of trees need to build

VI. learning_rate - Step size for updating the model's parameters during each boosting iteration

VII. Alpha - L1 regularization term in the objective function

VIII. Lambda - L2 regularization term in the objective function

3.6.3 Random Forest Regressor

Random Forest Regression is an ensemble method proficient in handling both regression and classification assignments. It accomplishes this by using multiple decision trees and a method known as Bootstrap and Aggregation, or bagging. The fundamental concept is to aggregate the outputs of multiple decision trees to arrive at the final result, as opposed to relying on individual trees.

The Random Forest comprises numerous decision trees as its foundational learning models. Through a process known as Bootstrap, we randomly select subsets of rows and features from the dataset to create sample datasets for each model [23].

The maximum number of features available for splitting at each node is constrained to a specified percentage of the total features (defined as a hyperparameter). This constraint is in place to prevent the ensemble model from excessively relying on any single feature and encourages a balanced utilization of all potentially predictive features.

Moreover, in the process of generating splits, each tree selects a random sample from the original dataset, introducing an additional layer of randomness. This randomness serves as a preventive measure against overfitting, contributing to the robustness of the ensemble model [24].

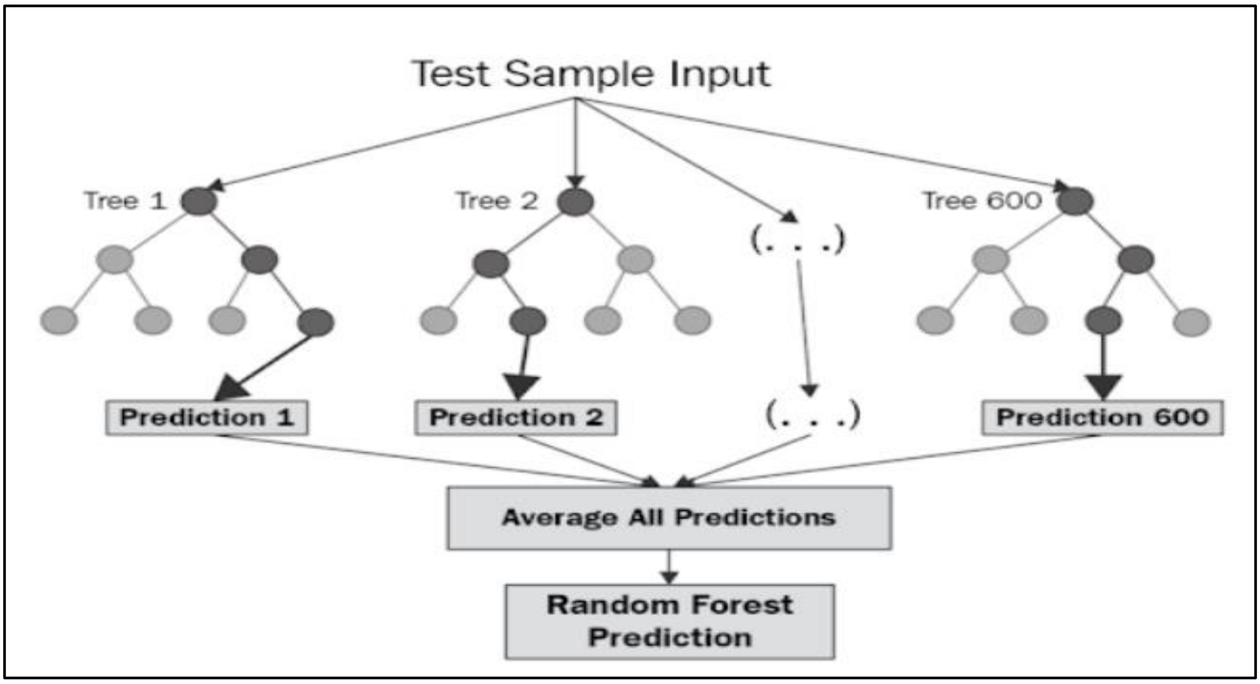


Figure 15: Mechanism of Random Forest Regressor

The below parameters of the Random Forest were tuned in this study.

Parameters

I. max_depth – The maximum length of the path from the root node to a leaf node.
II. min_samples_leaf – Significance attributed to the minimum size of samples in a leaf node.
III. min_samples_split – The minimum required number of observations in any given node to split
IV. n_estimators – Number of trees that need to be built

3.6.4 GridSearchCV

In machine learning, a method called GridSearchCV (Grid Search Cross Validation) is used to check and identify the best set of hyperparameters for a certain model. By methodically examining a predetermined range of hyperparameter values, it generates a "grid" of potential configurations. After that, it uses cross validation to assess each combination and chooses the one that produces the best results. GridSearchCV improves the model performance, eliminates manual trial-and-error, and automates the hyperparameter tuning process [25].

Hyperparameters:

Hyperparameters in ML are parameters that must be pre-determined before a model can be trained and are not learned from the data. Examples include the learning rate in a gradient descent algorithm, the depth of a decision tree, or the number of neighbors in a k-nearest neighbors classifier.

Cross-Validation in time series:

Cross-validation in time series analysis is used to identify the most effective set of hyperparameter values that reduce the risk of overfitting in the time series model. In contrast to traditional cross-validation techniques, time series cross-validation is capable of handling the temporal structure of the dataset. This method involves dividing the dataset into n number of consecutive folds, in a way that each fold represents a similar time period. The model is then trained on earlier folds and evaluated on subsequent folds, simulating the real-world scenario where the model must make predictions on unseen future data. Since the data is not randomly shuffled, this technique prevents the inclusion of data from the future in the training set. This mitigates data leakage and provides a better assessment of the model's performance over time. Time series cross-validation helps in identifying hyperparameter configurations that result in optimal model performance across different time periods.

Expanding window and rolling origin are the two most common time series cross-validation techniques. These methods involve iteratively moving the training and validation windows through the dataset, allowing the model to adapt to evolving patterns and trends in the time

The following figure illustrates the process of time series cross-validation.

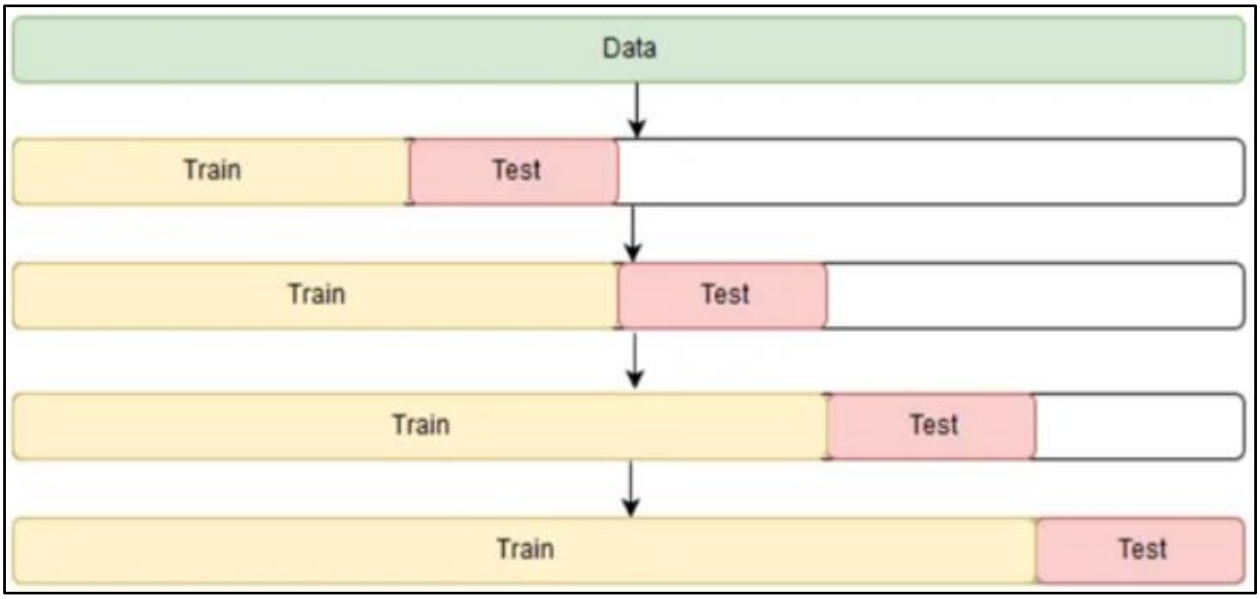


Figure 16: Cross Validation Process in Time Series

In this study, a comprehensive and reliable method for XGBoost and Random Forest model hyperparameter tuning was used. The GridSearchCV method was combined with the effect of five-fold cross-validation to methodically investigate and determine the ideal hyperparameters for XGBoost and Random Forest. It is important to note that in this study, we used TimeSeriesSplit() function to split the dataset for five folds. This methodology aimed to ensure the reliability and generalizability of the model's performance by rigorously evaluating various hyperparameter combinations. Various combinations for the hyperparameters of the XGBoost and Random Forest models mentioned in sections 3.6.2 and 3.6.3 were evaluated using GridSearchCV and the best hyperparameter set was chosen.

3.6.5 Feature Selection

A key component of all ML workflows is feature engineering, which includes feature selection as well as feature extraction. Despite the fact that they are sometimes confused, these are the essential elements of contemporary ML pipelines. Utilizing domain expertise, feature extraction creates additional variables from unprocessed data to enhance machine learning algorithms' performance. The goal of feature selection, on the other hand, is to identify the most reliable, pertinent, and unique features.

The following are the main objectives of feature selection in ML:

I. Simplifying models to make them more understandable for researchers and users.
II. Reducing training times.
III. Avoiding the challenges associated with high-dimensional data
IV. Enhancing the model's generalization by mitigating overfitting, which formally reduces variance.

Initially, both the XGBoost and Random Forest Regressor models were trained using differenced data containing lag variables, and their performances were evaluated. Subsequently, the most significant features for each model were identified using the feature importance functionality in both XGBoost and Random Forest Regressor. The models were then retrained using only the selected important features, and their performance was evaluated once again.

3.7 Comparison of the model output

Given that the models were trained using differenced data, the resulting predictions are also in differenced form. To assess the models, an inverse difference of the model's outputs was performed before evaluation, ensuring predictions are presented in the original scale.

The following error measures were used based on the literature to compare the models implemented for forecasting the monthly daily gold price.

Mean Square Error (MSE)

The average squared difference between the actual and the predicted value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i = Actual value

$\hat{y}_i$ = Predicted value

Root Mean Square Error (RMSE)

The square root value of MSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE)

The average of the absolute value of the difference between the actual and the predicted value.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}|$$

Coefficient of Determination (R^2)

The percentage of variability within the values that can be explained by the regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$\bar{y}$ – Mean of the actual values

Mean Absolute Percentage Error ((MAPE)

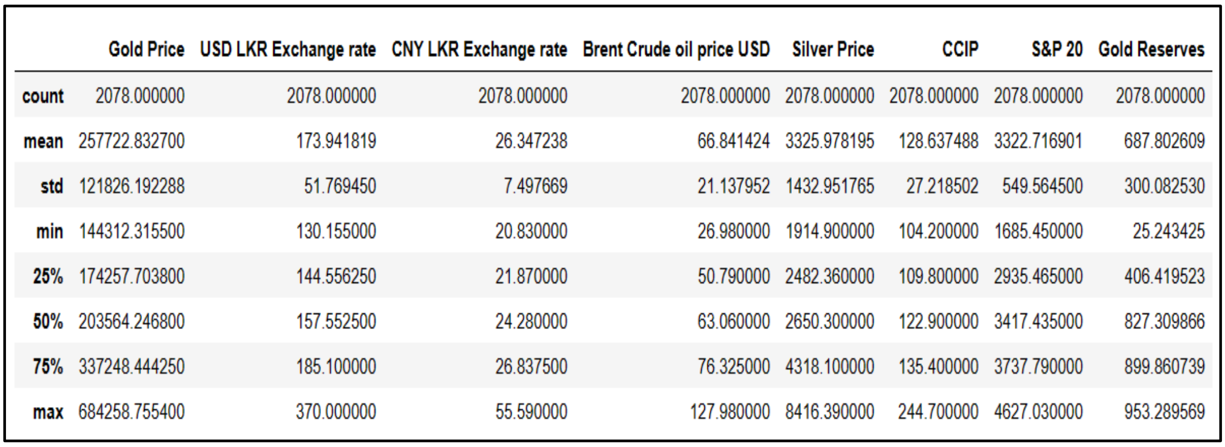
Calculate the relative error as a percentage of the actual value.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}|}{y_i}$$

Chapter 4
Evaluation and Results

This study focuses on predicting the gold price of Sri Lanka using exogenous variables including USD LKR Exchange rate, CNY LKR Exchange rate, Brent Crude oil price in USD, Silver Price in USD, CCIP, S&P 20. The price prediction was in a daily timeline based on the gold prices starting from January 2014 to September 2022.

A basic analysis of the variables given in the below table.



	Gold Price	USD LKR Exchange rate	CNY LKR Exchange rate	Brent Crude oil price USD	Silver Price	CCIP	S&P 20	Gold Reserves
count	2078.000000	2078.000000	2078.000000	2078.000000	2078.000000	2078.000000	2078.000000	2078.000000
mean	257722.832700	173.941819	26.347238	66.841424	3325.978195	128.637488	3322.716901	687.802609
std	121826.192288	51.769450	7.497669	21.137952	1432.951765	27.218502	549.564500	300.082530
min	144312.315500	130.155000	20.830000	26.980000	1914.900000	104.200000	1685.450000	25.243425
25%	174257.703800	144.556250	21.870000	50.790000	2482.360000	109.800000	2935.465000	406.419523
50%	203564.246800	157.552500	24.280000	63.060000	2650.300000	122.900000	3417.435000	827.309866
75%	337248.444250	185.100000	26.837500	76.325000	4318.100000	135.400000	3737.790000	899.860739
max	684258.755400	370.000000	55.590000	127.980000	8416.390000	244.700000	4627.030000	953.289569

Figure 17: Basic Analysis of Data

Based on the above data, the gold price and S&P 20 indicates a significant variability while CNY LKR exchange rate indicates a low variability. The substantial variability observed in historical gold prices, as indicated by the high standard deviation of USD 121,826.19, suggests that predicting future gold prices using a machine learning model be challenging.

4.1 Univariate Analysis

Gold Price

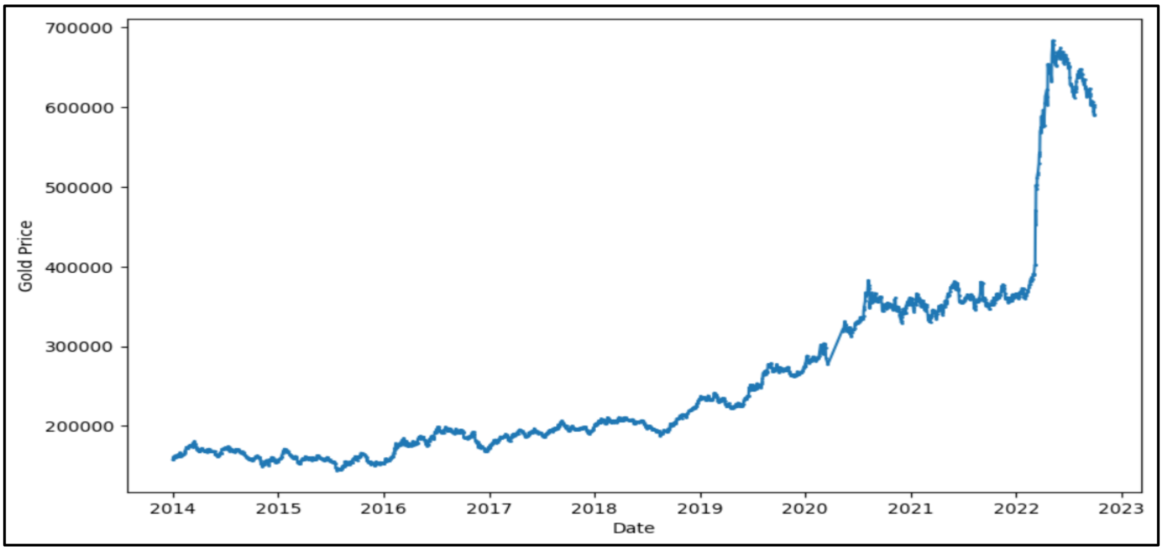


Figure 18: Time Plot of Gold Price

Central banks increased gold reserves in 2022, which led to a surge in prices in 2022. The purchase by central banks reflected their desire to increase gold purchases due to high levels of inflation globally, geo-political uncertainty driven by Russia's invasion of Ukraine and financial market turbulence.

In 2022, central banks acquired 1,136 tons of gold valued at approximately $70 billion, marking the largest acquisition in any year since 1967. The purchases made in the fourth quarter alone, totaling 417 tons, nearly equaled those of the entire year of 2021, which amounted to 450 tons.

USD LKR Exchange Rate / CNY LKR Exchange Rate

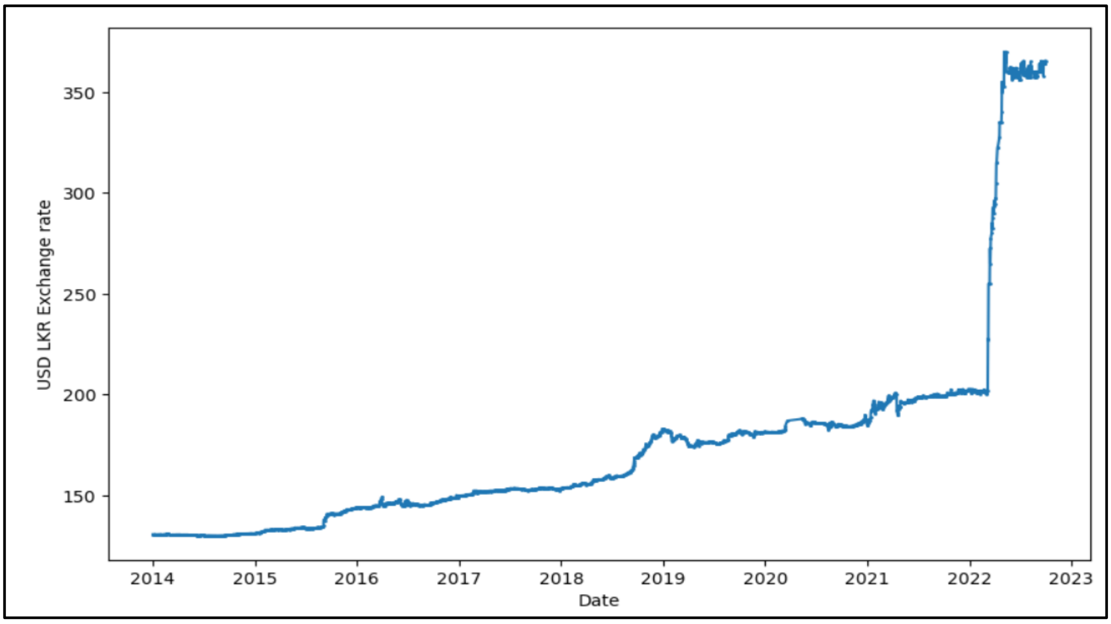


Figure 19: Time Plot of USD LKR Exchange Rate

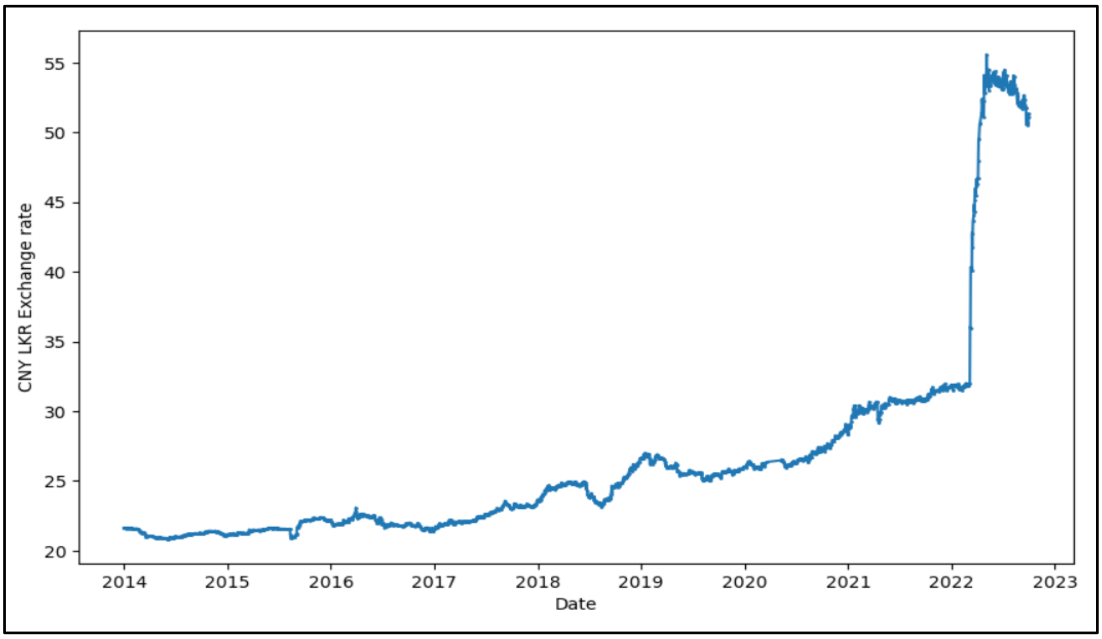


Figure 20: Time Plot of CNY LKR Exchange Rate

The Central Bank of Sri Lanka pegged the USD LKR rate at the range of LKR 200 – 203 and then subsequently gradually let the USD LKR float. This led to a depreciation of the rupee against the USD, due to market fundamentals. Post Covid, Sri Lanka was struggling with economic woes and saw a decrease in exports which impacted the country’s reserves balance.

The floating of the rupee led to the LKR depreciating against all other major currencies.

Silver Price

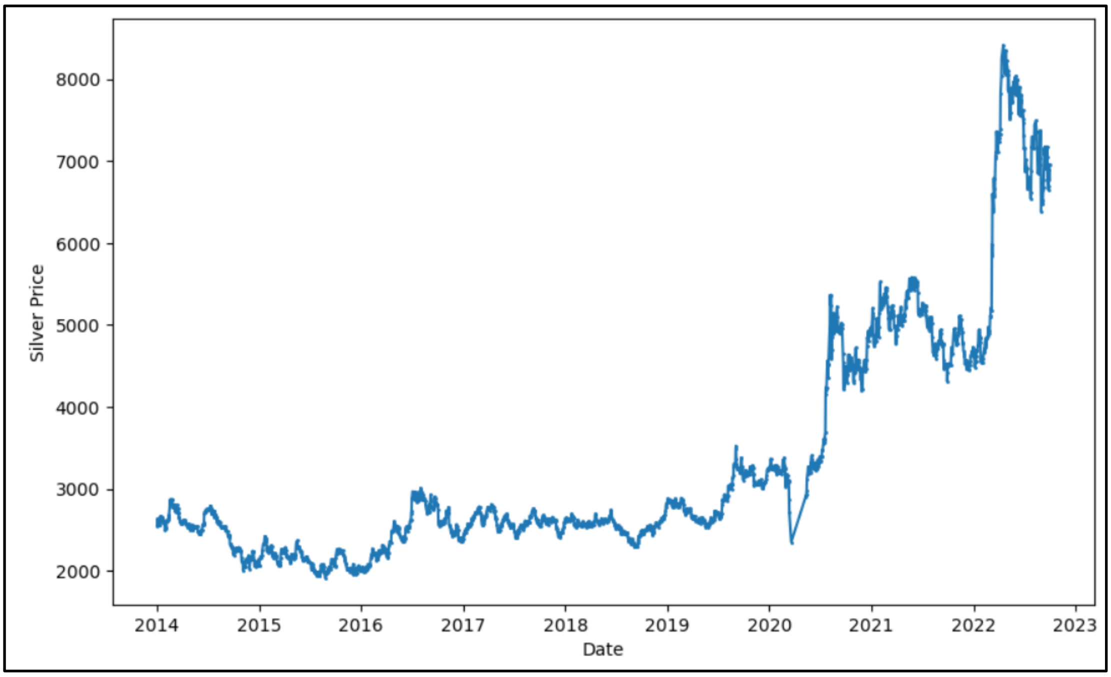


Figure 21: Time Plot of Silver Price

Silver prices too followed the same trajectory as gold prices due to high global inflation and geo-political uncertainty.

Brent Crude Oil Price

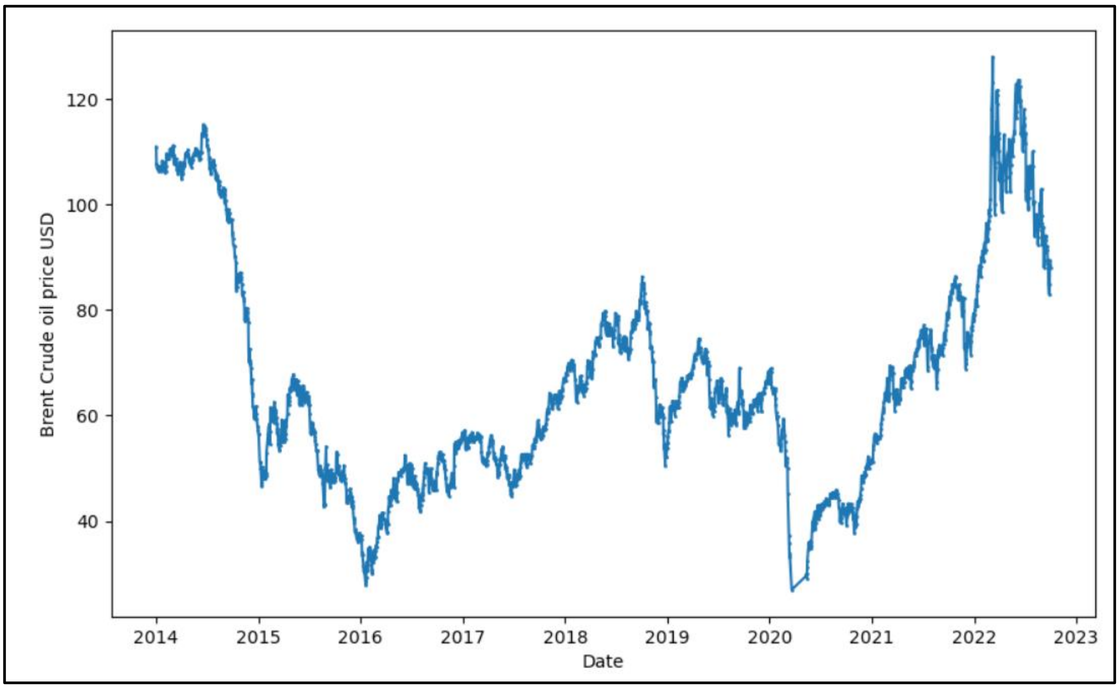


Figure 22: Time Plot of Brent Crude Oil Price

Crude oil prices increased from in 2022, mainly driven by the geo-political tensions between Russia and Ukraine. Further, oil prices were impacted in 2022, driven the production cuts by OPEC+.

CCPI

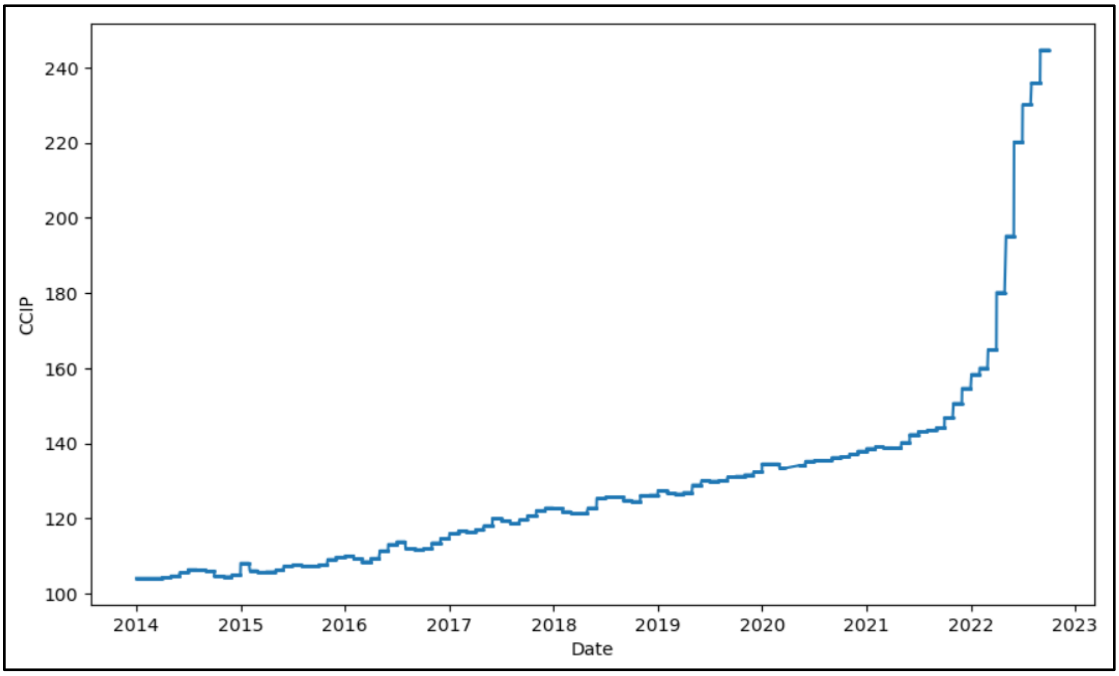


Figure 23: Time Plot of CCPI

The Colombo Consumer Price Index has been rising gradually since October 2021 and saw a significant increase due to the economic crisis in Sri Lanka. Sri Lanka abandoned the peg and in early 2022 and the Rupee depreciated against the dollar steeply from LKR 200 to LKR 360 post abandoning the page. This increased headline inflation and core inflation in the country. During 2022, headline inflation increased the highest by 73.7% in September 2022.

S&P 20 Index

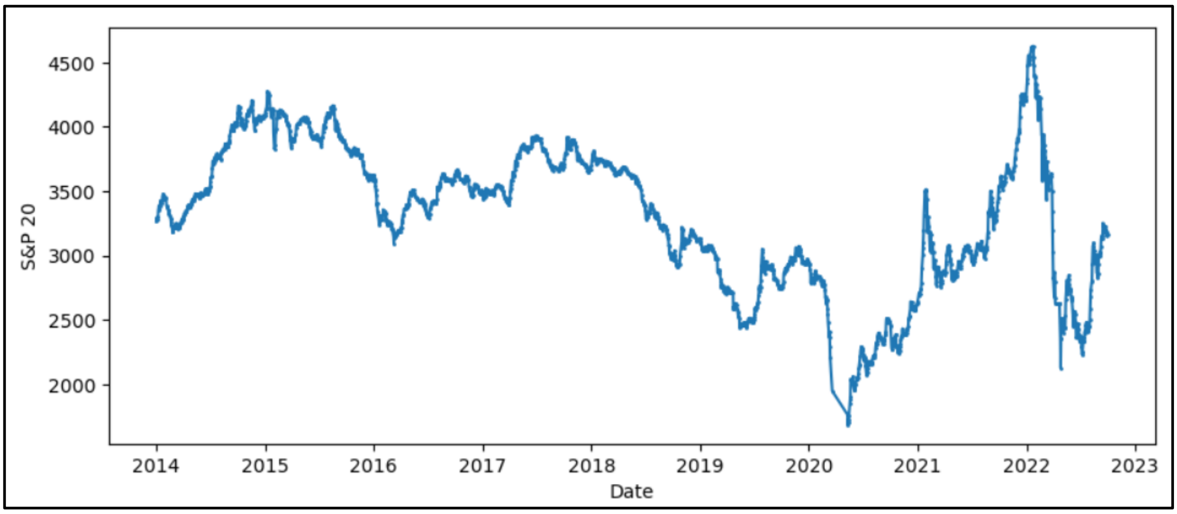


Figure 24: Time Plot of S&P 20 Index

The S&P 20 saw one of the lowest points in history in 2020 due to the country coming to a standstill due to the Covid-19 pandemic. The country's regulator issued a directive to close the stock market commencing 18th April and markets subsequently reopened on 11th May 2020.

In 2021 markets were pushed through primarily by the local investors who favored particular stocks which influenced market capitalization. Despite Covid-19 slowing down the economy there was disconnect between the ground business environment and listed stocks. Certain companies benefitted heavily due to the import restrictions as they became the main companies to supply goods which were otherwise imported. (Lanka Tiles, Rocell). The artificial disconnect continued throughout the year between small businesses and large listed companies. From October 2021 the Sri Lankan Rupee was pegged to the dollar at LKR 200 – 203 which also led to furthering the economic crisis.

In 2022, markets started to decline with the uncertainty of the political situation in the country which commenced with the Aragalaya. The protests commenced in March 2022 against the GoSL led by then president Gotabhaya Rajapakse resigned by April 2022, creating uncertainty in the economy. As mentioned above, Sri Lanka had artificially pegged the Rupee to the Green back for a few months and commenced floating the Rupee gradually from March 2022. The Aragalaya peaked with the resignation of the President Gotabaya Rajapakse in July 2022, and Parliament elected Ranil Wickremesinghe as the president of the country. Subsequently, the country went through severe economic crisis with the unavailability of fuel and gas which led to a complete slowdown in the economy. The Government had to request for a bailout package from the IMF for the 17th time in its history and received its first tranche of USD 330m in March 2023.

4.2 Bi-Variate Analysis

The depicted below figure illustrates a strong correlation between gold prices and some key variables, namely the USD LKR Exchange Rate, CNY LKR Exchange Rate, and silver prices. Furthermore, the observed correlations among these predictor variables indicate the presence of multicollinearity.

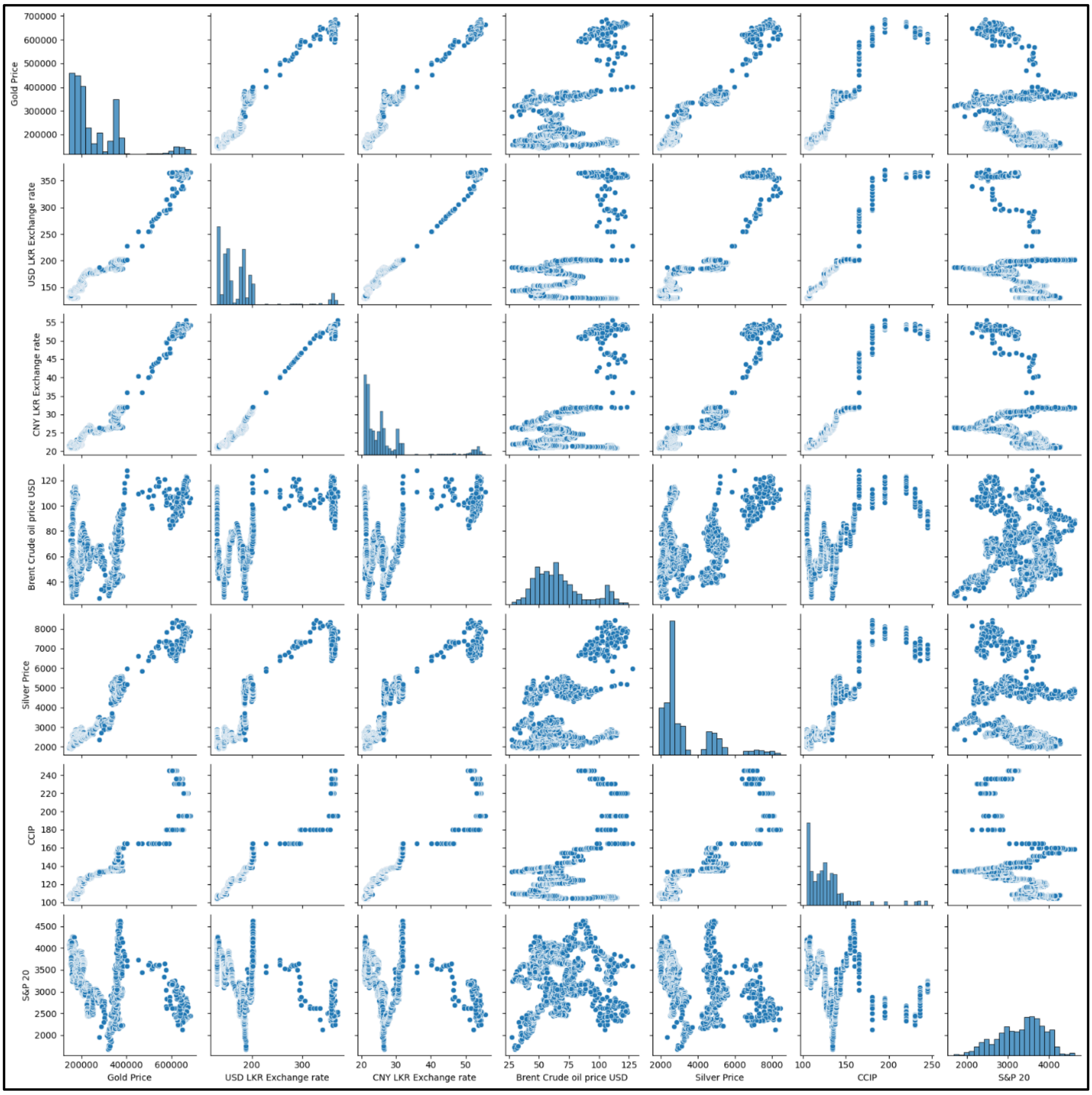


Figure 25: Plot of Correlations of Variables

4.3 Data Credibility

All the data gathered for this study is secondary data, and the sources for each variable are indicated below.

Table 3: Data Sources

Variable	Data Source
Gold Price	Central Bank of Sri Lanka
USD LKR Exchange Rate	Investing.com
CNY LKR Exchange Rate	Investing.com
Brent Crude Oil Price	Investing.com
Silver Price	bullion-rates.com
CCPI	Census and Statistics Department of Sri Lanka
S&P 20	spglobal.com
Gold Reserves	World Gold Council

4.4 Description of Data

Gold Price

This represents the gold prices in Sri Lanka per Troy ounce in LKR from January 2014 to September 2022.

USD LKR Exchange rate

This is the value of one U.S. Dollar in terms of Sri Lankan Rupees.

CNY LKR Exchange rate

This is the value of one Chinese Yuan in terms of Sri Lankan Rupees.

Brent Crude Oil Price

Two primary types of crude oil prices that are widely referenced in the global market are Brent crude and West Texas Intermediate (WTI) crude. This is named after the Brent oil field in the North Sea, Brent crude is extracted from several oil fields in the North Sea, including those in the United Kingdom and Norway. It represents the cost of a barrel of Brent crude oil on the global market. This is measured in USD.

Silver Price

This represents the gold prices in Sri Lanka per ounce in LKR from January 2014 to September 2022.

Colombo Consumer Price Index (CCPI)

This indicator serves as a gauge for inflation in Sri Lanka. The index is based on several years, including 2002, 2006, and 2013. This data is sourced from the economic data library of the Central Bank of Sri Lanka. The year 2013 was selected as the base year for constructing the CCPI variable in the dataset.

S&P SL 20

S&P SL 20 Index is a stock market index that represents the performance of the top 20 companies listed on the Colombo Stock Exchange (CSE) in Sri Lanka. It is designed to reflect the overall performance of the Sri Lankan stock market by including a diversified group of leading companies from various sectors. In April 2020, the CSE was shut down, leading to a halt in all operations and trading activities. Hence, the data set excludes the records for that particular month.

4.5 Advance Analytics and Results

In Chapter 3, the study unfolds in two distinct phases. In the initial phase, we trained two models, XGBoost and Random Forest, utilizing all the significant lag variables meticulously identified in the preceding chapter. Employing feature importance methods in both models, we discern and highlight the crucial features within each.

Moving on to the second phase, a refined approach is adopted. Both models are subjected to retraining, this time incorporating only the most important variables as identified in the earlier analysis. Subsequently, a thorough analysis is conducted on the results derived from this second phase, providing a comprehensive understanding of the impact and significance of the chosen variables.

4.5.1 Phase 1

4.5.1.1 XGBoost

Parameter Tuning

In the process of refining our model's performance, grid search in conjunction with a five-fold time series cross-validation strategy was used. This method allows to systematically explore and evaluate various combinations of hyperparameter values, ensuring that our model is fine-tuned for optimal performance across diverse subsets of the data. By incorporating five-fold time series cross-validation, it is aimed to obtain a robust assessment of the model's generalization capabilities, enhancing its reliability on unseen data.

The following table shows the best parameters obtained for the XGBoost model with all the predictor variables.

Table 4: Optimal Hyperparameter for XGBoost with all the Predictor Variables

Parameter	**Value**
n_estimators	875
learning_rate	0.15
alpha	1.5
reg_lambda	3
max_depth	4
sub_sample	1
gamma	0.5
colsample_bytree	1

The model was trained using the above optimal hyperparameters.

Measures of Accuracy

Values obtained for the accuracy measures are given in the following table.

Table 5: Table of Accuracy Measures for XGBoost with all the Predictor Variables

Measure	**Value**
R^2	0.88
MSE	1,914,180,697.98
RMSE	43,751.35
MAPE	5.80%
MAE	28,908.73

Feature Selection

The feature importance mechanism in XGBoost was used to identify the most influencing variables for the prediction in the model. The plot of important features is shown below.

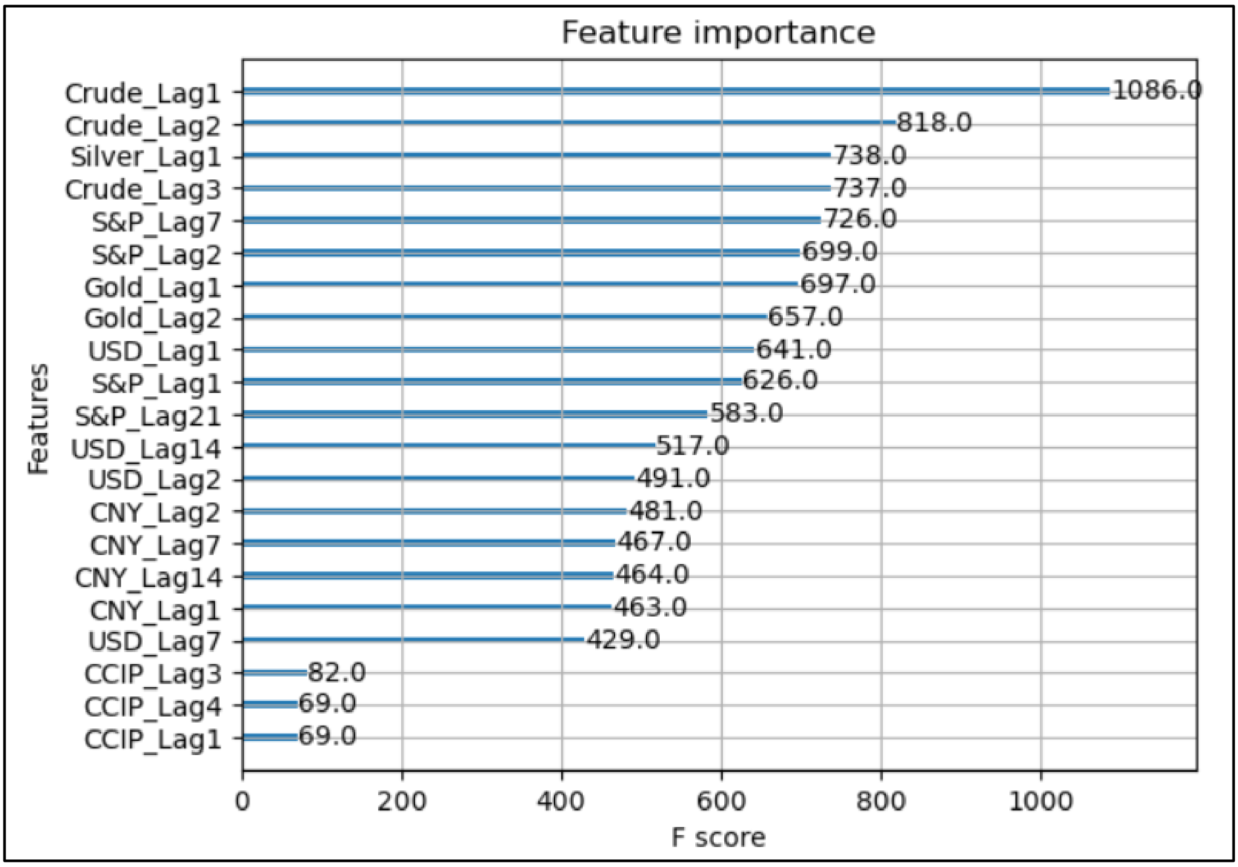


Figure 26: Feature Importance Plot of XGBoost

The variables with a F score above 500, were chosen as the most influencing features and XGBoost was retrained with these variables.

4.5.1.2 Random Forest

Parameter Tuning

The following table shows the best parameters obtained for the Random Forest Regressor model with all the predictor variables.

Table 6: Optimal Hyperparameters for Random Forest Regressor with all the Predictor Variables

Parameter	Value
n_estimators	890
max_depth	3
min_samples_leaf	2
min_samples_split	2

Measures of Accuracy

Values obtained for the accuracy measures are given in the following table.

Table 7: Table of Accuracy Measures for Random Forest Regressor with all the Predictor Variables

Measure	Value
R^2	0.88
MSE	1,937,515,837.57
RMSE	44,017.22
MAPE	8.40%
MAE	37,642.06

Feature Selection

The feature importance mechanism in Random Forest was used to identify the most influencing variables for the prediction in the model. The plot of important features is shown below.

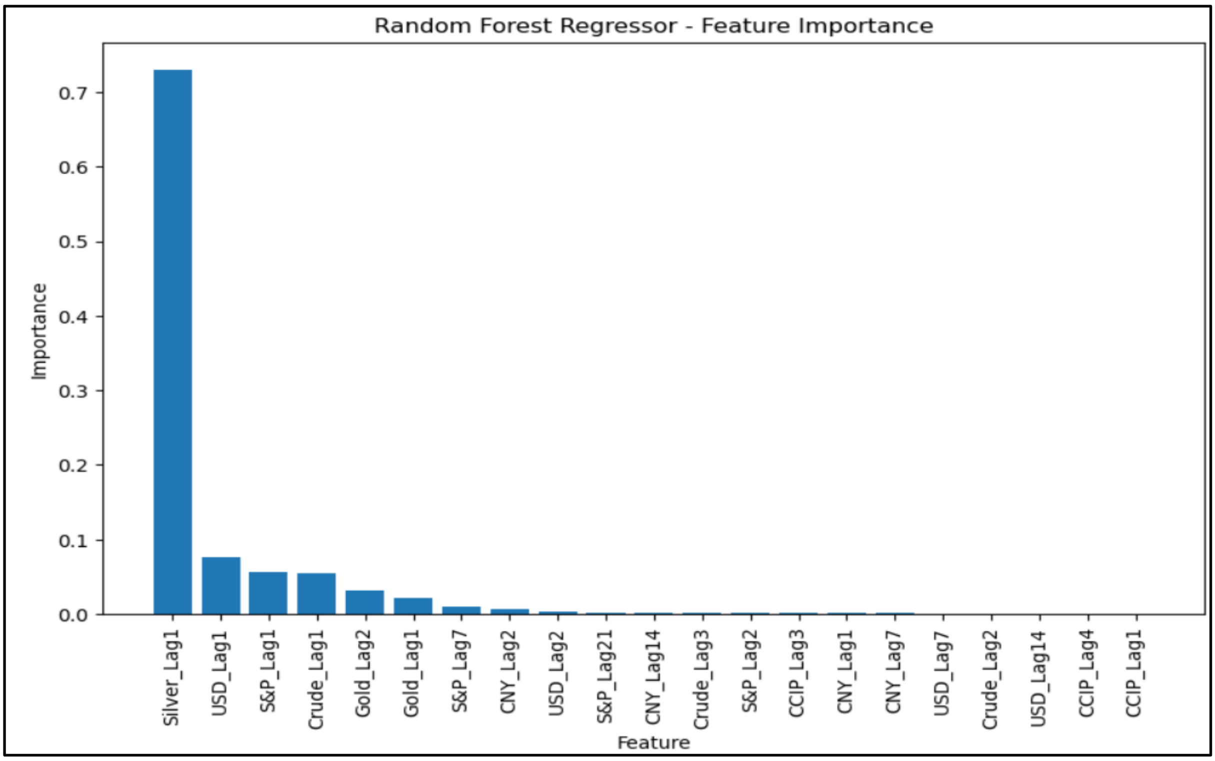


Figure 27: Feature Importance Plot of Random Forest Regressor

The variables which are most significant (From Silver_Lag1 to S&P_Lag7) were chosen as the most influencing factor under the Random Forest and the model was retrained using the selected variables.

4.5.2 Phase 2

4.5.2.1 XGBoost

Parameter Tuning

The following table shows the best parameters obtained for the XGBoost model with the selected predictor variables.

Table 8: Optimal Hyperparameters for XGBoost with Selected Predictor Variables

Parameter	**Value**
n_estimators	1,000
learning_rate	1
alpha	7
reg_lambda	6
max_depth	4
sub_sample	1
gamma	0.5
colsample_bytree	1

Measures of Accuracy

Values obtained for the accuracy measures are given in the following table.

Table 9: Table of Accuracy Measures for XGBoost with the Selected Predictor Variables

Measure	**Value**
R^2	0.89
MSE	1,696,944,054.71
RMSE	41,193.98
MAPE	5.38%
MAE	27,155.94

Important Features

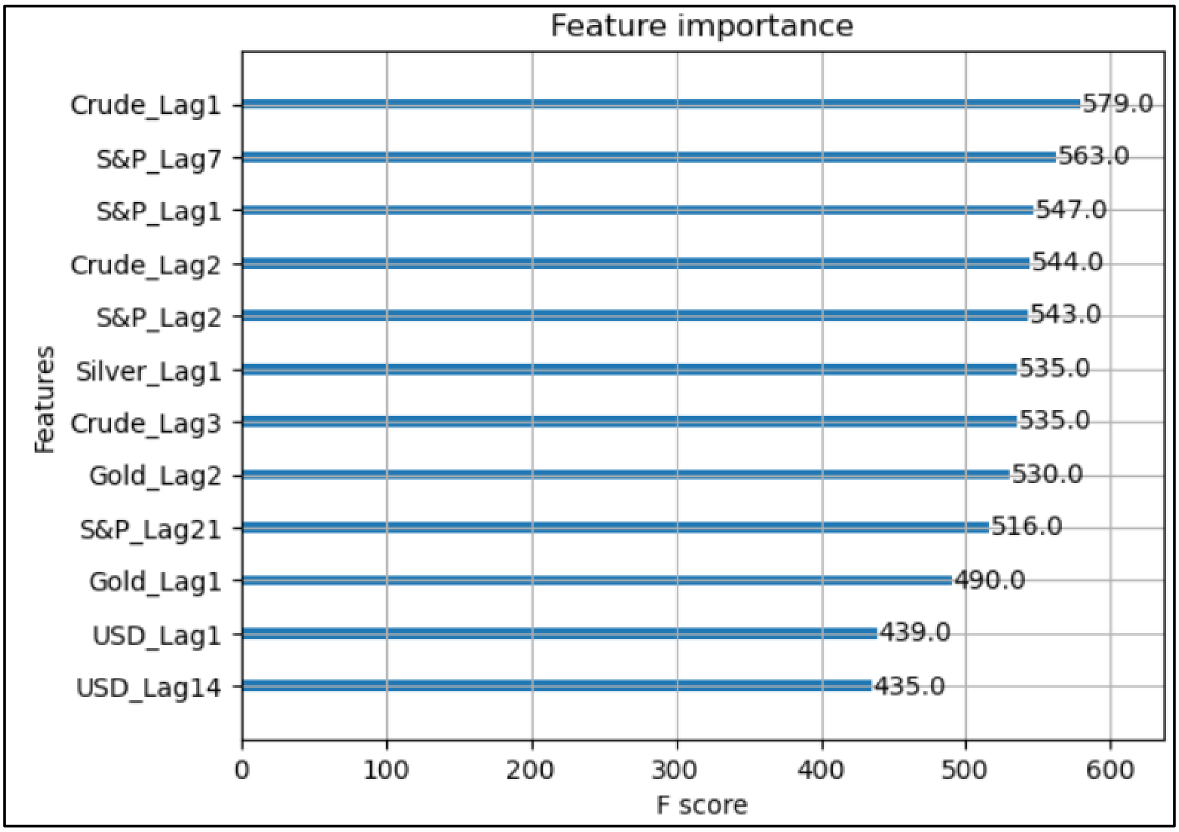


Figure 28: Feature Importance of the Selected Predictor Variables in XGBoost

4.5.2.2 Random Forest

Parameter Tuning

The following table shows the best parameters obtained for the Random Forest model with the selected predictor variables.

Table 10: Optimal Hyperparameters for Random Forest Regressor with the Selected Predictor Variables

Parameter	**Value**
n_estimators	300
max_depth	3
min_samples_leaf	2
min_samples_split	4

Measures of Accuracy

Values obtained for the accuracy measures are given in the following table.

Table 11: Table of Accuracy Measures for Random Forest Regressor with the Selected Predictor Variables

Measure	**Value**
R^2	0.81
MSE	3,045,017,182.36
RMSE	55,181.67
MAPE	7.73%
MAE	21,039.54

Important Features

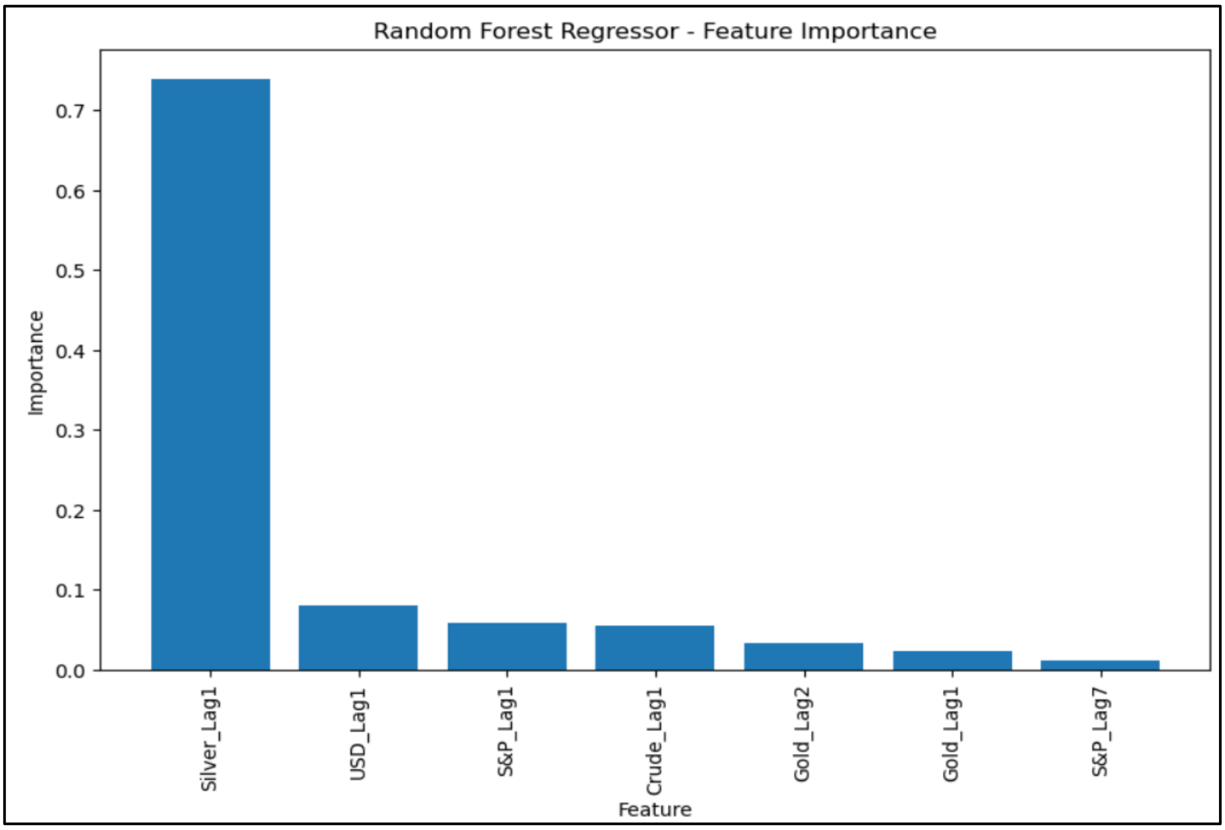


Figure 29: Feature Importance of the Selected Predictor Variables in the Random Forest

4.6 Summary of Measures of Accuracy

Table 12: Table of Summary of Accuracy Measures

	XGBoost		Random Forest	
Measure	Without Feature Selection	With Feature Selection	Without Feature Selection	With Feature Selection
R^2	0.88	0.89	0.88	0.81
MSE	1,914,180,697.98	1,696,944,054.71	1,937,515,837.57	3,045,017,182.36
RMSE	43,751.35	41,193.98	44,017.22	55,181.67
MAPE	5.80%	5.38%	8.40%	7.73%
MAE	28,908.73	27,155.94	37,642.06	21,039.54

4.7 Model Comparison

The evaluation of model performance encompassed a comprehensive set of metrics, including the root mean squared error (RMSE), mean square error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and the coefficient of determination (R^2). The high R^2 values obtained from both models, considering all predicted variables as well as the selected subset, signify their ability in explaining a substantial portion of variability in the dependent variable, namely the Sri Lankan gold price. The model with lowest values of MSE, RMSE, MAPE and MAE and the highest value of R^2 is considered to be the best model. From the above table, it is clearly evident that XGBoost is significantly better than the Random Forest model. Also, it is noticed that the model exhibited better performance when trained solely on the crucial features.

The change in feature importance scores when the XGBoost model was trained on all predictor variables compared to when it was trained only on the most important features can occur due to several reasons. XGBoost calculates feature importance based on the contribution of each feature to the model's performance. When the model is trained with all predictor variables, it considers the interactions and dependencies among features. Removing certain features may change the relative importance of the remaining features because their importance is now assessed in the context of a different set of variables. Also, training the model with fewer

features changes the complexity of the model. Feature importance is calculated based on how features contribute to the model's ability to reduce the error. If the model is simpler with fewer features, the importance scores may be distributed differently.
Based on the figure 28, it can be concluded that below lag features are important in predicting Sri Lankan gold price.

Table 13: Important Lags to Predict Sri Lankan Gold Price Using XGBoost

Variable	**Lag**
Crude Oil Price	1,2,3
S&P 20	1,2,7,21
Silver Price	1
Gold Price	1,2
USD LKR Exchange Rate	1,14

Chapter 5
Conclusion and Future Work

5.1 General Discussion

Predicting gold prices holds paramount importance due to its implications for financial markets, investment strategies, and economic stability. Gold serves as a crucial indicator of global economic health, often responding to geopolitical events, inflation, and currency fluctuations. The prediction of gold prices is challenging due to the intricate diverse factors impacting the value of this precious metal. This study pursues a dual objective. Firstly, to identify the factors affecting the gold price, and secondly, to construct an effective ML model for predicting daily gold prices in Sri Lanka using these identified variables. The XGBoost and Random Forest were used in this study. The results showed that XGBoost outperforms Random Forest. Also, the most significant factors in predicting the gold price were chosen using the feature importance mechanism in XGBoost.

5.2 Applications to the Field of Study

When looking at the goals of this study, it's clear that it stands out from other studies in the local context that focus on predicting gold prices. The results derived from the analysis offer stakeholders a comprehensive understanding of the domain. Specifically, variables such as crude oil price, USD LKR exchange rate, CNY LKR exchange rate, CCPI, and silver price emerge as potent predictors, significantly influencing the daily gold prices in Sri Lanka. Intriguingly, the study underscores that gold reserves in the central bank do not have any influence on gold prices in the Sri Lankan context. Analysts and experts can enhance the accuracy of the findings of this study by incorporating additional data from diverse sources. This approach ensures a more substantial contribution from this framework to the decision-making processes of investors and other economic activities.

5.3 Limitations of Study

I. The research encountered a challenge as certain variables initially intended for analysis were not available on a daily basis. Notably, some factors like gold demand and supply are reported in quarterly terms, making their conversion to a daily frequency prone to substantial errors. Consequently, the researcher made the decision to forgo the inclusion of these variables to ensure the integrity and precision of the analysis.

II. The study was constrained to the time frame, spanning from January 2014 to September 2022, due to limitations in the data available from various resources.

III. Certain variables suggested in previous literature were omitted from the forecasting of the gold price due to unavailability of the required data.

IV. The gold price is influenced not only by the independent variables proposed in this study but also by other factors like macroeconomic news announcements and money supply, which play a significant role in affecting gold prices.

V. The research utilized data obtained from secondary sources, and as a result, the inherent limitations associated with secondary data are expected to impact this study.

VI. The findings of this study are relevant exclusively to the Sri Lankan market and provide benefits primarily for Sri Lankan policymakers and participants in the local market.

5.4 Suggestions for Future Work

The upward trend of gold prices and the economic uncertainty in recent years have prompted numerous researchers to research on gold-related topics. Consequently, this research study serves as a foundational resource for future researchers interested in solving this issue.

Firstly, despite the determinants proposed in this study, it is recommended that future researchers consider incorporating additional variables that significantly impact gold prices. Factors such as money supply, macroeconomic announcements, and jewelry demand could provide a more comprehensive understanding of the dynamics influencing gold prices. Given gold's dual nature as a safe haven asset and investment tool, policymakers may require a more intricate framework to inform their investment decisions. Hence, future researchers are encouraged to explore gold in different contexts, such as its volatility and returns rather than just its price. This approach could offer valuable insights into the dynamics of individual assets and portfolios, leading to improved risk and portfolio management strategies for financial planners.

Secondly, future researchers are advised to investigate related issues under different economic conditions, particularly during economic crises and periods of stability. Different factors may significantly influence gold prices under varying economic circumstances, and a comprehensive examination can provide decision-makers with tailored insights into investment strategies.

Further research studies can be conducted to explore the application of advanced models, such as Neural Network, GARCH, ARCH, hybrid models and similar approaches, to predict gold prices.

Furthermore, as this study focuses solely on Sri Lanka, future research should adopt a broader perspective by considering countries with diverse economic statuses, encompassing developed, developing, and emerging economies. By selecting multiple countries, researchers can uncover unique findings from different regions, contributing more meaningful and relevant information for policymakers, economists, and investors worldwide.

5.5 Conclusion

Predicting daily gold price is a bit challenging due to the high volatility in the price. This study, predicts Sri Lanka's daily gold price using exogenous variables such as Brent crude oil price, USD LKR exchange rate, CNY LKR exchange rate, S&P SL 20 index, silver price, and CCPI. The Augmented Dickey-Fuller test showed that all the variables including the response variable were non-stationary. Hence first differencing was used on the full data set in order to make the variables stationary. The significant lags of the predictor variables were identified using the CCF plots while the significant lags of the response variable were found using the PACF. The XGBoost and Random Forest models were trained and tested using the selected lag variables of differenced data. Then inverse differencing was performed on the predicted differenced data in order to predict on the original scale. Finally, R^2, MSE, MAPE, RMSE and MAE were used to evaluate the performance of the model with feature selection and without feature selection.

The highest R^2 value and lowest MSE, MAPE, RMSE and MAE values proved that the best ML model to predict the gold price in Sri Lanka is XGBoost, between XGBoost and Random Forest. The proposed XGBoost model is a moderately good technique due to its ability to predict the daily gold price with an approximate error of 5.38%. Given the high volatility in the Sri Lankan gold price, this model can be considered moderately good.

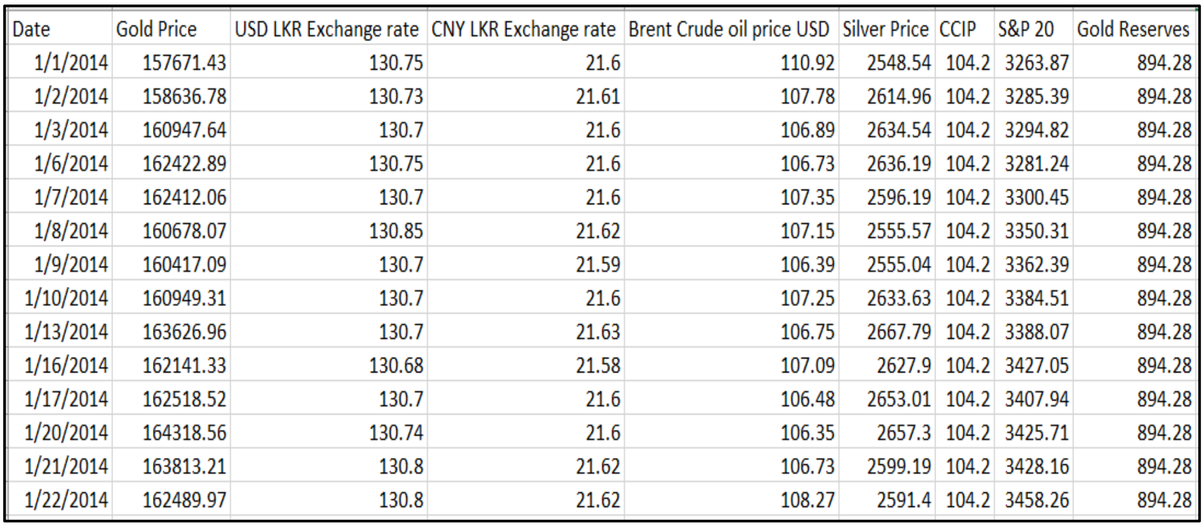
Furthermore, it was identified that following day's gold price is highly affected by the previous day's Brent crude oil price. Also, it is identified previous day's Brent crude oil price is affecting more to the following day's gold price than the previous day's gold price.

The following features given are required in predicting the gold price in Sri Lanka.

I. Brent crude oil previous 3 days' prices
II. S&P SL 20 previous two days, previous week same day value, previous 3 weeks' same day value
III. Silver previous day price
IV. Previous 2 days' gold prices
V. Previous day and previous 2 weeks' same day USD LKR exchange rate

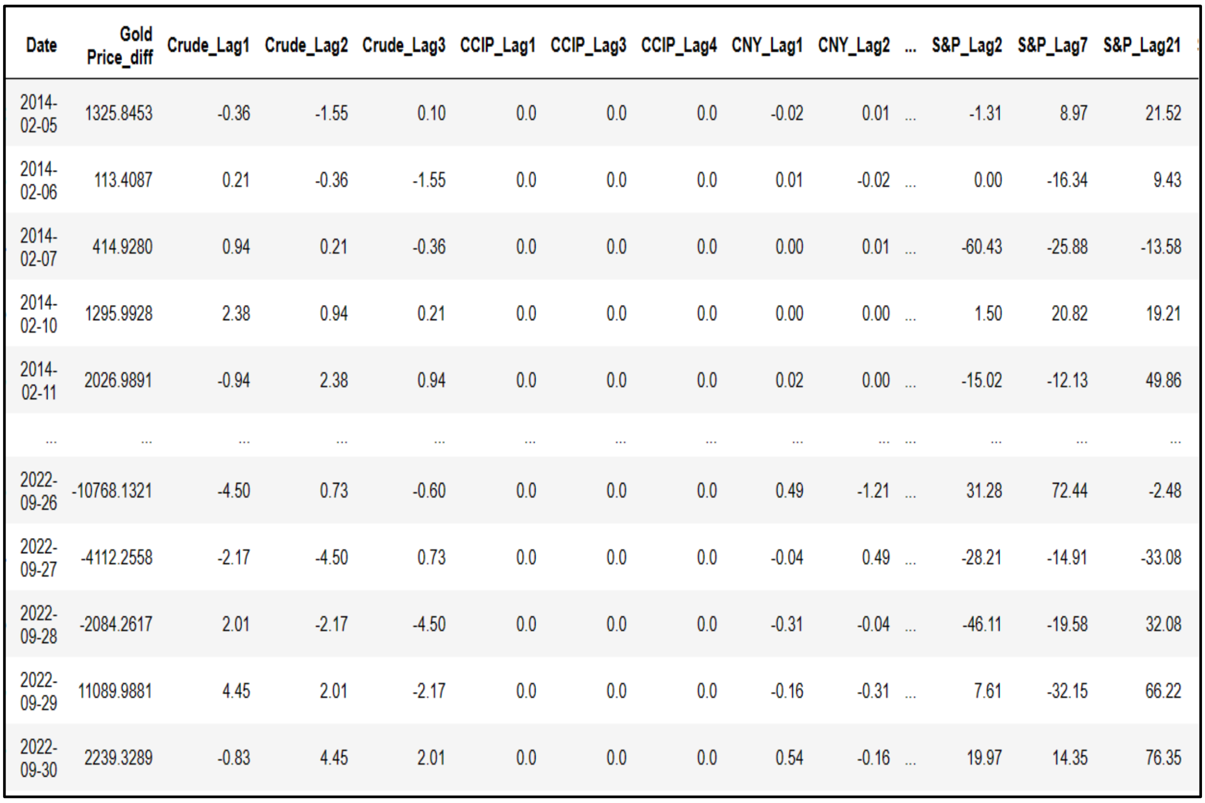
Appendices

Appendix A: Snapshots of the initial data sets



Date	Gold Price	USD LKR Exchange rate	CNY LKR Exchange rate	Brent Crude oil price USD	Silver Price	CCIP	S&P 20	Gold Reserves
1/1/2014	157671.43	130.75	21.6	110.92	2548.54	104.2	3263.87	894.28
1/2/2014	158636.78	130.73	21.61	107.78	2614.96	104.2	3285.39	894.28
1/3/2014	160947.64	130.7	21.6	106.89	2634.54	104.2	3294.82	894.28
1/6/2014	162422.89	130.75	21.6	106.73	2636.19	104.2	3281.24	894.28
1/7/2014	162412.06	130.7	21.6	107.35	2596.19	104.2	3300.45	894.28
1/8/2014	160678.07	130.85	21.62	107.15	2555.57	104.2	3350.31	894.28
1/9/2014	160417.09	130.7	21.59	106.39	2555.04	104.2	3362.39	894.28
1/10/2014	160949.31	130.7	21.6	107.25	2633.63	104.2	3384.51	894.28
1/13/2014	163626.96	130.7	21.63	106.75	2667.79	104.2	3388.07	894.28
1/16/2014	162141.33	130.68	21.58	107.09	2627.9	104.2	3427.05	894.28
1/17/2014	162518.52	130.7	21.6	106.48	2653.01	104.2	3407.94	894.28
1/20/2014	164318.56	130.74	21.6	106.35	2657.3	104.2	3425.71	894.28
1/21/2014	163813.21	130.8	21.62	106.73	2599.19	104.2	3428.16	894.28
1/22/2014	162489.97	130.8	21.62	108.27	2591.4	104.2	3458.26	894.28

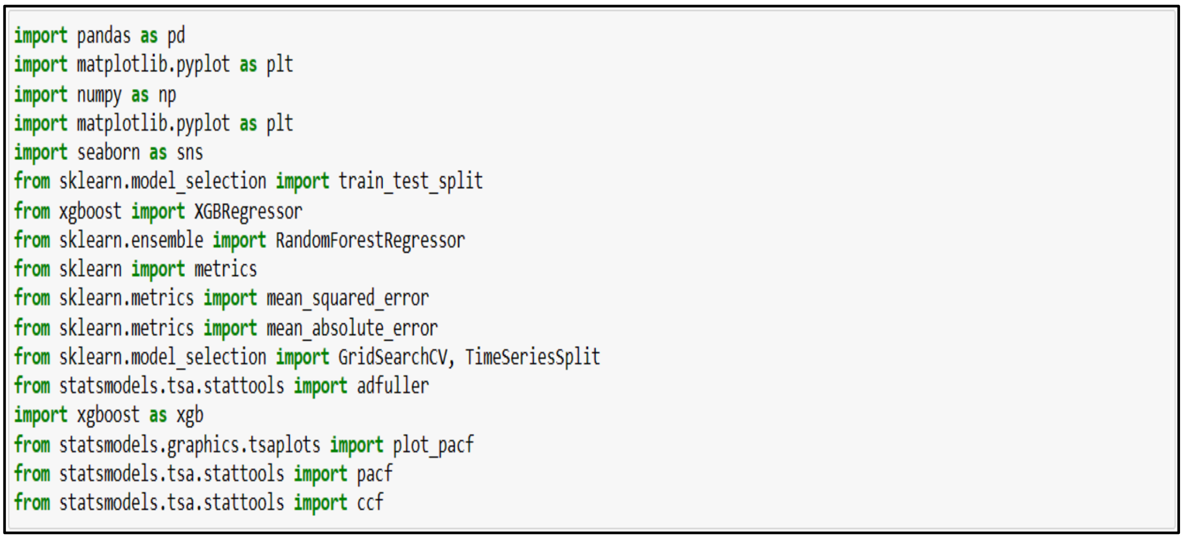
Appendix Figure A-1: Snapshot of the Initial Dataset



Date	Gold Price_diff	Crude_Lag1	Crude_Lag2	Crude_Lag3	CCIP_Lag1	CCIP_Lag3	CCIP_Lag4	CNY_Lag1	CNY_Lag2	...	S&P_Lag2	S&P_Lag7	S&P_Lag21
2014-02-05	1325.8453	-0.36	-1.55	0.10	0.0	0.0	0.0	-0.02	0.01	...	-1.31	8.97	21.52
2014-02-06	113.4087	0.21	-0.36	-1.55	0.0	0.0	0.0	0.01	-0.02	...	0.00	-16.34	9.43
2014-02-07	414.9280	0.94	0.21	-0.36	0.0	0.0	0.0	0.00	0.01	...	-60.43	-25.88	-13.58
2014-02-10	1295.9928	2.38	0.94	0.21	0.0	0.0	0.0	0.00	0.00	...	1.50	20.82	19.21
2014-02-11	2026.9891	-0.94	2.38	0.94	0.0	0.0	0.0	0.02	0.00	...	-15.02	-12.13	49.86
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2022-09-26	-10768.1321	-4.50	0.73	-0.60	0.0	0.0	0.0	0.49	-1.21	...	31.28	72.44	-2.48
2022-09-27	-4112.2558	-2.17	-4.50	0.73	0.0	0.0	0.0	-0.04	0.49	...	-28.21	-14.91	-33.08
2022-09-28	-2084.2617	2.01	-2.17	-4.50	0.0	0.0	0.0	-0.31	-0.04	...	-46.11	-19.58	32.08
2022-09-29	11089.9881	4.45	2.01	-2.17	0.0	0.0	0.0	-0.16	-0.31	...	7.61	-32.15	66.22
2022-09-30	2239.3289	-0.83	4.45	2.01	0.0	0.0	0.0	0.54	-0.16	...	19.97	14.35	76.35

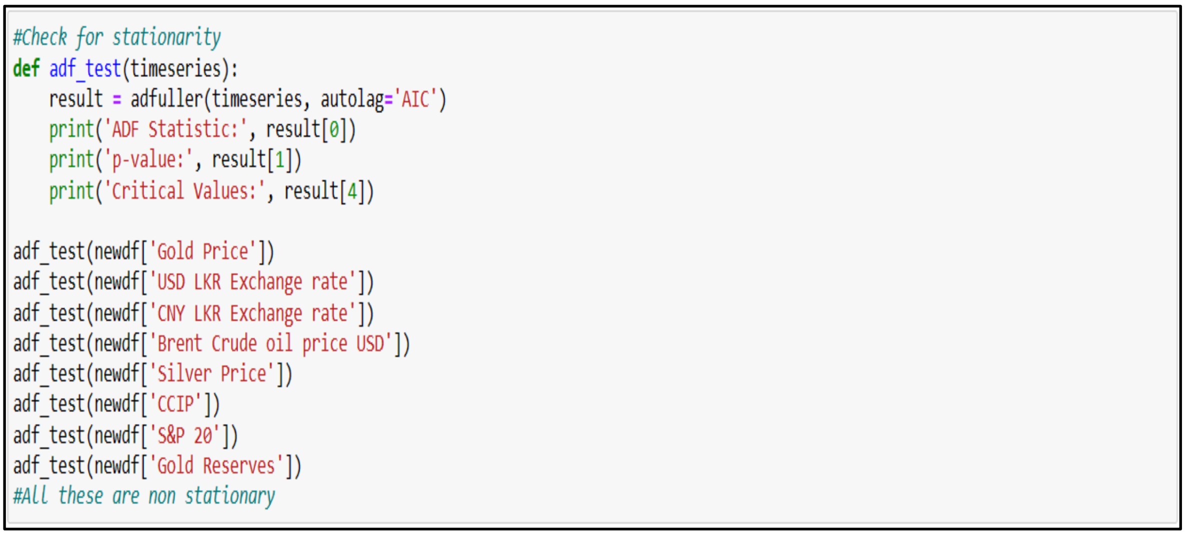
Appendix Figure A-2: Snapshot of the Final Dataset

Appendix B: Snapshots of the Python Code



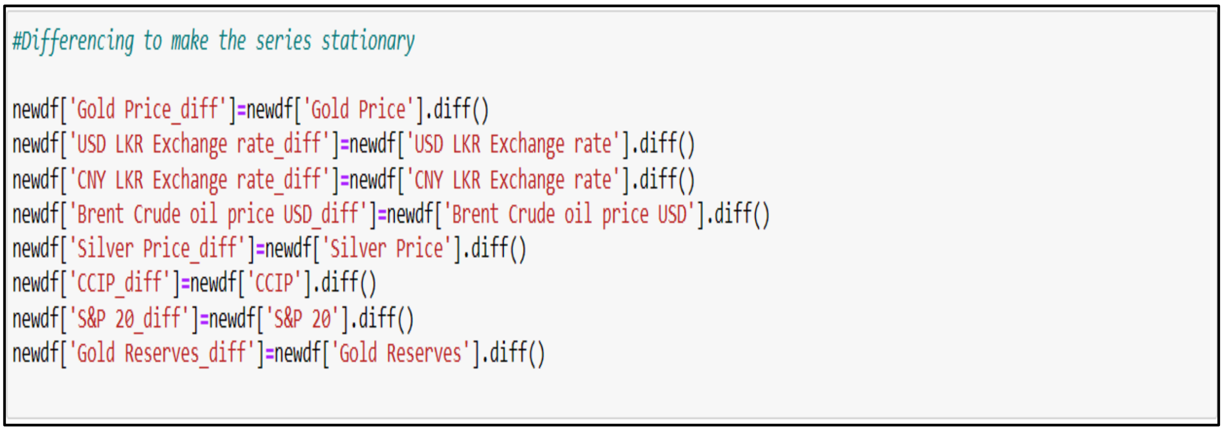
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from xgboost import XGBRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import GridSearchCV, TimeSeriesSplit
from statsmodels.tsa.stattools import adfuller
import xgboost as xgb
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.tsa.stattools import pacf
from statsmodels.tsa.stattools import ccf
```

Appendix Figure B-1: Sample of Python Libraries Used



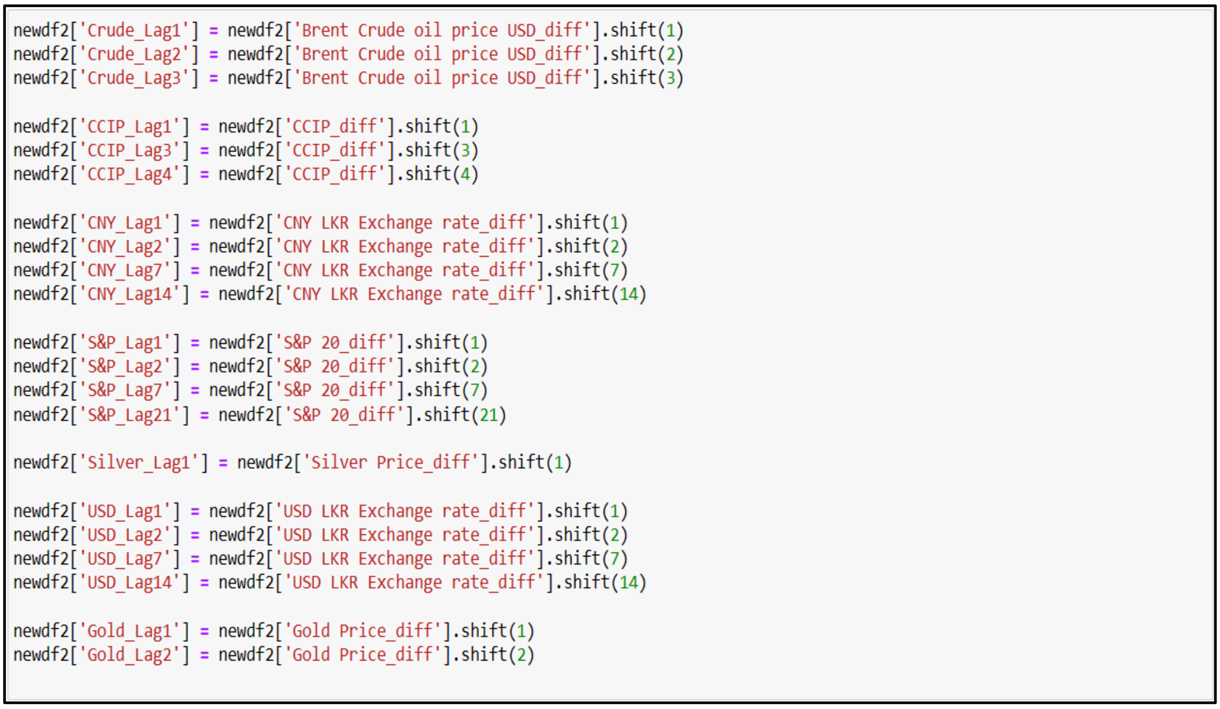
```
#Check for stationarity
def adf_test(timeseries):
    result = adfuller(timeseries, autolag='AIC')
    print('ADF Statistic:', result[0])
    print('p-value:', result[1])
    print('Critical Values:', result[4])

adf_test(newdf['Gold Price'])
adf_test(newdf['USD LKR Exchange rate'])
adf_test(newdf['CNY LKR Exchange rate'])
adf_test(newdf['Brent Crude oil price USD'])
adf_test(newdf['Silver Price'])
adf_test(newdf['CCIP'])
adf_test(newdf['S&P 20'])
adf_test(newdf['Gold Reserves'])
#All these are non stationary
```

Appendix Figure B-2: Sample Code to Check Stationarity



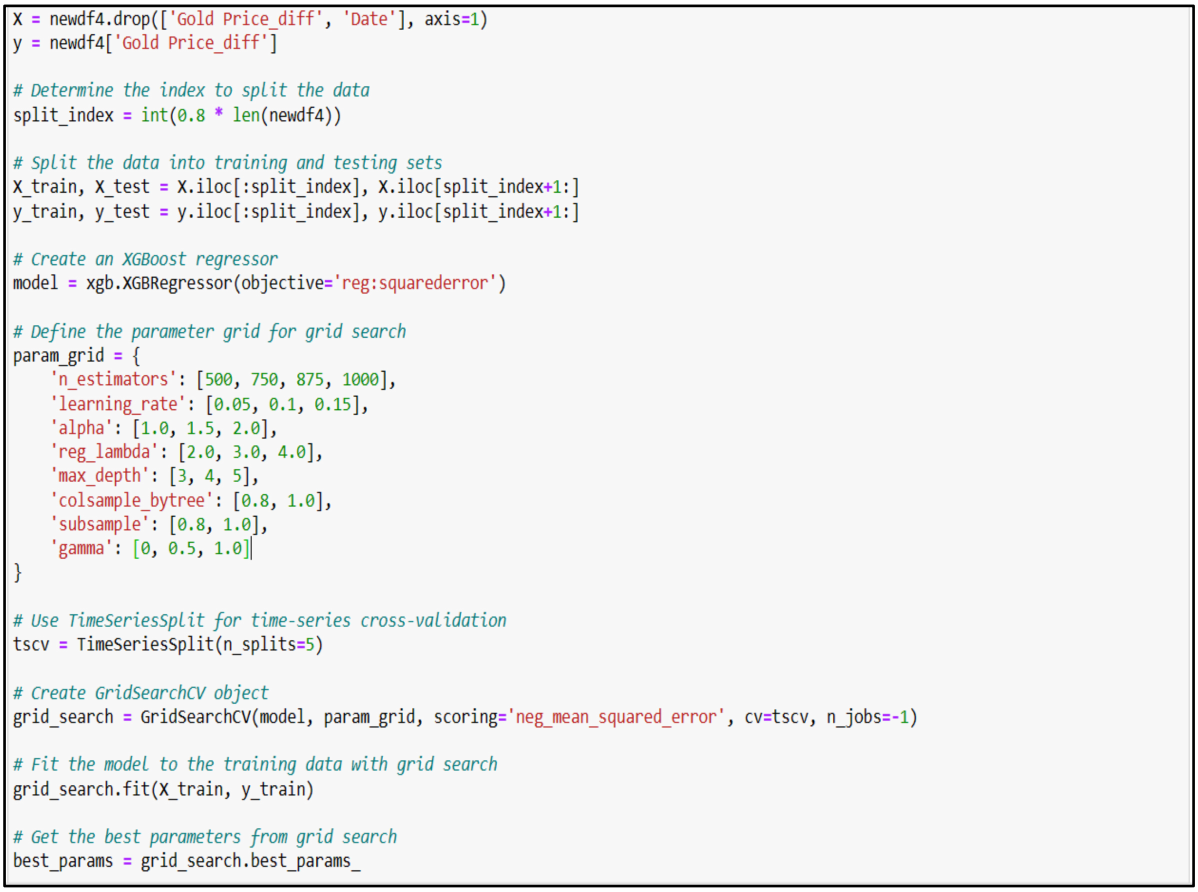
```
#Differencing to make the series stationary

newdf['Gold Price_diff']=newdf['Gold Price'].diff()
newdf['USD LKR Exchange rate_diff']=newdf['USD LKR Exchange rate'].diff()
newdf['CNY LKR Exchange rate_diff']=newdf['CNY LKR Exchange rate'].diff()
newdf['Brent Crude oil price USD_diff']=newdf['Brent Crude oil price USD'].diff()
newdf['Silver Price_diff']=newdf['Silver Price'].diff()
newdf['CCIP_diff']=newdf['CCIP'].diff()
newdf['S&P 20_diff']=newdf['S&P 20'].diff()
newdf['Gold Reserves_diff']=newdf['Gold Reserves'].diff()
```

Appendix Figure B-3: Sample Code for Differencing of the Data

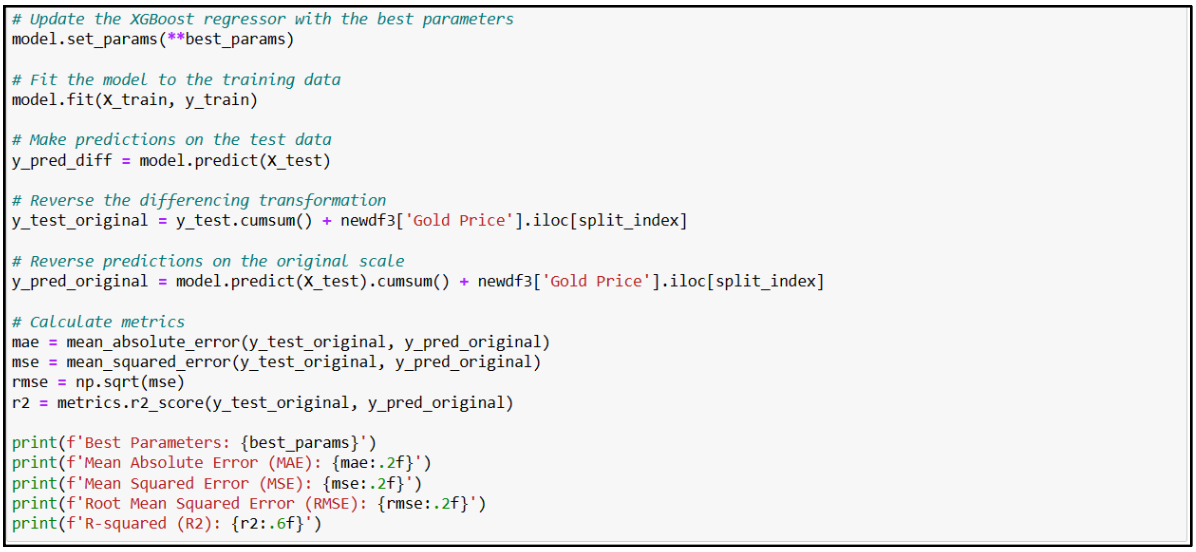


```
newdf2['Crude_Lag1'] = newdf2['Brent Crude oil price USD_diff'].shift(1)
newdf2['Crude_Lag2'] = newdf2['Brent Crude oil price USD_diff'].shift(2)
newdf2['Crude_Lag3'] = newdf2['Brent Crude oil price USD_diff'].shift(3)

newdf2['CCIP_Lag1'] = newdf2['CCIP_diff'].shift(1)
newdf2['CCIP_Lag3'] = newdf2['CCIP_diff'].shift(3)
newdf2['CCIP_Lag4'] = newdf2['CCIP_diff'].shift(4)

newdf2['CNY_Lag1'] = newdf2['CNY LKR Exchange rate_diff'].shift(1)
newdf2['CNY_Lag2'] = newdf2['CNY LKR Exchange rate_diff'].shift(2)
newdf2['CNY_Lag7'] = newdf2['CNY LKR Exchange rate_diff'].shift(7)
newdf2['CNY_Lag14'] = newdf2['CNY LKR Exchange rate_diff'].shift(14)

newdf2['S&P_Lag1'] = newdf2['S&P 20_diff'].shift(1)
newdf2['S&P_Lag2'] = newdf2['S&P 20_diff'].shift(2)
newdf2['S&P_Lag7'] = newdf2['S&P 20_diff'].shift(7)
newdf2['S&P_Lag21'] = newdf2['S&P 20_diff'].shift(21)

newdf2['Silver_Lag1'] = newdf2['Silver Price_diff'].shift(1)

newdf2['USD_Lag1'] = newdf2['USD LKR Exchange rate_diff'].shift(1)
newdf2['USD_Lag2'] = newdf2['USD LKR Exchange rate_diff'].shift(2)
newdf2['USD_Lag7'] = newdf2['USD LKR Exchange rate_diff'].shift(7)
newdf2['USD_Lag14'] = newdf2['USD LKR Exchange rate_diff'].shift(14)

newdf2['Gold_Lag1'] = newdf2['Gold Price_diff'].shift(1)
newdf2['Gold_Lag2'] = newdf2['Gold Price_diff'].shift(2)
```

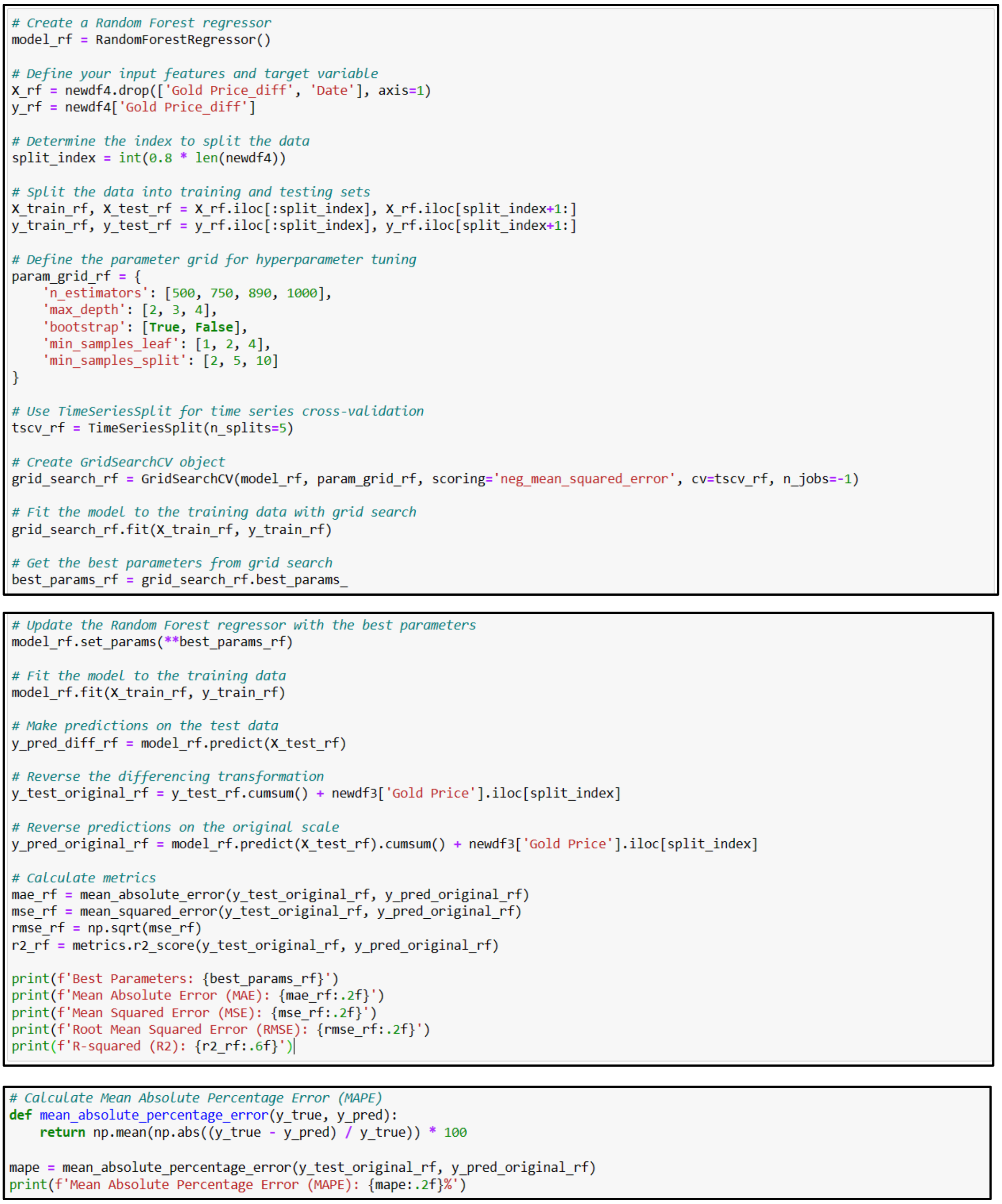
Appendix Figure B-4: Sample Code for Creating the Log Variables



```
X = newdf4.drop(['Gold Price_diff', 'Date'], axis=1)
y = newdf4['Gold Price_diff']

# Determine the index to split the data
split_index = int(0.8 * len(newdf4))

# Split the data into training and testing sets
X_train, X_test = X.iloc[:split_index], X.iloc[split_index+1:]
y_train, y_test = y.iloc[:split_index], y.iloc[split_index+1:]

# Create an XGBoost regressor
model = xgb.XGBRegressor(objective='reg:squarederror')

# Define the parameter grid for grid search
param_grid = {
    'n_estimators': [500, 750, 875, 1000],
    'learning_rate': [0.05, 0.1, 0.15],
    'alpha': [1.0, 1.5, 2.0],
    'reg_lambda': [2.0, 3.0, 4.0],
    'max_depth': [3, 4, 5],
    'colsample_bytree': [0.8, 1.0],
    'subsample': [0.8, 1.0],
    'gamma': [0, 0.5, 1.0]
}

# Use TimeSeriesSplit for time-series cross-validation
tscv = TimeSeriesSplit(n_splits=5)

# Create GridSearchCV object
grid_search = GridSearchCV(model, param_grid, scoring='neg_mean_squared_error', cv=tscv, n_jobs=-1)

# Fit the model to the training data with grid search
grid_search.fit(X_train, y_train)

# Get the best parameters from grid search
best_params = grid_search.best_params_
```



```
# Update the XGBoost regressor with the best parameters
model.set_params(**best_params)

# Fit the model to the training data
model.fit(X_train, y_train)

# Make predictions on the test data
y_pred_diff = model.predict(X_test)

# Reverse the differencing transformation
y_test_original = y_test.cumsum() + newdf3['Gold Price'].iloc[split_index]

# Reverse predictions on the original scale
y_pred_original = model.predict(X_test).cumsum() + newdf3['Gold Price'].iloc[split_index]

# Calculate metrics
mae = mean_absolute_error(y_test_original, y_pred_original)
mse = mean_squared_error(y_test_original, y_pred_original)
rmse = np.sqrt(mse)
r2 = metrics.r2_score(y_test_original, y_pred_original)

print(f'Best Parameters: {best_params}')
print(f'Mean Absolute Error (MAE): {mae:.2f}')
print(f'Mean Squared Error (MSE): {mse:.2f}')
print(f'Root Mean Squared Error (RMSE): {rmse:.2f}')
print(f'R-squared (R2): {r2:.6f}')
```



```python
# Create a Random Forest regressor
model_rf = RandomForestRegressor()

# Define your input features and target variable
X_rf = newdf4.drop(['Gold Price_diff', 'Date'], axis=1)
y_rf = newdf4['Gold Price_diff']

# Determine the index to split the data
split_index = int(0.8 * len(newdf4))

# Split the data into training and testing sets
X_train_rf, X_test_rf = X_rf.iloc[:split_index], X_rf.iloc[split_index+1:]
y_train_rf, y_test_rf = y_rf.iloc[:split_index], y_rf.iloc[split_index+1:]

# Define the parameter grid for hyperparameter tuning
param_grid_rf = {
    'n_estimators': [500, 750, 890, 1000],
    'max_depth': [2, 3, 4],
    'bootstrap': [True, False],
    'min_samples_leaf': [1, 2, 4],
    'min_samples_split': [2, 5, 10]
}

# Use TimeSeriesSplit for time series cross-validation
tscv_rf = TimeSeriesSplit(n_splits=5)

# Create GridSearchCV object
grid_search_rf = GridSearchCV(model_rf, param_grid_rf, scoring='neg_mean_squared_error', cv=tscv_rf, n_jobs=-1)

# Fit the model to the training data with grid search
grid_search_rf.fit(X_train_rf, y_train_rf)

# Get the best parameters from grid search
best_params_rf = grid_search_rf.best_params_
```

```python
# Update the Random Forest regressor with the best parameters
model_rf.set_params(**best_params_rf)

# Fit the model to the training data
model_rf.fit(X_train_rf, y_train_rf)

# Make predictions on the test data
y_pred_diff_rf = model_rf.predict(X_test_rf)

# Reverse the differencing transformation
y_test_original_rf = y_test_rf.cumsum() + newdf3['Gold Price'].iloc[split_index]

# Reverse predictions on the original scale
y_pred_original_rf = model_rf.predict(X_test_rf).cumsum() + newdf3['Gold Price'].iloc[split_index]

# Calculate metrics
mae_rf = mean_absolute_error(y_test_original_rf, y_pred_original_rf)
mse_rf = mean_squared_error(y_test_original_rf, y_pred_original_rf)
rmse_rf = np.sqrt(mse_rf)
r2_rf = metrics.r2_score(y_test_original_rf, y_pred_original_rf)

print(f'Best Parameters: {best_params_rf}')
print(f'Mean Absolute Error (MAE): {mae_rf:.2f}')
print(f'Mean Squared Error (MSE): {mse_rf:.2f}')
print(f'Root Mean Squared Error (RMSE): {rmse_rf:.2f}')
print(f'R-squared (R2): {r2_rf:.6f}')
```

```python
# Calculate Mean Absolute Percentage Error (MAPE)
def mean_absolute_percentage_error(y_true, y_pred):
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

mape = mean_absolute_percentage_error(y_test_original_rf, y_pred_original_rf)
print(f'Mean Absolute Percentage Error (MAPE): {mape:.2f}%')
```

Appendix Figure B-5: Sample Code for Modeling XGBoost and Random Forest

REFERENCES

[1] “The History of Gold as a Currency and Store of Value Throughout Human Civilization.” Accessed: Aug. 01, 2023. [Online]. Available: https://www.atfx.com/en/analysis/trading-strategies/the-history-of-gold-as-a-currency-and-store-of-value-throughout-human-civilization

[2] “When and Why Do Gold Prices Plummet?” Accessed: Aug. 01, 2023. [Online]. Available: https://www.investopedia.com/articles/investing/071414/when-and-why-do-gold-prices-plummet.asp

[3] “What Is the Gold Standard? Advantages, Alternatives, and History.” Accessed: Aug. 01, 2023. [Online]. Available: https://www.investopedia.com/ask/answers/09/gold-standard.asp

[4] “Gold Price Prediction using Machine Learning - Javatpoint.” Accessed: Aug. 02, 2023. [Online]. Available: https://www.javatpoint.com/gold-price-prediction-using-machine-learning

[5] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, “Forecasting gold price with the XGBoost algorithm and SHAP interaction values,” *Ann Oper Res*, 2021, doi: 10.1007/s10479-021-04187-w.

[6] P. Zhang and B. Ci, “Deep belief network for gold price forecasting,” *Resources Policy*, vol. 69, Dec. 2020, doi: 10.1016/j.resourpol.2020.101806.

[7] D. Liu and Z. Li, “Gold price forecasting and related influence factors analysis based on random forest,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2017, pp. 711–723. doi: 10.1007/978-981-10-1837-4_59.

[8] D. Makala and Z. Li, “Prediction of gold price with ARIMA and SVM,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1767/1/012022.

[9] I. ul and K. Nazir, “Predicting Future Gold Rates using Machine Learning Approach,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, 2017, doi: 10.14569/ijacsa.2017.081213.

[10] M. YURTSEVER, “Gold Price Forecasting Using LSTM, Bi-LSTM and GRU,” *European Journal of Science and Technology,* Dec. 2021, doi: 10.31590/ejosat.959405.

[11] Manjula and Karthikeyan, *Gold Price Prediction using Ensemble based Machine Learning Techniques*.

[12] I. E. Livieris, E. Pintelas, and P. Pintelas, “A CNN–LSTM model for gold price time-series forecasting,” *Neural Comput Appl*, vol. 32, no. 23, pp. 17351–17360, Dec. 2020, doi: 10.1007/s00521-020-04867-x.

[13] S. Patalay, “Gold Price Prediction Using Machine Learning Model Trees.” [Online]. Available: https://www.researchgate.net/publication/366000728

[14] Z. Alameer, M. A. Elaziz, A. A. Ewees, H. Ye, and Z. Jianhua, “Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm,” *Resources Policy*, vol. 61, pp. 250–260, Jun. 2019, doi: 10.1016/j.resourpol.2019.02.014.

[15] D. N. Gono, H. Napitupulu, and Firdaniza, “Silver Price Forecasting Using Extreme Gradient Boosting (XGBoost) Method,” *Mathematics*, vol. 11, no. 18, Sep. 2023, doi: 10.3390/math11183813.

[16] G. Astudillo, R. Carrasco, C. Fernández-Campusano, and M. Chacón, “Copper price prediction using support vector regression technique,” *Applied Sciences (Switzerland)*, vol. 10, no. 19, Oct. 2020, doi: 10.3390/APP10196648.

[17] C. Pierdzioch and M. Risse, “Forecasting precious metal returns with multivariate random forests,” *Empir Econ*, vol. 58, no. 3, pp. 1167–1184, Mar. 2020, doi: 10.1007/s00181-018-1558-9.

[18] S. Urolagin, N. Sharma, and T. K. Datta, “A combined architecture of multivariate LSTM with Mahalanobis and Z-Score transformations for oil price forecasting,” *Energy*, vol. 231, Sep. 2021, doi: 10.1016/j.energy.2021.120963.

[19] M. Appuhamilage and K. Sriyalatha, “Does the All Share Price Index represent the Colombo Stock Market ?”

[20] “XGBoost for Time Series extrapolation: You’re gonna need a bigger boat | by Saupin Guillaume | Towards Data Science.” Accessed: Dec. 20, 2023. [Online]. Available: https://towardsdatascience.com/xgboost-for-time-series-youre-gonna-need-a-bigger-boat-9d329efa6814

[21] “What is Machine Learning? Definition, Types, Tools & More | DataCamp.” Accessed: Nov. 04, 2023. [Online]. Available: https://www.datacamp.com/blog/what-is-machine-learning

[22] “XGBoost - GeeksforGeeks.” Accessed: Nov. 05, 2023. [Online]. Available: https://www.geeksforgeeks.org/xgboost/

[23] “Random Forest Regression in Python - GeeksforGeeks.” Accessed: Dec. 22, 2023. [Online]. Available: https://www.geeksforgeeks.org/random-forest-regression-in-python/

[24] “Random Forest Regression in Python Explained | Built In.” Accessed: Dec. 21, 2023. [Online]. Available: https://builtin.com/data-science/random-forest-python

[25] “GridSearchCV |Tune Hyperparameters with GridSearchCV.” Accessed: Nov. 05, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/