Cocoa Price Prediction Model for Ghana*

Forecasting Cocoa Price Flutuation Using Time Series

Shanjie Jiao Edward Hong Lilian Sun Haoya Wang

April 2, 2025

Table of contents

1 Model

This study aims to develop a predictive model to capture future fluctuations in cocoa prices. To enhance the model's forecasting accuracy, a range of exogenous variables are considered, spanning both agricultural asepcts and macroeconomic dimensions. Climatic factors such as precipitation and temperature are considered due to their indirect influence on market prices through their effects on cocoa yield, which is considered as the most important factor pushing cocoa price. In addition, agricultural indicators—including labor input, cultivated area, yield

*Code and data are available at: https://github.com/Jie-jiao05/Cocoa_price_preditcion.

per hectare—as well as productivity-related metrics such as total factor productivity (TFP), are integrated into the framework to comprehensively evaluate their potential impact on price. By incorporating these variables, we hope to explore how these environmental and economic variables explain their impact on cocoa prices.

To investigate the potential impact of external variables on cocoa prices, the Generalized Additive Model (GAM), Autoregressive Integrated Moving Average (ARIMA), and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models are considered as candidate approaches.

1.1 Model Set-up

1.1.1 Generalized Additive Model (GAM)

Price, as the response variable in this study, is continuous, strictly positive, and reflects actual measured values rather than frequencies or binary outcomes for decision-making purposes. Thus, the Gamma distribution is selected. The use of a log link function ensures that predicted prices remain positive and allows the model to capture nonlinear and multiplicative relationships between the response and explanatory variables. This makes the Gamma distribution a theoretically appropriate and practically robust choice for modeling the influence of external factors on cocoa prices.

Since the dataset is organized by month (from January 2015 to December 2023) and includes only the Ghana region, there is no hierarchical or nested structure in the data. Furthermore, the temporal dimension is explicitly available through the monthly time variable. Therefore, random effects are not included in the model; instead, we focus on fixed effects, along with a smooth function of time. The smooth term is incorporated to capture nonlinear trends in the response over time. Additionally, since the outcome variable is cocoa price, a continuous quantity rather than a rate or count so offset term will not be considered in the model.

The model is defined as follows:

$$\begin{aligned}
Y_t \mid U &\sim \text{Gamma}(\mu_t, \theta), \quad g(\mu_t) = X_t\beta + U(t) \\
g(\mu_t) = \log(\mu_t) &= \beta_0 + s_1(\text{Month\_Index}_t) + s_2(\text{Temp}_t) + s_3(\text{Fert}_t) + s_4(\text{TFP\_Index}_t) \\
&\quad + s_5(\text{Capital\_Index}_t) + s_6(\text{Land\_Q}_t) + s_7(\text{Labor\_Q}_t) + s_8(\text{Cropland\_Q}_t) \\
&\quad + s_9(\text{prep}_t) + \beta_{10} \cdot \text{Production\_tonnes}_t + \beta_{11} \cdot \text{Yield\_tonnes\_per\_hectare}_t \\
&\quad + U(t) \\
U(t) &\sim \text{IWP}_2(\sigma) \quad \text{(Smooth Trend)}
\end{aligned}$$

1.1.2 Autoregressive Integrated Moving Average (ARIMA)

The second model we select is an ARIMAX model. Our dataset provides accurate monthly records from January 2015 to December 2023, along with a range of potentially influential external variables such as temperature, fertilizer use, and productivity indicators. For the standard ARIMA model, which accounts only for a univariate time series, the ARIMAX framework enhances forecasting accuracy by incorporating both temporal dependencies and external factors. By addressing non-stationarity via differencing, it help to stabilizes the data and facilitates more reliable model construction.

From the initial plot of the cocoa price data, there is no clear evidence of a seasonal trend, as the series appears to fluctuate irregularly over time. However, the ACF and PACF plots of the original (undifferenced) series reveal signs of non-stationarity, as the ACF decay slowly. To address this, we apply first-order differencing, which yields a series that appears stationary. The ACF and PACF plots of the differenced series indicate an moving average structure of order 2. Based on these diagnostics, we propose an ARIMA(0,1,2) with external regressors model for the cocoa price series.

The model is defined as follows:

$$\begin{aligned}\Delta y_t = {} & \theta_1 w_{t-1} + \theta_2 w_{t-2} \\ & + \beta_1 \cdot \mathrm{Temp}_t + \beta_2 \cdot \mathrm{Fert}_t + \beta_3 \cdot \mathrm{TFP\_Index}_t \\ & + \beta_4 \cdot \mathrm{Capital\_Index}_t + \beta_5 \cdot \mathrm{Land\_Q}_t + \beta_6 \cdot \mathrm{Labor\_Q}_t \\ & + \beta_7 \cdot \mathrm{Cropland\_Q}_t + \beta_8 \cdot \mathrm{prep}_t + \beta_9 \cdot \mathrm{Production\_tonnes}_t \\ & + \beta_{10} \cdot \mathrm{Yield\_tonnes\_per\_hectare}_t + w_t, \quad w_t \sim \mathcal{N}(0, \sigma^2)\end{aligned}$$

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

1.1.3 Machine Learning — XGBoost

Under this research, we aim focus on accurately forecasting cocoa prices under conditions of complex and volatile market fluctuations. While traditional time series models are effective at capturing linear relationships and structured temporal dependencies, they often fall short when modeling the nonlinear dynamics commonly found in agricultural commodity markets. To address these limitations, we adopt a machine learning approach by employing an XGBoost model with lagged price features to forecast future cocoa prices. By incorporating twelve lagged values of the price series, the model effectively utilizes historical information to learn intricate patterns, address the nonlinearities previously identified, and contribute to a more robustic model.

The XGBoost framework is defined as follows:

$$\hat{y}_t = \sum_{k=1}^{K} f_k(\mathbf{x}_t), \quad f_k \in \mathcal{F}$$

- $\hat{y}_t$: predicted cocoa price at time t
- $\mathbf{x}_t$: feature vector consisting of lagged cocoa prices
- f_k: the k-th regression tree
- $\mathcal{F}$: the functional space of all possible regression trees
- K: total number of boosting iterations (trees)

1.2 Final Model

The performance metrics for the three models—ARIMAX, GAM, and the machine learning-based XGBoost—are summarized in Table 1. Among these, the XGBoost model demonstrates superior predictive performance, achieving the lowest RMSE (148.94) and the highest R^2 (0.92), illustrating that approximately 92% of the variation in cocoa prices in the test set can be explained by this model. In addition, the XGBoost model achieves a considerably low MAE (107.15) and MAPE (4.08%), further validating its accuracy and reliability. The consistently strong performance across multiple evaluation metrics highlights XGBoost's capability to deliver precise and robust forecasts. In contrast, the traditional statistical models exhibit higher error rates and negative R^2 values, suggesting poor generalization to out-of-sample data. Given these findings, the XGBoost model is selected as the final forecasting model for this study

Table 1: Model Result

Model	AIC	RMSE	MAE	MAPE	R_squared
XGBoost	NA	148.940	107.15	4.08%	0.920
ARIMAX	1357.807	1588.224	NA	NA	-11.864
GAM	953.568	682.315	NA	NA	-1.374

2 Results

2.1 Model Performance Validation

Figure 1 The predicted cocoa prices generated by our XGBoost model align closely with the actual observed values across the entire time period. This result highly align with the model's statistical performance, as reflected by a high R^2 value of 0.92. Where our model demonstrates

strong capability in capturing both short-term fluctuations and long-term trends in cocoa price.

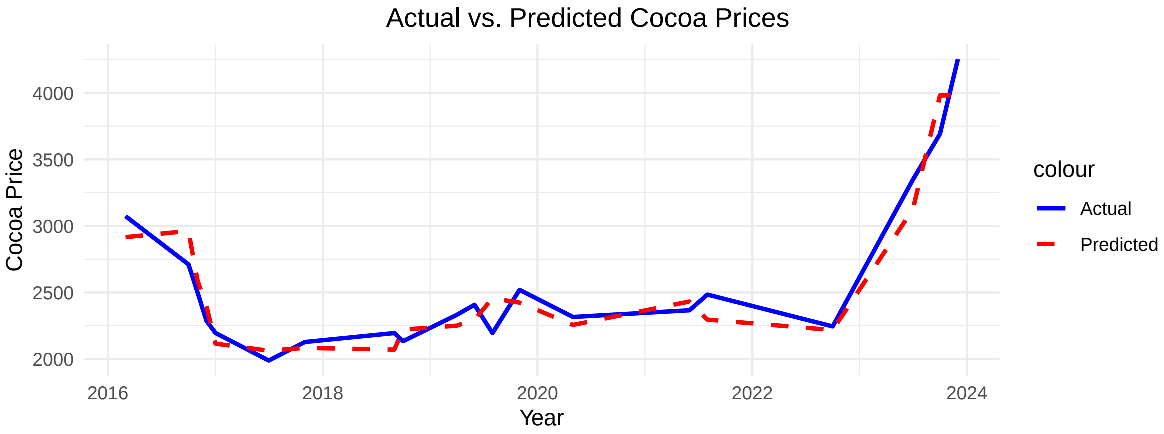

Figure 1: Actual vs. Predicted Cocoa Prices

2.2 Model Lag Check

It is evident from Figure 2 that lag_1 by far has the highest importance, our model heavily relieson the cocoa price from one month ago. This finding aligns with real-world financial behavior, where more recent observations typically have a stronger influence on current prices, while distant past data tends to carry less predictive power. Conversely, some lag features such as lag_11 and lag_2 show near-zero contribution, suggesting that the model did not find time at that lag is significant for prediction.

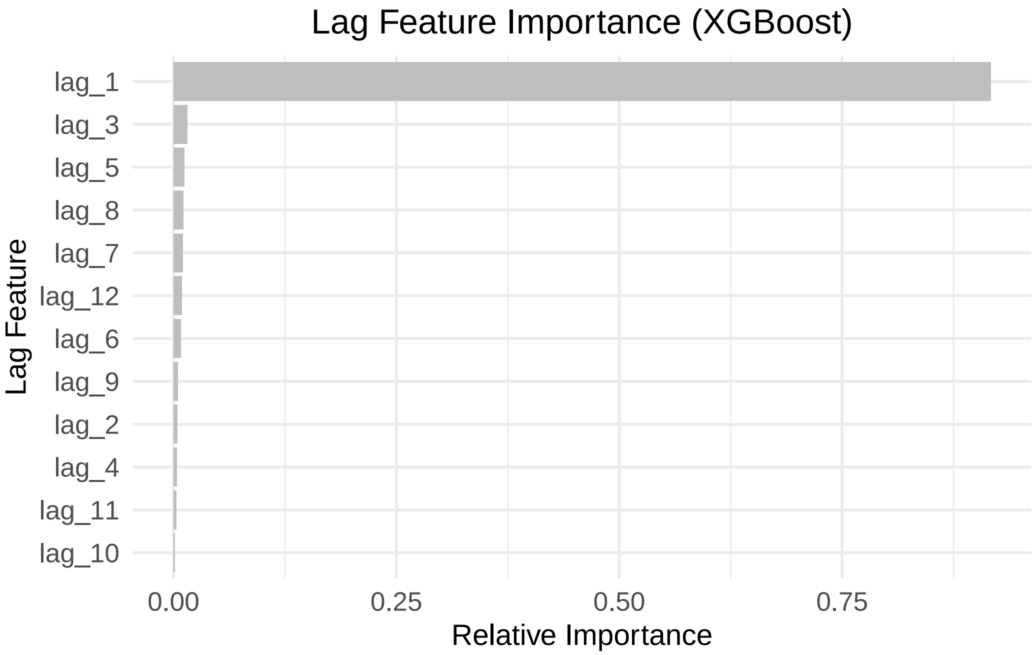

Figure 2: Model lag information

2.3 Residual Check

To evaluate the adequacy of the XGBoost model, we conducted a residual diagnostic analysis using a histogram, residuals vs. predicted plot, and the autocorrelation function (ACF) of the residuals.

Figure 3 shows that the residuals are approximately symmetrically distributed around zero, indicating that the model does not systematically over or under predict. The Figure 4 residuals vs. predicted plot reveals no discernible pattern or signs of heteroskedasticity, suggesting that the variance of the errors remains consistent across different levels of predicted values. However, it is important to note that due to the limited size of the dataset, only a small portion was allocated to the test set, which may affect the robustness of these diagnostics. Lastly, Figure 5 the ACF plot of the residuals shows no significant autocorrelation beyond lag zero, confirming that the model has effectively captured the temporal structure in the data.

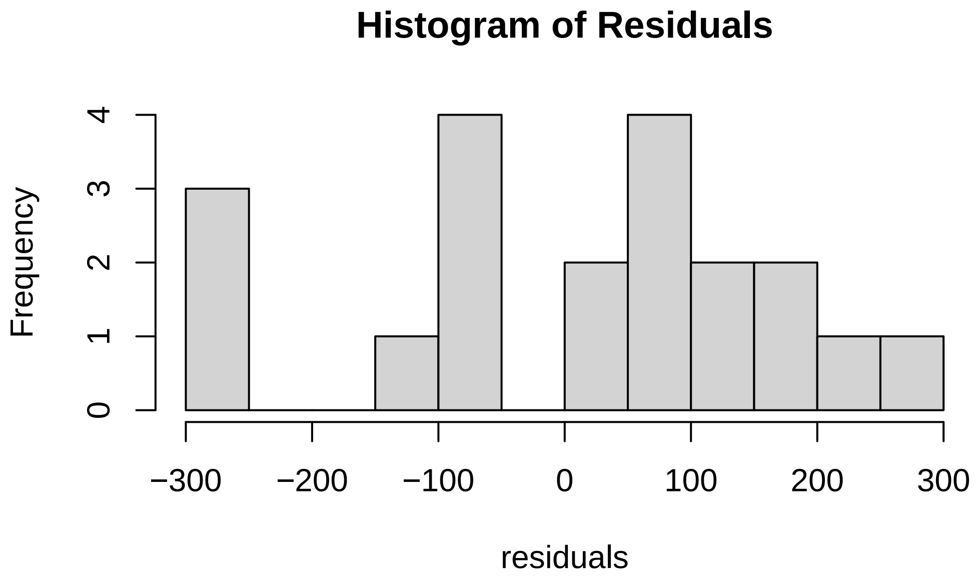

Figure 3: Residual Histogram

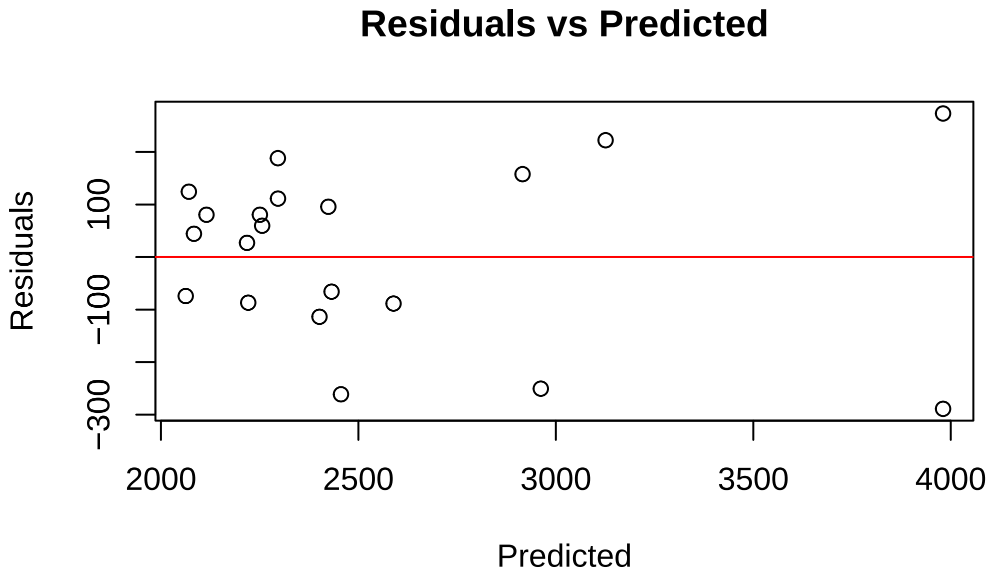

Figure 4: Dot Plot of Residuals vs Predicted Value

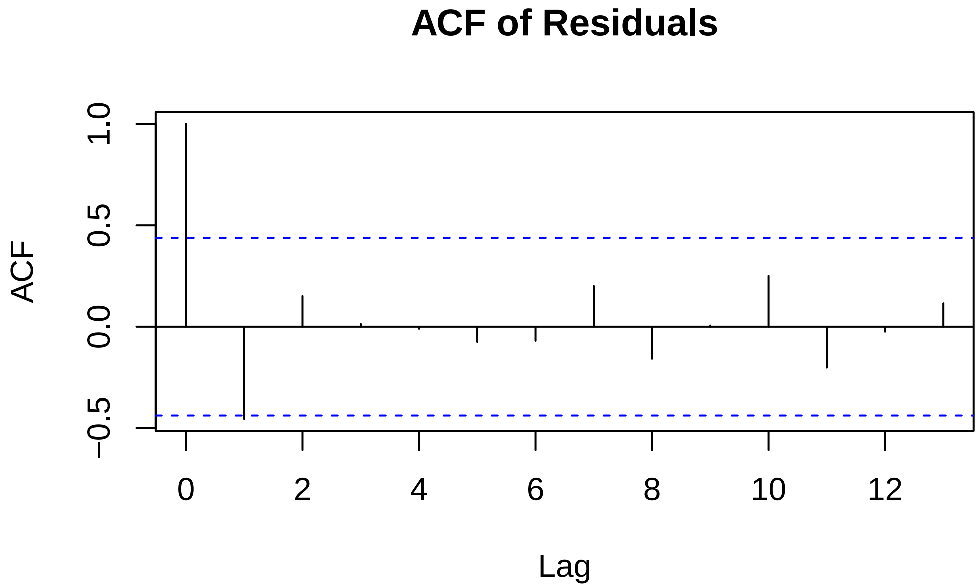

Figure 5: Residual ACF

2.4 Forecasting

Figure 6 illustrates the projected cocoa prices generated by the XGBoost model for the next 2 years (2025-02 to 2026-12), with a 95% confidence interval shaded in red. The forecast demonstrates a downward trend following the recent price peak, while the confidence band captures the expected range of uncertainty in future price movements.

Following a sharp increase in cocoa prices toward the end of 2023, the model forecasts a downward trend, with prices gradually returning to levels comparable to those seen in 2023-01. In early 2025, the model suggests a brief rebound, followed by another period of decline. It is important to note that future price in realistic will inevitably be influenced by seasonal fluctuations in production, changes in labor and transportation costs, and broader macroeconomic conditions affecting global commodity markets. These predictions are based solely on the available dataset and should be interpreted as model-based estimations rather than a definitive forecasts.

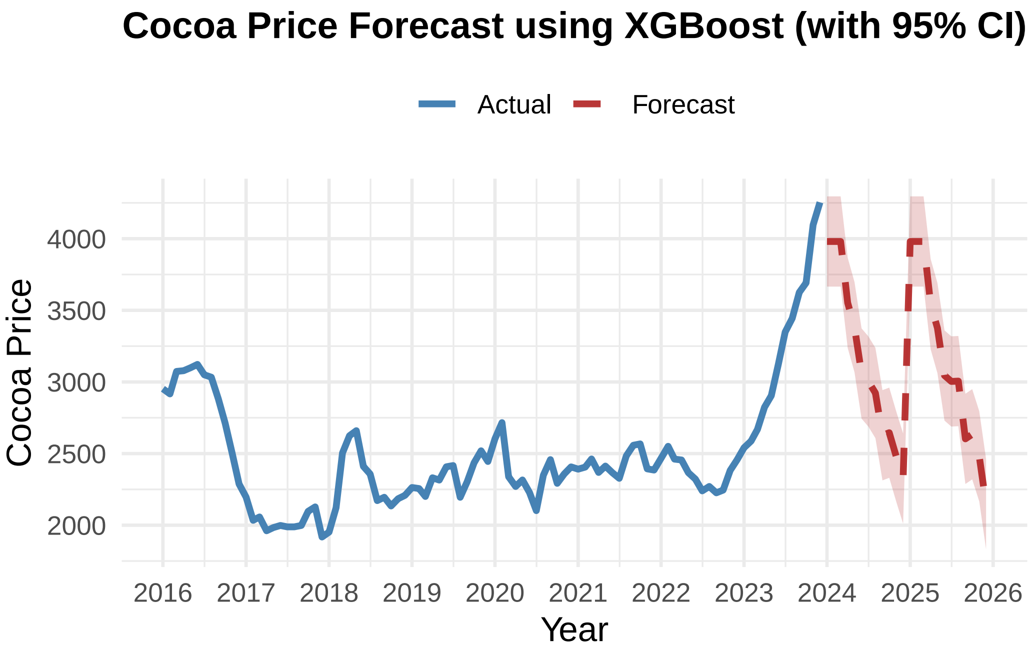

Figure 6: Cocoa Price Prediction Using XGBoost with Confidence Interval

Appendix