Future Cocoa Price Prediction

Peush Gomes, Yiyi Li, and Ashka Shah

2025-04-04

Introduction

Cocoa beans are an enormously essential agricultural crop worldwide, with the global demand for cocoa beans increasing by nearly a million tons in just ten years. Countries in West Africa, specifically Ghana and Cote d'Ivoire, are responsible for nearly 70% of the world's cocoa production (Wessel and Quist-Wessel, 2015), with cocoa being the largest export from Ghana to the global market. Cocoa requires a specific humid climate and temperature to grow and is extremely susceptible to disease and pests, thus the amount of cocoa that cocoa farmers can grow each year is heavily affected by any environmental changes that may occur. Due to the risky nature of cocoa production, being able to predict how the cocoa prices may vary across a time period is incredibly important for cocoa farmers (Sukiyono et al., 2018) and the overall economies of countries that rely heavily on cocoa production.

Historically, the times series method has been used to forecast future prices for various agricultural commodities, including sugar, grain, and cotton (Ouyang et al. 2019). The objective of this project is to use various time series and machine learning methods to model commodity prices for cocoa, as well as forecast future cocoa prices. For our study, we will be using two main datasets: the Cocoa Futures Prices dataset from the International Cocoa Organization (ICCO) that contains cocoa exchange prices from 1994 to 2025, and the Ghana Climate dataset from the National Centers for Environmental Information (NCEI), which contains the daily precipitation and temperature levels in Ghana. Because currency exchange rates may heavily influence the contracts regarding the prices that cocoa beans are sold at, we also have used an additional Ghana-US Exchange Rate dataset to track exchange rates from 1990 to 2024.

In order to determine which time series model works the best at forecasting cocoa prices, we will first look at previous research methods used to forecast commodity prices, and determine which ones we will apply to our own data. From that, we will test four different models to see which one provides the most accurate prediction of future cocoa prices using Akaike Information Criterion (AIC) and Root Mean Squared Error (RMSE) values. Using the model we deem most fit, we will then present our predictions for future cocoa prices.

As there are many different models that have attempted to measure cocoa prices using a variety of statistical methods, the key challenge for this study is to find a model that is able to provide a reasonable prediction for future cocoa prices that builds upon other models, and identify what the most significant factors are in influencing future cocoa prices. Furthermore, the highly volatile nature of cocoa market prices and the unpredictable nature of today's climate due to various factors including climate change may all affect how we can fit a model and predict cocoa prices based on provided datasets.

Literature Review

Modelling the price fluctuations of commodities has been widely studied, particularly in financial markets. Future prices and their fluctuations rely on historical data as much as they rely on current events, making Time Series models an important tool in this investigation. The Autoregressive Integrated Moving Average (ARIMA) model is a simple statistical model that uses past values (autoregressive component) and past errors (moving average component) to forecast, allowing it to capture short-term linear dependencies between future and past values quite effectively. There is an abundance of literature that has researched the use of univariate ARIMA models to forecast commodity prices, including annual cocoa production predictions in Nigeria from 2019-2025 (Oni et al, 2021) and cocoa bean price predictions in Indonesia from 2008-2016 (Sukiyono et al., 2018). Specifically, research by Sukiyono et al. found that ARIMA models have worked better than other time series models like Exponential Smoothing State Space Models (ETS) and Decomposition models. Yet the ARIMA model is limited in its utility as it requires the time series to be weakly stationary, where the mean and variance are constant, and its autocovariance depends only on the lag between two time points, rather than the specific time at which the data is observed (Shumway and Stoffer, 2017).

Cocoa prices have been known to exhibit seasonality that could be due to weather patterns or market cycles (Geman and Sarfo, 2012). Unfortunately, ARIMA models are unable to capture seasonality. When seasonality needs to be captured, an extension of the ARIMA model called Seasonal ARIMA (SARIMA) is used (Shumway and Stoffer, 2017). The SARIMA model includes additional seasonal differencing, autoregressive and moving average parameters making it suitable for capturing seasonality. It must be noted that the SARIMA model also requires that the time series be stationary. The traditional ARIMA/SARIMA models assume that future prices solely depend on past observations, but this is not always the case. While current and future cocoa prices are influenced by historical cocoa prices, they are also influenced by many exogenous factors including weather, supply, demand, currency exchange rates etc. Studies have shown that including macroeconomic indicators (Ye et al., 2021), weather conditions (Zhang et al., 2024) have resulted in improved commodity price forecasts. This is where SARIMAX models come in. They can incorporate exogenous variables in the model to enhance forecasting.

Apart from the time series model, classic regression models like Multiple Linear Regression with Lagged Variables are some of the simplest forecasting models. Here, the dependent variable is modelled as a function of explanatory variables like weather, exchange rate, supply etc. Regression models are easy to interpret and are widely used in forecasting financial and economic data. Yet regression models can only capture linear relationships between the dependent variable and the explanatory variables, limiting applicability to more complex relationships (Shumway and Stoffer, 2017). They can still be used as baseline models alongside other machine learning and time series models to understand relative performance.

ARIMA/SARIMA and linear regression models are good at capturing short-term linear dependencies, but may not be ideal when it comes to more complex situations. Machine learning techniques such as Random Forests (RF) have addressed the limitation of linear models and have gained popularity in financial forecasting. Random Forests have outperformed ARIMA models in cases where external factors were incorporated (Zhao, 2023). Unlike the ARIMA/SARIMA models, Random forests do not assume stationarity or distribution of the data. It is an ensemble model that uses a bunch of uncorrelated decision trees to make a prediction (Zhao, 2023). However, Random Forests are black boxes and can be very hard to interpret requiring extensive validation.

Each of these forecasting methods have their own set of advantages and limitations. Many studies have applied them to predict cocoa production, cocoa prices, and other commodity or stock prices. This study aims to evaluate and compare the performances of ARIMA, SARIMAX, Multiple Linear Regression, and Random Forest in the context of Cocoa Futures in order to choose the best possible model. Additionally, by incorporating external variables of temperature, precipitation and exchange rates, we aim to understand their impact on Cocoa Futures. Our approach differs from current literature because our aim is to both focus on the variety of factors that affect cocoa prices, but also compare different models to choose the best possible model for cocoa futures prices in Ghana. Through this study, we aim to contribute to the knowledge of climatic and economic factors that can affect Cocoa Futures prediction with a broader applicability to commodities price forecasting.

Data

For the purposes of modeling cocoa futures prices we used three datasets. Two of the datasets were already provided to us by the instructors of STA457H/STA2202H, while one dataset was externally sourced by our team. There are multiple factors that can affect the price of cocoa. Thus, we tried to find external datasets to include those factors in the models for Cocoa Futures Price Prediction. After all our search and data considerations we decided to use one external dataset on the Cedi (Ghana currency) to USD exchange rate from the International Monetary Fund. Ghana and Ivory Coast are the top producers of cocoa beans in the world followed by Nigeria. Between 1978 and 1982 an upsurge in exchange rates reduced the quantity of cocoa exports (Uduh, 2017). Literature suggests that exchange rate volatility can negatively affect production activities and export activities Uduh, 2017). Thus, Cedi exchange rates become important to consider for modelling Cocoa prices.

There were two other datasets from Food and Agriculture Organization of the United Nations we wanted to incorporate (1) Production, Yield and Harvest data for Cocoa Beans in Ghana[1], to model the supply of cocoa in Ghana (2) Cocoa Beans Import in the World[2], to model the demand of cocoa in the world. Unfortunately, we were unable to use data from these sets due to two reasons - (i) These datasets only had yearly data, while our other datasets have at least monthly data. We would have to convert all our series to monthly and we would miss out on any seasonal trends in other series. (ii) These datasets only had data until 2023. While all the other series carry on till at least 2024. The datasets used for this project are as follows:

1. Cocoa Futures Price Dataset - This dataset provided by the instructors consists of daily Closing Prices for Cocoa Futures contracts and was obtained from the International Cocoa Organization (ICCO). It consists of data from Oct 2, 1994 to Feb 27, 2025. This dataset is stored under the file name `Cocoa_Daily_Prices.csv.`
2. Ghana Climate Dataset - This dataset provided by the instructors consists of daily Temperature and Precipitation in Ghana and was obtained from the National Centers for Environmental Information (NCEI). It consists of data from Jan 1, 1990 to Nov, 28, 2024. This dataset is stored under the file name `Ghana_Climate_Data.csv`.
3. Ghana Currency Exchange Rate Dataset - This is the dataset we sourced externally which consists of monthly Average Exchange Rate values for Cedi to USD obtained from the International Monetary Fund[3]. It ranges from Jan 1990 to Sep 2024. The data is stored under the filename `Ghana_Exchange_Rates.csv`.

Methodology

Dataset Processing

All datasets were converted into monthly intervals as traditional time series models like SARIMA, GARCH etc assume equally spaced time series. Missing values were handled according to the context of each dataset. The data cleaning and processing steps for each time series are outlined below:

1. **Cocoa Futures Price Dataset** - This series consisted of daily data. It had duplicate prices for 2023-12-15 and 2024-01-09 which were removed. It had two price values for 2024-01-30 and 2024-01-31, so we kept the price that was closer to the prices around the neighbouring dates. All the prices were then grouped by *month* and *year* and averaged. As a result, the daily time series dataset became a monthly time series dataset.
2. **Ghana Currency Exchange Rate Dataset** - This series was already monthly and had no duplicates. There were 4 different types of exchange rates in this series but we only considered the Average Period Exchange Rate for the Domestic Currency (Cedi) per U.S. Dollar.
3. **Ghana Climate Dataset** - This series consisted of daily data. A lot of dates had 'NA' values for precipitation, so we replaced them with 0 (given in project description). It had temperature and precipitation recorded for different locations in Ghana for the same dates. To accommodate for this kind of duplication, we averaged out all the temperatures and precipitations for a particular date over

[1] The data can be downloaded from https://www.fao.org/faostat/en/#data/QCL
[2] The data can be downloaded from https://www.fao.org/faostat/en/#data/TCL
[3] The data can be downloaded from https://data.imf.org/regular.aspx?key=61545850

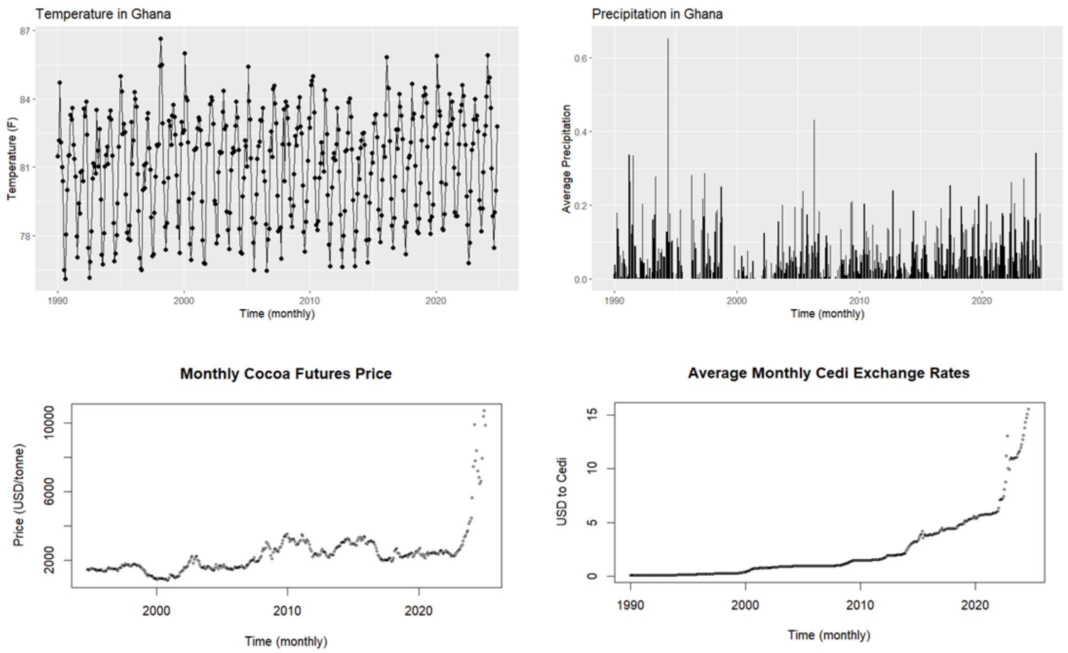

Figure 1: The plotted graphs illustrating the characteristics of the time series once cleaned and converted to monthly intervals (1) Temperature series (top-left), (2) Precipitation series (top-right), (3) Cocoa Futures Price series (bottom-left) and (4) Cedi Exchange Rate series (bottom-right).

different locations. The temperature and precipitation data was then again grouped by *month* and *year* and averaged. As a result, the daily time series data became a monthly time series data set. There were 3 months missing temperature and precipitation data namely December 1994, October 2001 and December 2001. Since only 3 data points were missing we ended up using linear interpolation to impute the missing values.

Lastly, the clean data set was split into train and validation sets depending on the model being considered. For time series and regression models, the last 4 data points were used for validation/forecast while all the other previous points were used for training.

Exploratory Data Analysis

At the end of the Data Processing phase, there were four monthly series in the dataset (1) Cocoa Futures Price Series (2) Exchange Rate Series (3) Temperature Series (4) Precipitation Series, all of which were non-stationary. Since the time series models assume stationarity, we took the log of the future prices and exchange rate series to account for heteroskedasticity, and applied a first-order difference with a lag of 1 to achieve stationarity. For the temperature and precipitation series, the first order difference was taken with a lag of 12 to achieve stationarity as the ACF's of both series showed seasonality that repeated after lag 12. To confirm stationarity, we generated the ACFs and PACFs for all the series and saw that the ACF's were indeed tailing off. The ACF of future prices series was significant at lag 1 and lag 7, at lag 12 for precipitation and temperature and lag 1 for exchange rates. Since we planned to use temperature, precipitation, and exchange rates as exogenous/explanatory variables in cocoa futures price modeling, we generated CCFs (cross-correlation functions) for each of them. The CCF plots for future prices and temperature, future prices and precipitation and future prices and exchange rates had significant lags at 24, 22 and 17 respectively. The SARIMAX, Random Forests and Multiple Linear Regression models used the lags from the ACFs and CCFs as exogenous/explanatory variables.

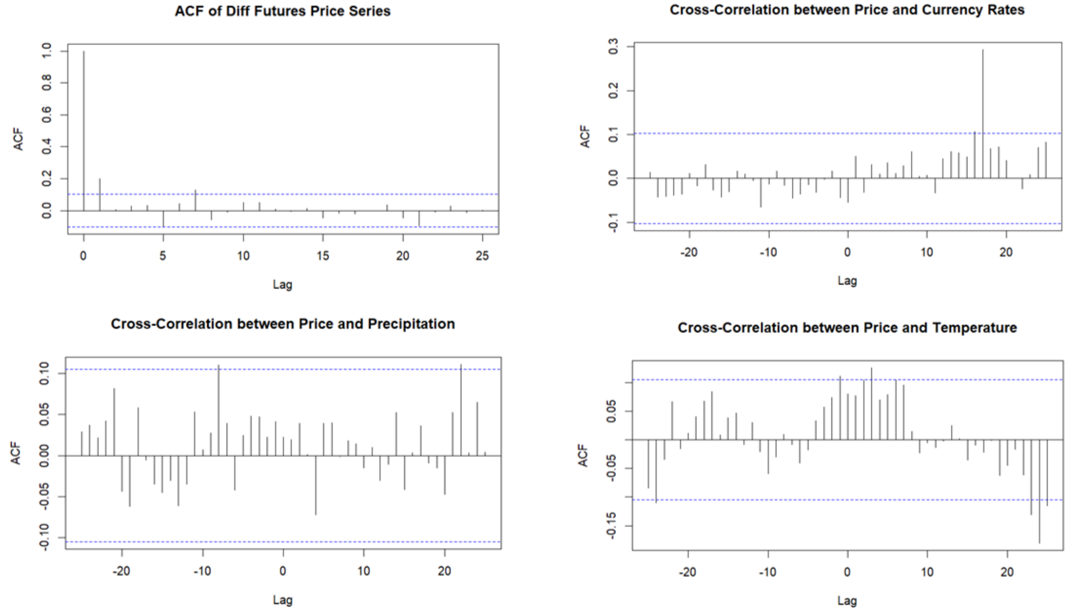

Figure 2: The ACF of Cocoa Future Price (top-left) shows significant value at lag 1 and 7. The CCF of Cocoa Future Price and Exchange Rates shows a significant value at lag 16 and 17, while both can be considered the most significant lag i.e. 17 was used for modelling. The CCF of Cocoa Future Price and Temperature shows a significant value at lag 3,23 and 14, while all can be considered the most significant lag i.e. 24 was used for modelling. The CCF of Cocoa Future Price and Precipitation only shows a significant value at lag 22. Cross-correlation Functions were to investigate how changes in Temperature, Precipitation and Exchange Rates affect Cocoa Future Prices up to lag 30, thus negative lags were not considered.

Modelling

1. Univariate ARIMA model

The first model we chose to compare was the univariate ARIMA model based solely on the cocoa futures price dataset. In order to find the p and q values for the ARIMA model while only looking at the price dataset, we analyzed the ACF/PACF plots to find that the ACF model cut off after lag 1 and the PACF model also cut off after lag 1. As a result, we tried fitting an ARIMA(0,1,1) model and an ARIMA(1,1,0) model to see which one had the lowest AIC value.

2. Multiple Linear Regression

Using the lags mentioned in the exploratory data analysis, we also fit a Multiple Linear Regression model for cocoa future prices prediction. The differencing of the temperature and precipitation time series with a lag of 12 led to a loss of 11 months worth of initial data. Since the largest lag from all series was 24, the lag transformations resulted in the loss of another 24 months worth of data. Thus, we removed 35 months of initial data to ensure that no NA values were present in our dataset. While the cleaned dataset had data from Oct 1994, this model was only trained on data starting from October 1996. The model performance was then validated using the last 4 data points in the series. The metrics used to compare this model with others were AIC and RMSE.

3. Random Forests

The random forest model was fitted with the use of lagged cocoa prices from the initial dataset, its variables of prediction were dictated by the exogenous variables that were used in the prior models, which were found by ACF/PACF and residual analysis. As a random forest model needs to be trained and tested upon, the time series data set was tested on the last 4 data points and was trained on the rest. After fitting the random forest model on the test set, we compared this RMSE value to the RMSE of the chosen ARIMA model.

4. SARIMA Model

This model was created off the basis of the ARIMA models, through ACF/PACF and residual analysis to determine the values of p, q, d, P, Q, D, and S. By analyzing the ACF/PACF plots of each of the series, we concluded that the best model to use was the SARIMA (0,1,1) x (0,1,1)12, as it resulted in the best case scenario for fit as seen through its AIC. To increase the validity and fit of the SARIMA model, we included the use of exogenous variables of average precipitation, temperature, and exchange rate. Although we attempted to fit another model with all the same significant lags, it yielded a very similar result. Finally, the validity of our model was concluded through residual analysis and calculating RMSE by taking our actual values and re-forecasting the last four months of the dataset.

Forecasting and Results

When it came to forecasting and assessing the results of different models on future cocoa prices, we had implemented, and evaluated different approaches. We did forecasting for cocoa prices with time series analysis via the prediction of a Seasonal ARIMA model, a classical regression model, and a machine learning Random Forest model

ARIMA Models

A few different ARIMA models were inspected and applied for forecasting. Each model was modeled and selected based on diagnosis of the time series' ACF and PACFs and observations of the residual diagnostics. While the first ARIMA models tested were univariate, and the final SARIMA model included included exogenous variables such as average precipitation, temperature, and exchange rate.

Table 1: Model vs AIC

Model	AIC
ARIMA (0,1,1)	-989.4392
ARIMA (1,1,0)	-989.8947
SARIMA (0,1,1) x (0,1,1)[12]	-2.650322

To conclude which of these models would be of best fit for our forecasting purpose, AIC was used for comparison. Although the univariate ARIMA models had a smaller AIC values, we realized that it was important for the model to be able to predict based on multiple variables, and not just the price time series dataset. The SARIMA model worked the best when looking at all the factors going into determining which model we should select. After considering first differencing, ACF/PACF analysis, and residual diagnosis, it was concluded the final SARIMA model would be that of SARIMA(0,1,1)(0,1,1)[12]. This model included the exogenous regressors of average precipitation, temperature, and exchange rate. (It is to be noted that the same SARIMA model with exogenous lag variables was also considered however yielded a similar result).

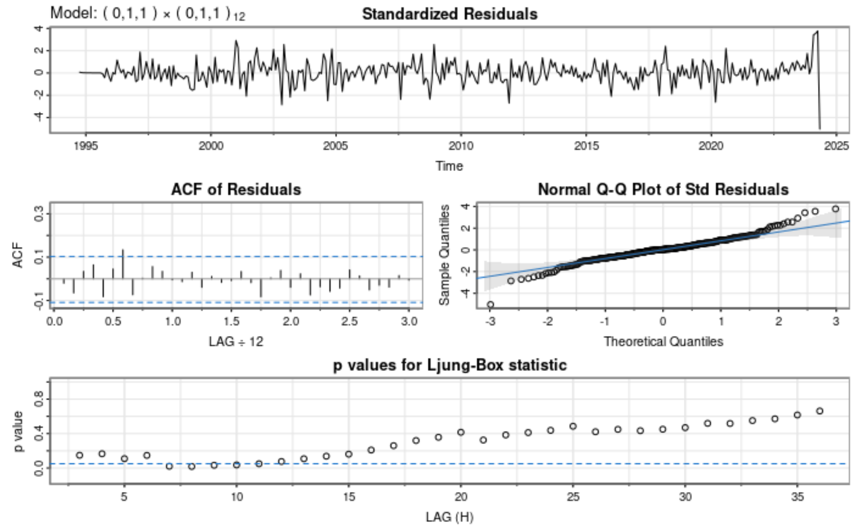

Figure 3: Model

As we can see from Figure 3, the residual diagnosis plot for the modeled sarima model illustrates that the standardized residuals are relatively stable over time. The ACF of residuals stays within the bounds and thus no significant autocorrelations, and the Ljung-Box test p-values do exceed the 0.05 threshold for the most part. Signifying that the residuals are not significantly different from white noise. Thus an adequate time series model. When compairing this model's AIC with the other explored variations that we had modeled earlier, the current model produces an AIC of -2.650322, reinforcing that this will be our sarima model.

Random Forest Model

We have also developed a Random Forest regression model to forecast the cocoa prices. This model was created using lagged prices and the same exogenous variables as the sarima model (precipitation, temperature, and exchange rate). A single lag of the prices was used in conjunction with the exogenous variables as the predictors for the random forest. Trained and tested upon the dataset that was split 80/20, where the value the performance of this model we will consider the RMSE.

Forecast Test Compairson

To compare the SARIMA and the Random Forest models together, we will consider RMSE along side actual values vs test forecasting of the last four months of our time series data set.

Table 2: SARIMA vs Forest RMSE

Model	RMSE
SARIMA	769.4567
Random Forest	5032.889
Multiple Linear Regression	1367.60524881816

As we can see in Table 2: The RMSE of the Random Forest is actually less than that of the SARIMA model, thus meaning the model is of greater fit. However, when we actually model the actual values vs the forecasted values of both models we get an interesting result.

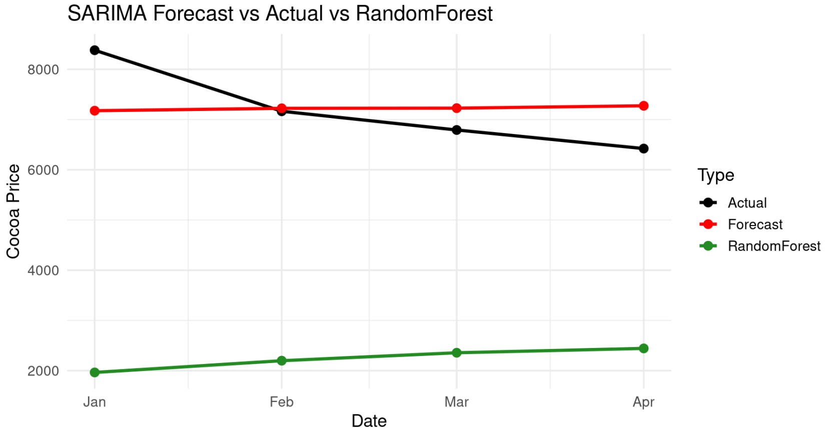

Figure 4: Forecasting vs Actual (Last 4 Months)

As we can see in Figure 4, when we compare the forecasted values given to us by our sarima model and random forest, and we compare it to the actual values of the last four months. We can see that the SARIMA with the lower RMSE gives us a better more closely tracked forecasted values to the actual cocoa prices than that of our random forest model. It is possible that the Random Forest while having an worse RMSE is potentially over fitting the earlier data of the time series. As we know the greatest change in actual cocoa prices was more 'recent' thus the forest was more inline with the trend of the earlier data rather than the important latter values.

Both models demonstrated reasonable forecasting performance, however, the SARIMA model provided better interpretability and less over fitting to the earlier data. Thus we move forward from these two with the SARIMA model.

Multiple Linear Regression Model

In addition to the time series sarima models and the machine learning random forest model that we had fitted, we have also implemented a multiple linear regression model. This model used a combination of predictors of lagged variables, those being price lags, currency rate lag, temperature lag, and precipitation lag. Diagnostic plots and summaries had indicated that several of these predictors were significant, leaving out that of the precipitation lag predictor. For this model, we had forecasted the last 5 months of actual data to see if the model was of best fit, to calculate performance we also look at RMSE for this case. However, looking back to Table 2, we can see that this linear regression model had produced a RMSE number of 1367.60524881816.

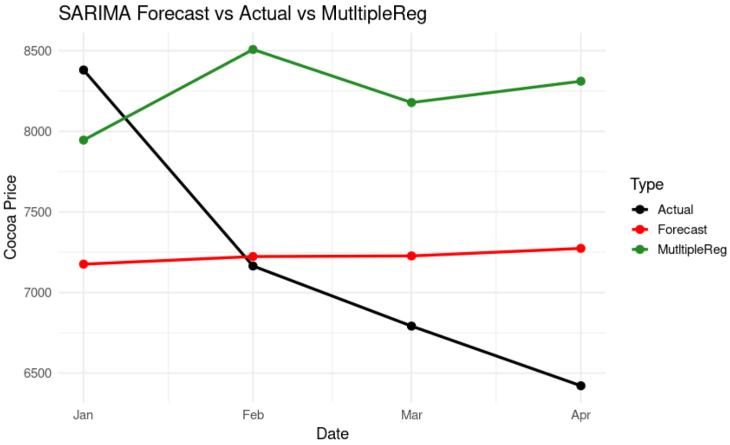

Figure 5: Forecasting vs Actual vs MLR (Last 4 Months)

As observed from Figure 5, we can see that there is a difference between the forecasted values of the time series by the multiple linear regression model and that of the actual data. With the low RMSE, large distance from actual and forecasted values, as well as that it seems that the forecasted values almost begin to oscillate. We lean towards using the SARIMA model for our proper forecasting.

Forecasting Results

As we have concluded with our forecasting model to be the SARIMA(0,1,1)(0,1,1)[12] model, we then forecast the next four months of prices based on our time series data. As we can see in figure 6, the next four months have been forecasted with an somewhat upwards trend. Should be taken into account that we are forecasting the log prices here as we had taken log of the prices when created the time series data set.

Table 3: Forecasted Values

Month	Forecasted Value
October (2024)	6242.38
November (2024)	6211.10
December (2024)	6285.50
January (2025)	6362.22

Table 3 gives us the real forecasted values, after converting them back from our log prices. The next four forecasted months show a positive trend in price.

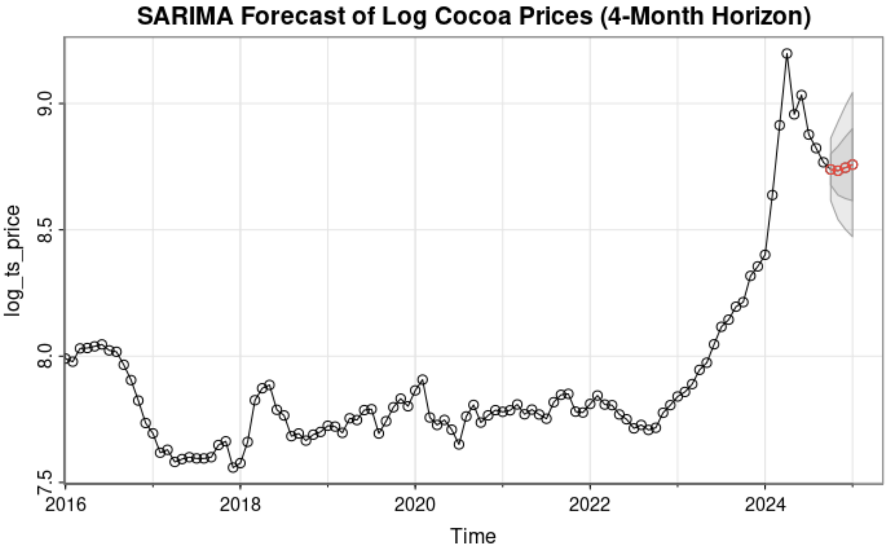

Figure 6: Forecasting

Discussions and Conclusions

Discussion

Although it had a larger RMSE value than the random forest model, the SARIMA (0,1,1) x (0,1,1)12 model ended up being the best candidate to predict future cocoa prices. It gave us better forecasted cocoa price values as compared to the random forest model, as shown in Figure 4. Looking at our forecasted cocoa future prices in Figure 6, we can see that cocoa prices for the end of 2024 are predicted to slowly climb upwards, though they have declined from their peak prices in April 2024. Nonetheless, as prices have still increased substantially from just 3-4 years ago, this may imply that cocoa prices will continue to increase in the future due to high global demand and decreased production. Additionally, if our climate continues to worsen and become more volatile while demand for cocoa only continues to increase, this could lead to even higher prices as cocoa bean production becomes more difficult and cocoa becomes more scarce.

Limitations

When creating our models, we also recognized several limitations in our approach that could affect the accuracy of our model. One limitation comes from the datasets we chose to forecast future cocoa prices with (future prices, climate, and currency exchange data). In reality, there are a variety of other factors that could also affect cocoa production and therefore, cocoa futures prices. Other factors include disease/pest data, political instability, and disruptions in the cocoa production process, though this data is hard to find as data is typically not collected because it is hard to quantify. Nonetheless, as focusing on too many variables could lead to overfitting in the model, we decided to just focus on futures prices, temperature, precipitation, and currency exchange rate as the main variables to fit the best model possible based on data available online.

When using the cocoa future prices data to create our own model, we also recognized that the future prices are not only based on historical data, but are also heavily influenced by changes in news sentiment and the economy which are hard to predict. Additionally, there was missing data in the precipitation dataset that could affect the future model when accumulated. Due to inconsistencies in how data was collected, we were forced to take the monthly averages of the daily reported temperature, precipitation, and currency exchange rate data. Although taking the averages of these values made it easier for us to create our models, it potentially eliminated specific details that could also be used to create our models and therefore forecast cocoa prices.

Future potential applications for this research could be to forecast cocoa prices with different variables, such as looking at how global demand for chocolate affects its pricing. Another application could be applying a similar SARIMA model to the one we used to predict future cocoa prices in other countries with high levels of cocoa production, such as Cote d'Ivoire or Indonesia.

Conclusion

In this study, we researched past literature and determined four different methods to forecast future cocoa prices in Ghana: a univariate ARIMA model, a SARIMAX model, multiple linear regression, and the random forests model. After checking for stationary assumptions and finding four different models from this data, we determined that the SARIMA (0,1,1) x (0,1,1)12 model was the best fit for predicting future cocoa prices, as it utilized multiple variables that affect the pricing of cocoa, including temperature, precipitation, and currency exchange rates. Furthermore, the SARIMA model had the lowest RMSE value compared to the random forest model and multiple linear regression model. As a result, we chose this model to forecast the cocoa futures prices for four months in the future, forecasting prices of 6242.38, 6211.10, 6285.50, and 6362.22 for the months of October 2024 to January 2025. Given that cocoa production is a significant contributor to Ghana's economy, this model provides valuable information to cocoa farmers and the Ghanaian government for predicting how much they can expect to make through cocoa production and therefore maximizing revenue through cocoa contracts signed in the future.

References

Geman, H. & Sarfo, S. (2012). Seasonality in cocoa spot and forward markets: Empirical evidence. *Journal of Agricultural Extension and Rural Development*, 4(8), 164-180. https://doi.org/10.5897/JAERD11.123

Oni O.V., Oni O.A., Akanle Y.O., Ogunleye T.B.. (2021). Modelling Annual Cocoa Production Using ARIMA Time Series Model. *African Journal of Mathematics and Statistics Studies (AJMSS)*, 4(3), 135-144. https://doi.org/10.52589/AJMSS-0LPATDNK

Ouyang, H., Wei, Xiaolu, & and Wu, Q. (2019). Agricultural commodity futures prices prediction via long- and short-term time series network. *Journal of Applied Economics*, 22(1), 468–483. https://doi.org/10.1080/15140326.2019.1668664

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications with R Examples* (4th ed.). Springer International Publishing.

Uduh, D. M. (2017). Impact of Exchange Rate on Cocoa Export in Nigeria. *International Journal of Economics, Commerce and Management*, 5(5). https://ijecm.co.uk/wp-content/uploads/2017/05/5531a.pdf

Wessel, M., & and Quist-Wessel, P. M. F. (2015). Cocoa production in West Africa, a review and analysis of recent developments. *NJAS: Wageningen Journal of Life Sciences*, 74–75(1), 1–7. https://doi.org/10.1016/j.njas.2015.09.001

Ye, W., Guo, R., Deschamps, B., Jiang, Y., & Liu, X. (2021). Macroeconomic forecasts and commodity futures volatility. *Economic Modelling*, 94, 981–994. https://doi.org/10.1016/j.econmod.2020.02.038

Zhang, D., Dai, X., & Xue, J. (2024). Incorporating weather information into commodity portfolio optimization. *Finance Research Letters*, 66, 105672. https://doi.org/10.1016/j.frl.2024.105672

Zhao, Y. (2023). Comparison of Stock Price Prediction in Context of ARIMA and Random Forest Models. *BCP Business & Management*, 38, 1880-1885. https://doi.org/10.54691/bcpbm.v38i.3996

Appendix

```
# Ghana Exchange Rate Data
exch_rate_avg <- read.table("Ghana_Exchange_Rates.csv", sep=",", header=TRUE)
exch_rate_avg$YearMonth <- paste0(gsub("M", "-",exch_rate_avg$YearMonth), "-01")
exch_rate_avg$YearMonth <- as.Date(exch_rate_avg$YearMonth,"%Y-%m-%d")
# Drop all other columns
exch_rate_avg$SDR_End_Period <- NULL
exch_rate_avg$SDR_Period_Average <- NULL
exch_rate_avg$Dom_USD_End_Period <- NULL
tail(exch_rate_avg,10)
# Check for duplicates
duplicates <- exch_rate_avg[duplicated(exch_rate_avg[c("YearMonth")]), ] # Checking for duplicates
# Plotting the values
plot(x=exch_rate_avg$YearMonth, y=exch_rate_avg$Dom_USD_Period_Average, main = "Average Monthly Cedi Ex
# Ghana Climate Data
climate <- read.table("Ghana_Climate_Data.csv", sep=",", header=TRUE)
climate$DATE <- as.Date(climate$DATE,"%Y-%m-%d")
climate$PRCP[is.na(climate$PRCP)] <- 0
climate$TMAX <- NULL
climate$TMIN <- NULL
duplicates <- climate[duplicated(climate[c("DATE")]), ] # Checking for duplicates
# Climate Data without location bias
climate_cmloc <- climate %>%
  group_by(DATE) %>%
  summarize(AVG_PRCP = mean(PRCP),AVG_TAVG = mean(TAVG))
sum(is.na(climate_cmloc$AVG_TAVG))
sum(is.na(climate_cmloc$AVG_PRCP))
# Assign all NaN's to 0 as they mean
climate_cmloc$YearMonth <- as.Date(format(climate_cmloc$DATE, "%Y-%m-01"))
climate_avg <- climate_cmloc %>%
  group_by(YearMonth) %>%
  summarize(
    PrcpAvg = mean(AVG_PRCP), # Calculating how many days in the month it rained
    TAvg = mean(AVG_TAVG) # Calculated the months average temperature)
sum(is.na(climate_avg$PrcpAvg))
sum(is.na(climate_avg$TAvg))
# Plotting the series
ggplot(climate_avg, aes(x = YearMonth, y = TAvg)) +
  geom_line(color = 'black') +  # Line graph
  geom_point(color = 'black') +  # Points on the line
  labs(title = "Temperature in Ghana", x = "Time (monthly)", y = "Temperature (F)")
ggplot(climate_avg, aes(x = YearMonth, y = PrcpAvg)) +
  geom_bar(stat = "identity", fill = "black")+
  labs(title = "Precipitation in Ghana", x = "Time (monthly)", y = "Average Precipitation")
# Cocoa Daily Prices
prices <- read.table("Cocoa_Daily_Prices.csv", sep=",", header=TRUE) # US dollar/tonne
colnames(prices)[2] <- "Price" # renaming the column to make it easier
prices$Date <- as.Date(prices$Date,"%d/%m/%Y")
prices$YearMonth <- as.Date(format(prices$Date, "%Y-%m-01"))
prices$Price <- as.numeric(gsub(",", "",prices$Price)) # removing commas from the values
# Checking for the presence of NA values
sum(is.na(prices))
# Let's check if there are duplicates
```

```r
dups <- prices[duplicated(prices$Date),1]
print(dups)
for (d in dups){print(prices[prices$Date == as.Date(d),])}
#Let's remove the duplicates
prices <- prices[-c(279,281,297,312),]
# Check if we have duplicates
print(prices[duplicated(prices$Date),1])
# Making this monthly data
prices_avg <- prices %>%
  group_by(YearMonth) %>%
  summarize(PriceAvg = mean(Price) # Calculated the months average temperature)
# Plotting the values
plot(x=prices_avg$YearMonth, y=prices_avg$PriceAvg, main = "Monthly Cocoa Futures Price",
     xlab = "Time (monthly)", ylab = "Price (USD/tonne)", cex=0.4)
# Let's make sure all years have 12 months with the exception of 2024
checked <- climate_avg %>%
  mutate(Year = format(YearMonth, "%Y")) %>%   # Extract year as a new column
  group_by(Year) %>% summarise(Months = n())

print(checked$Year[checked$Months <12])
checkedd <- prices_avg %>%
  mutate(Year = format(YearMonth, "%Y")) %>%   # Extract year as a new column
  group_by(Year) %>% summarise(Months = n())
print(checkedd$Year[checkedd$Months <12])
checkeddd <- exch_rate_avg %>%
  mutate(Year = format(YearMonth, "%Y")) %>%   # Extract year as a new column
  group_by(Year) %>% summarise(Months = n())
print(checkeddd$Year[checkeddd$Months <12])
print(climate_avg[format(climate_avg$YearMonth, "%Y") == "2024",])
print(climate_avg[format(climate_avg$YearMonth, "%Y") == "2001",])
print(climate_avg[format(climate_avg$YearMonth, "%Y") == "1994",])
# Create new rows to add these months (they don't exist in the dataset at all so we need to add them)
add_rows <- data.frame(
  YearMonth = as.Date(c('1994-12-01', '2001-10-01', '2001-12-01')),  # New rows (missing months)
  PrcpAvg = c(NA, NA, NA),  # Precipitation average (NA)
  TAvg = c(NA, NA, NA))     # Temperature average (NA)
climate_avg <- rbind(climate_avg, add_rows)
climate_avg <- climate_avg[order(climate_avg$YearMonth), ]
# Interpolating the three missing values
climate_avg$PrcpAvg <- approx(climate_avg$YearMonth, climate_avg$PrcpAvg, xout = climate_avg$YearMonth)$
climate_avg$TAvg <- approx(climate_avg$YearMonth, climate_avg$TAvg, xout = climate_avg$YearMonth)$y
final_data <- prices_avg %>%
  inner_join(climate_avg, by = "YearMonth") %>%
  inner_join(exch_rate_avg, by = "YearMonth")
# Drop any crows with Date < 1994-10-01
final_data <- final_data %>% filter(YearMonth >= as.Date("1994-10-01"))
# Let's make sure all years have 12 months with the exception of 2024
checker <- final_data %>%
  mutate(Year = format(YearMonth, "%Y")) %>%   # Extract year as a new column
  group_by(Year) %>% summarise(Months = n())
print(final_data)
write.csv(final_data, file = "cleaned_data.csv", row.names = FALSE)
print(checker)
```

MLR

```r
data_x <- read.table("cleaned_data.csv", sep=",", header=TRUE)
# Get the separate time series
temp <- data_x$TAvg
precip <- data_x$PrcpAvg
price <- data_x$PriceAvg
idx <- data_x$Dom_USD_Period_Average
#plot the ACFs
acf(temp, main="ACF of Temperature Series")
acf(precip, main="ACF of Precipitation Series")
acf(price,main="ACF of Futures Price Series")
acf(idx,  main="ACF of Cedi Exchange Rate Series")
```

```r
l_temp <-  diff(temp, lag = 12) # There seems to be some annual seasonality
l_precip <- diff(precip, lag = 12) # There seems to be some annual seasonality
l_price <- diff(log(price))
l_idx <- diff(log(idx))
#plot the ACFs
acf(l_temp, main="ACF of Diff Temperature Series")
acf(l_precip, main="ACF of Diff Precipitation Series")
acf(l_price, main="ACF of Diff Futures Price Series")
acf(l_idx,  main="ACF of Diff Cedi Exchange Rate Series")
ccf(l_price, l_temp, lag.max = 25, main = "Cross-Correlation between Price and Temperature")
ccf(l_price, l_precip, lag.max = 25, main = "Cross-Correlation between Price and Precipitation")
ccf(l_price, l_idx, lag.max = 25, main = "Cross-Correlation between Price and Currency Rates")
# Editing the length of temp and precip
l_temp <-  c(rep(NA, 11),l_temp) # There seems to be some annual seasonality
l_precip <- c(rep(NA, 11),l_precip) # There seems to be
# Lagging the data
lag_cr_17 <- c(rep(NA, 17), l_idx[1:(length(l_idx) - 17)])
lag_pr_1 <- c(rep(NA, 1), l_price[1:(length(l_price) - 1)])
lag_pr_7 <- c(rep(NA, 7), l_price[1:(length(l_price) - 7)])
lag_pcp_22 <- c(rep(NA, 22), l_precip[1:(length(l_precip) - 22)])
lag_tmp_24 <- c(rep(NA, 24), l_temp[1:(length(l_temp) - 24)])
# Let's drop the first 35 rows as lag 24 is our max lag and we already diff
#the temp/prcp series with lag 12. So, we will have data only from 1997 Aug.
edited_data <- cbind(l_price, l_temp, l_precip, l_idx, lag_pr_1, lag_pr_7,
                     lag_cr_17, lag_tmp_24, lag_pcp_22)
test_set <- data.frame(tail(edited_data,4))
train_set <- head(edited_data, n = nrow(edited_data) - 4)
# Now let's adjust the train_set for lags
train_set <- data.frame(edited_data[36:nrow(train_set), ])
head(train_set)
model <- lm(l_price ~ lag_pr_1 + lag_pr_7 + lag_cr_17 + lag_pcp_22 + lag_tmp_24,
            data=train_set)
summary(model)
AIC(model)
checkresiduals(model)
plot(model$residuals)
# Q-Q plot of residuals
qqnorm(residuals(model))
qqline(residuals(model))
Box.test(residuals(model),lag=10, type = "Ljung-Box")
```

```r
forecasted <- predict(model, newdata = test_set[, -which(names(test_set) == "l_price")])
print(forecasted)

for_prices  <- c(log(data_x$PriceAvg[length(data_x$PriceAvg)-3]))
# Get the last 5th log price from the training set

for (i in 1:length(forecasted)) {
  for_prices <- c(for_prices, for_prices[length(for_prices)] + forecasted[i])
}
for_prices <- for_prices[-1] # Removing the first price
for_prices <- exp(for_prices)

rmse <- sqrt(mean((for_prices - data_x$PriceAvg[(length(data_x$PriceAvg)-3):length(data_x$PriceAvg)])^2
print(paste("RMSE: ", rmse))
print("Forecasted Prices")
print(for_prices)
print("Actual Prices")
print(data_x$PriceAvg[(length(data_x$PriceAvg)-3):length(data_x$PriceAvg)])
# Plot Original vs Forecasted Values
for_prices <- c(rep(NA, length(data_x$PriceAvg) - length(for_prices)), for_prices)
ggplot() +
  geom_line(aes(x=as.Date(data_x$YearMonth),y=data_x$PriceAvg), color = "black") +# Line plot
  geom_line(aes(x=as.Date(data_x$YearMonth),y=for_prices), color = "red") +
  labs( x = "Date", y = "Cocoa Future Prices")
```

ARIMA

```r
data <- read_csv("cleaned_data.csv")

price_ts <- ts(data$PriceAvg, start=c(1994, 10), frequency=12)
temp_ts <- ts(data$PrcpAvg, start=c(1994, 10), frequency=12)
rain_ts <- ts(data$TAvg, start=c(1994, 10), frequency=12)
currency_ts <- ts(data$Dom_USD_Period_Average, start=c(1994, 10), frequency=12)
```

```r
acf2(price_ts)
acf2(temp_ts)
acf2(rain_ts)
acf2(currency_ts)

adf.test(price_ts)
adf.test(temp_ts)
adf.test(rain_ts)
adf.test(currency_ts)
```

```r
# ARIMA model
arima_model1 <- arima(log_price, order = c(1,1,0))
arima_model2 <- arima(log_price, order = c(0,1,1))
arima_model3 <- arima(log_price, order = c(1,1,1))
AIC(arima_model1, arima_model2, arima_model3)
```

SARIMA + FORECAST

```r
cleands <- read.csv("cleaned_data.csv")

ts_price <- ts(cleands$PriceAvg, frequency = 12, start = c(1994, 10))

#Taking log of the prices to see if we can help with heteroskeadicity and fit,
#more relative change, but will have to unlog after forecasting

log_ts_price <- ts(log(cleands$PriceAvg), frequency = 12, start = c(1994, 10))
plot(ts_price)
plot(log_ts_price)
```

```r
par(mfrow = c(1,2))
acf(ts_price, main = "ACF")
pacf(ts_price, main = "PACF")
```

```r
par(mfrow = c(1,2))
acf(log_ts_price, main = "ACF")
pacf(log_ts_price, main = "PACF")
```

```r
#try diff
dprice <- diff(log_ts_price)

plot(dprice, type = "l",
     main = "First Difference of Data",
     ylab = expression(Delta * "Price"),
     xlab = "Time")

acf(dprice, main = "ACF")
pacf(dprice, main = "PACF")

#exogenous vars
exog_vars <- cleands %>%
  select(PrcpAvg, TAvg, Dom_USD_Period_Average) %>%
  as.matrix()

#sarmia modeling

ma_1_models <- sarima(ts_price, 0,0,1)

arimax_model <- auto.arima(ts_price, xreg = exog_vars)
summary(arimax_model)
checkresiduals(arimax_model)
```

```r
#logARIMAX MODEL
arimax_model_log <- auto.arima(log_ts_price, xreg = exog_vars)
summary(arimax_model_log)
checkresiduals(arimax_model_log)

#sarmia modeling with vars includede

sarima_model <- sarima(log_ts_price, 0,1,1, xreg = exog_vars)
sarima_model
sarima_model_111 <- sarima(ts_price, 1,1,1, xreg = exog_vars)
```

```r
AIC(arimax_model, arimax_model_log)
#logedpriced ARIMAX model (0,1,1) beats out ARIMAX model (0,1,0) via AIC
sarima_model$ICs
```

```r
#nonseasonalARIMAseems to underfit try seasonal
sarima_model_011_withseason <- sarima(log_ts_price, 0,1,1,0,1,1,12, xreg = exog_vars)

sarima_model_011_univariate <- sarima(log_ts_price, 0,1,1,0,1,1,12)
```

```r
#randomforest attempt

data_ml <- cleands %>%
  mutate(PriceLag1 = lag(PriceAvg, 1)) %>%
  drop_na()
```

```r
# Split train/test (80/20)
set.seed(123)
train_idx <- createDataPartition(data_ml$PriceAvg, p = 0.8, list = FALSE)
train <- data_ml[train_idx, ]
test <- data_ml[-train_idx, ]
```

```r
# Fit random forest
rf_model <- randomForest(
  PriceAvg ~ PriceLag1 + PrcpAvg + TAvg + Dom_USD_Period_Average,
  data = train
)
```

```r
rf_preds <- predict(rf_model, newdata = test)

postResample(rf_preds, test$PriceAvg)
#Will need to compair RMSE of the random forest preds, with the forecasted
#data from the ARIMA/SARIMA/ARIMAX model of our choosing.
```

```r
#RF forecasting

# 1. Last known value (log price) to kick off the loop
last_log_price <- tail(log_ts_price, 1)

# 2. Create dummy exogenous values (e.g., use last known or rolling mean)
last_row <- cleands %>%
  summarise(
    PrcpAvg = mean(PrcpAvg, na.rm = TRUE),
    TAvg = mean(TAvg, na.rm = TRUE),
    Dom_USD_Period_Average = mean(Dom_USD_Period_Average, na.rm = TRUE)
  )

# 3. Build 4-month future input set
future_rf_input <- last_row[rep(1, 4), ]
future_rf_input$PriceLag1 <- NA

# 4. Iteratively forecast next 4 values
rf_forecast_log <- c()

for (i in 1:4) {
  # Step 1: Set lag
  future_rf_input$PriceLag1[i] <- if (i == 1) last_log_price else rf_forecast_log[i - 1]

  # Step 2: Predict
  rf_forecast_log[i] <- predict(rf_model, newdata = future_rf_input[i, ])
}

# Print results
rf_forecast_log

rf_forecast_act_log <- log(rf_forecast_log)
```

```
newxreg <- tail(exog_vars, 4)  # now 4 rows for 4 months ahead

forecast_log <- sarima.for(log_ts_price, n.ahead = 4,
           p = 0, d = 1, q = 1,
           P = 0, D = 1, Q = 1,
           S = 12,
           newxreg = newxreg, main = "SARIMA Forecast of Log Cocoa Prices (4-Month Horizon)")

print(forecast_log)

forcased <- exp(forecast_log$pred)
print(forcased)
```

```
#splitting ts into test (4 months) and train
train_size <- length(log_ts_price) - 4
tstrain <- head(log_ts_price, train_size)
tstest <- tail(log_ts_price,4)


#exogenous vars
#exog_vars <- cleands %>%
#  select(PrcpAvg, TAvg, Dom_USD_Period_Average) %>%
 # as.matrix()

xreg_train <- as.matrix(head(exog_vars, train_size)[, c("PrcpAvg", "TAvg", "Dom_USD_Period_Average")])
xreg_test <- as.matrix(tail(exog_vars,4)[, c("PrcpAvg", "TAvg", "Dom_USD_Period_Average")])

train_sarima_model <- sarima(tstrain, 0,1,1,0,1,1,12, xreg = xreg_train)

test_sarima_forecast <- sarima.for(tstrain, n.ahead = 4,
           p = 0, d = 1, q = 1,
           P = 0, D = 1, Q = 1,
           S = 12,
           newxreg = xreg_test)

test_forecast_price <- exp(test_sarima_forecast$pred)
print(test_forecast_price)

test_actual <- exp(tstest)

test_rmse <- RMSE(test_actual,test_forecast_price)
```

```
test_rmse

rf_rmse <- RMSE(test_actual,rf_forecast_log)

rf_rmse
```

```
#plotvalues

n_test <- 4
log_actual <- tail(log_ts_price, n_test)
log_forecast <- test_sarima_forecast$pred  # from sarima.for()

# Create date sequence for x-axis (adjust if needed)
forecast_dates <- seq(as.Date("2024-01-01"), by = "month", length.out = n_test)

# Build dataframe for plotting
plot_df <- data.frame(
  Date = forecast_dates,
  Actual = as.numeric(exp(log_actual)),
  Forecast = as.numeric(exp(log_forecast)),
  RandomForest = as.numeric(rf_forecast_log)
)

# Convert to long format
plot_long <- plot_df %>%
  pivot_longer(cols = c("Actual", "Forecast", "RandomForest"), names_to = "Type",
               values_to = "LogPrice")

# Plot
ggplot(plot_long, aes(x = Date, y = LogPrice, color = Type)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  labs(title = "SARIMA Forecast vs Actual vs RandomForest",
       x = "Date", y = "Cocoa Price") +
  scale_color_manual(values = c("Actual" = "black", "Forecast" = "red",
                                "RandomForest" = "forestgreen")) +
  theme_minimal(base_size = 14)
```

```
# Build dataframe for plotting
plot_df <- data.frame(
  Date = forecast_dates,
  Actual = as.numeric(exp(log_actual)),
  Forecast = as.numeric(exp(log_forecast)),
  MutltipleReg = as.numeric(c(7945.434, 8507.443, 8178.286, 8310.560))
)

# Convert to long format
plot_long <- plot_df %>%
  pivot_longer(cols = c("Actual", "Forecast", "MutltipleReg"), names_to = "Type",
               values_to = "LogPrice")

# Plot
ggplot(plot_long, aes(x = Date, y = LogPrice, color = Type)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  labs(title = "SARIMA Forecast vs Actual vs MutltipleReg",
       x = "Date", y = "Cocoa Price") +
  scale_color_manual(values = c("Actual" = "black", "Forecast" = "red",
                                "MutltipleReg" = "forestgreen")) +
  theme_minimal(base_size = 14)
```

```
#trying lagged vars
data_x <- read.table("cleaned_data.csv", sep=",", header=TRUE)
temp <- data_x$TAvg
precip <- data_x$PrcpAvg
price <- data_x$PriceAvg
idx <- data_x$Dom_USD_Period_Average

l_temp <-  diff(temp, lag = 12) # There seems to be some annual seasonality
l_precip <- diff(precip, lag = 12) # There seems to be some annual seasonality
l_price <- diff(log(price))
l_idx <- diff(log(idx))
# Editing the length of temp and precip
l_temp <-  c(rep(NA, 11),l_temp) # There seems to be some annual seasonality
l_precip <- c(rep(NA, 11),l_precip) # There seems to be
# Lagging the data
lag_cr_17 <- c(rep(NA, 17), l_idx[1:(length(l_idx) - 17)])
lag_pr_1 <- c(rep(NA, 1), l_price[1:(length(l_price) - 1)])
lag_pr_7 <- c(rep(NA, 7), l_price[1:(length(l_price) - 7)])
lag_pcp_22 <- c(rep(NA, 22), l_precip[1:(length(l_precip) - 22)])
lag_tmp_24 <- c(rep(NA, 24), l_temp[1:(length(l_temp) - 24)])
# Let's drop the first 35 rows as lag 24 is our max lag and we already diff the
#temp/prcp series with lag 12. So, we will have data only from 1997 Aug.
edited_data <- cbind(l_price, l_temp, l_precip, l_idx, lag_pr_1, lag_pr_7, lag_cr_17,
                     lag_tmp_24, lag_pcp_22)
edited_data <- as.data.frame(edited_data)
target <- edited_data$l_price
exog <- as.matrix(edited_data[, c("lag_pr_1", "lag_pr_7", "lag_cr_17", "lag_tmp_24",
                                  "lag_pcp_22")])
sarima_fit <- sarima(target, p = 0, d = 1, q = 1,
                     P = 0, D = 1, Q = 1, S = 12,
                     xreg = exog)
```