

Overcoming Tree Extrapolation Limits in Gold Price Forecasting: A Stationary Return-Reconstructed XGBoost Approach

Imam Saputra^{1*}, Zulfikar², Juni Ismail³, Alfry Aristo Jansen Sinlae⁴, Yanto Saputra⁵, Raja Anan Nasution⁶, Elsy Rahajeng⁷, Mesran¹

¹ Prodi Manajemen, Sekolah Tinggi Ilmu Manajemen Sukma Medan, Sumatera Utara, Indonesia

² Program Studi Teknologi Rekayasa Logistik, Politeknik Kampar, Riau, Indonesia

³ Teknik Komputer, Politeknik Bisnis Indonesia, Sumatera Utara, Indonesia

⁴ Fakultas Teknik, Program Studi Ilmu Komputer, Universitas Katolik Widya Mandira, Nusa Tenggara Timur, Indonesia

⁵ Program Studi Sistem dan Teknologi Informasi, Institut Teknologi Bisnis Riau, Riau, Indonesia

⁶ Program Studi Manajemen Informatika, Akademi Manajemen Informatika dan Komputer ITMI, Sumatera Utara, Indonesia

⁷ Fakultas Sains dan Teknologi, Prodi Sistem Informasi, Universitas Islam Negeri Syarif Hidayatullah, Banten, Indonesia

Email: ^{1*}saputraimam69@gmail.com, ²zulfikar.hc@gmail.com, ³juniismaill@gmail.com, ⁴alfry.aj@unwira.ac.id,

⁵yantosaputra1984@gmail.com, ⁶rajanasutionnnn@gmail.com, ⁷elsy.rahajeng@uinjkt.ac.id, ⁸mesran.skom.mkom@gmail.com

Correspondence Author Email: saputraimam69@gmail.com@gmail.com

Received: 05/09/2026, Revised: 21/09/2026, Accepted: 27/09/2026, Published: 28/09/2026

Abstract—This study addresses the fundamental tree extrapolation limit inherent in gradient-boosted decision tree ensembles when forecasting non-stationary financial time series during historical market rallies. Standard raw-price regression models fail to project unseen price regimes because decision trees assign static scalar predictions to terminal leaf nodes, capping predictions at historical maximums. To overcome this spatial limitation, we propose a Stationary Return-Reconstructed XGBoost framework that shifts the machine learning target from non-stationary spot prices to stationary daily relative returns before deterministically re-integrating predicted returns into nominal price levels. Evaluated on an out-of-sample testing set spanning 2022 to 2024, the return-reconstructed framework completely eliminates flatline prediction plateaus during record-setting gold price rallies exceeding 2,800 USD per troy ounce. The proposed model achieves an R-squared score of 0.9905, a Root Mean Squared Error of 28.17 USD per troy ounce, and a Mean Absolute Percentage Error of 1.12 percent, outperforming the raw-price baseline model by 92.75 percent in average error reduction. Furthermore, game-theoretic SHAP analysis validates that daily return forecasts are driven by macroeconomically sound transmission channels, predominantly led by the US Dollar Index and monthly historical price memory.

Keywords: Gold Price Forecasting; XGBoost; Tree Extrapolation Limit; Stationary Return Reconstruction; Explainable AI (SHAP)

1. INTRODUCTION

Gold occupies a paramount position in global financial markets as a primary safe-haven asset and a strategic hedge against macroeconomic volatility, systemic inflation, and geopolitical turbulence. Consequently, accurately forecasting international gold prices has become an essential endeavor for institutional investors, central banks, and risk management practitioners worldwide [1]. Historically, financial time-series modeling relied heavily on classical econometric and statistical techniques, such as Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) frameworks [2]. While these linear paradigms provide rigorous mathematical guarantees under stationarity assumptions, they frequently fail to capture complex, non-linear dynamics inherent in modern asset price behaviors [3]. The rapid evolution of computational intelligence has catalyzed a paradigm shift toward data-driven machine learning models capable of mapping intricate feature interactions without rigid distributional assumptions. Machine learning techniques have demonstrated superior adaptive capacity in processing high-dimensional financial inputs across varying economic regimes [4]. Among these computational paradigms, ensemble learning algorithms have gained widespread adoption due to their exceptional generalization performance on tabular data [5]. Modern financial decision-making systems increasingly depend on predictive architectures that can assimilate both endogenous price history and exogenous macroeconomic variables. However, financial time series remain notoriously difficult to model owing to non-stationarity, structural breaks, and heavy-tailed distribution characteristics [5]. Establishing a reliable predictive architecture thus requires balancing non-linear representation capacity with mathematically sound temporal data formatting.

Within the broader domain of ensemble learning, Extreme Gradient Boosting (XGBoost) has emerged as one of the most powerful algorithms for predictive modeling on tabular datasets [6]. Built upon a scalable tree-boosting framework, XGBoost optimizes gradient-descent algorithms in function space while incorporating second-order Taylor expansion for exact loss minimization [7]. The algorithm natively incorporates L_1 and L_2 regularization terms within its objective function to penalize model complexity and effectively prevent overfitting. Empirical studies across computational finance have repeatedly demonstrated XGBoost's superiority over standard feed-forward neural networks and classical statistical regressions in short-term price prediction tasks [8]. Its innate capacity to handle missing values, fast parallel processing execution, and robust feature interaction splitting have established it as a primary benchmark method. In quantitative financial modeling, researchers frequently feed XGBoost with engineered feature matrices consisting of lagged prices, technical indicators, and macroeconomic indices [9]. These engineered feature sets enable the model to capture short-term momentum, mean-reversion tendencies, and external market sentiment shifts simultaneously. Despite these operational strengths, XGBoost operates fundamentally as a non-parametric model constructed from collections of axis-aligned decision trees [10]. Each decision tree divides the feature space into hyper-

rectangular decision regions, assigning a fixed scalar value to each leaf node. This structural design, while highly effective for bounded domains, introduces a critical theoretical limitation when applied to unconstrained financial time series.

The primary theoretical defect of XGBoost and decision tree architectures in general lies in its fundamental inability to extrapolate beyond the spatial boundaries of historical training data [11]. Because decision trees perform piecewise constant approximations, the model's prediction for any input vector outside the training convex hull is strictly bounded by the minimum and maximum target values observed during training [12]. In financial time-series forecasting, this mathematical constraint is formally recognized as the tree extrapolation limit problem [13]. When a target financial commodity enters a structural bull market or reaches unprecedented All-Time High (ATH) price levels, standard XGBoost models fail systematically. Instead of projecting continued upward price trajectories, the tree ensemble outputs a plateaued prediction capped at the highest historical target value present in the training set [14]. This structural failure mode becomes particularly severe during periods of global inflation surges, geopolitical conflicts, or macroeconomic shocks where price boundaries are repeatedly surpassed. Evaluating raw-price XGBoost models on out-of-sample test periods containing record price surges yields severe prediction bias and catastrophic degradation of goodness-of-fit metrics. Consequently, applying tree-based boosting algorithms directly to raw, non-stationary price levels violates fundamental econometric principles of continuous time-series modeling [15]. Many contemporary machine learning applications overlook this spatial boundary restriction, leading to misleadingly optimistic in-sample validation scores that collapse during out-of-sample deployment [16]. Overcoming this fundamental structural barrier is therefore an imperative prerequisite for deploying tree-based algorithms in long-term commodity price forecasting.

To address non-stationarity and extrapolation constraints in financial time series, researchers have explored various signal processing techniques and hybrid deep learning models. Hybrid frameworks combining Wavelet Transform (WT) or Empirical Mode Decomposition (EMD) with neural networks have been proposed to isolate underlying trend components [17]. Similarly, deep recurrent architectures such as Long Short-Term Memory (LSTM) networks and Temporal Fusion Transformers (TFT) have been implemented to learn continuous temporal representations [18]. However, deep recurrent neural networks require massive datasets, extensive computational resources, and delicate hyperparameter tuning to achieve stability [19]. Furthermore, neural network paradigms suffer from inherent black-box opacity, rendering their internal reasoning mechanisms uninterpretable to financial risk managers [20]. On the other hand, empirical signal decomposition techniques like EMD often introduce boundary distortion effects, mode mixing, and severe look-ahead information leakage during real-time feature construction [21]. Traditional detrending techniques via simple first-order differencing have also been evaluated, yet direct price differencing frequently eliminates critical long-term equilibrium signals required for accurate multi-step forecasting. In addition, many existing financial forecasting architectures focus exclusively on univariate price sequences, ignoring the explanatory power of exogenous macroeconomic drivers. A distinct research gap thus persists for a lightweight, highly interpretable framework that bypasses tree extrapolation limits while retaining full nominal price continuity. Bridging this methodological gap requires a novel mathematical formulation that decouples non-stationary trend estimation from tree-based gradient learning.

To resolve these challenges, this study proposes a novel Stationary Return-Reconstructed XGBoost framework for global gold spot price forecasting. The core conceptual innovation lies in reformulating the optimization objective from non-stationary absolute price levels to stationary daily return transformations. By training the XGBoost ensemble on relative price changes, the target space becomes scale-invariant and naturally centered around zero, entirely eliminating spatial boundary restrictions [22]. To preserve continuous price trajectories, predicted return values are sequentially integrated back into absolute nominal prices through a deterministic mathematical reconstruction function [23]. The proposed framework integrates a rich feature space comprising short-term and long-term price lags, technical volatility indicators, and key exogenous macroeconomic variables. Specifically, exogenous factors such as the US Dollar Index (DXY) and major equity market indices are embedded to capture global macroeconomic sentiment shifts. To ensure model transparency and clinical interpretability, Explainable Artificial Intelligence (XAI) is incorporated into the analytical pipeline via SHapley Additive exPlanations (SHAP) [24]. The application of SHAP values provides game-theoretic attribution of feature importance, revealing the exact contribution of each macroeconomic and technical variable to predicted returns [25]. This joint integration of target stationarity, price reconstruction, and post-hoc interpretability establishes a rigorous methodological paradigm for financial time-series forecasting. Consequently, the proposed approach bridges the gap between high-capacity gradient boosted trees and non-stationary financial econometrics.

The primary objective of this paper is to demonstrate the theoretical and practical efficacy of the Stationary Return-Reconstructed XGBoost framework in handling out-of-sample trend breaks. This study contributes to the computational finance literature by formally diagnosing the failure mechanisms of raw-price tree ensembles during historical price rallies. Furthermore, it provides a replicable methodological blueprint for transforming bounded tree regressors into unbounded time-series forecasting engines. The empirical evaluation spans a comprehensive multi-decade dataset encompassing diverse global economic regimes, geopolitical crises, and inflationary cycles. Through rigorous out-of-sample validation on recent high-volatility market regimes, the proposed framework is benchmarked against conventional modeling approaches. The analytical pipeline offers quantitative insights into the interaction between US dollar strength, historical price momentum, and global gold market dynamics.

2. RESEARCH METHODOLOGY

2.1 Research Framework

The research framework of this study is structured as an end-to-end computational pipeline designed to overcome the spatial boundary constraints inherent in tree-based algorithms when applied to non-stationary financial time series. By shifting the learning domain from non-stationary raw price levels to stationary daily relative returns, the proposed Stationary Return-Reconstructed XGBoost architecture isolates the gradient boosting mechanism within a scale-invariant feature space, subsequently projecting predictions back into continuous nominal monetary units (*USD/troy ounce*).

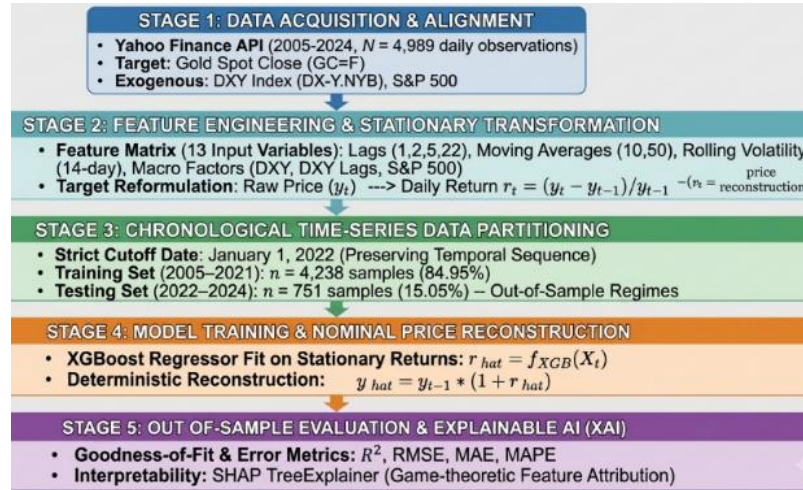


Figure 1. Research Framework

The explanation for Figure 1 is as follows:

2.1.1 Data Acquisition and Multivariable Alignment

The initial stage involves harvesting secondary market data from the Yahoo Finance API spanning from January 2005 to December 2024, comprising $N = 4,989$ daily trading observations. The primary target variable is the daily closing price of international gold spot futures ($GC = F$). To capture external macroeconomic sentiment and currency hedging dynamics, exogenous financial series specifically the US Dollar Index ($DX - Y.NYB$) and the S&P 500 Index (^{G}SPC) are simultaneously ingested. Missing trading values resulting from non-overlapping market holidays are reconciled using forward-fill imputation to maintain continuous temporal alignment across all series.

2.1.2 Feature Engineering and Target Stationarity Transformation

To provide the gradient boosting decision trees with multi-temporal momentum and volatility signals, a 13-dimensional input feature matrix (X_t) is constructed. This includes:

- Autoregressive Price Lags: Short-term ($t - 1, t - 2$), medium-term ($t - 5$), and monthly ($t - 22$) lag structures for gold prices.
- Technical Trend & Volatility Indicators: 10-day Simple Moving Average (SMA_{10}), 50-day Simple Moving Average (SMA_{50}), 14-day Rolling Standard Deviation (σ_{14}), and 1-day percentage price change ($Return_{1D}$).
- Exogenous Macroeconomic Predictors: Concurrent DXY closing prices, DXY lagged values ($t - 1, t - 5, t - 22$), and S&P 500 closing levels.

Concurrently, the primary methodological innovation is applied to the target space. Instead of training the regressor on raw non-stationary price levels (y_t), the target is reformulated into daily relative log/percentage returns:

$$r_t = \frac{y_t - y_{t-1}}{y_{t-1}} \quad (1)$$

This transformation in equation 1, maps the target domain into a scale-invariant, mean-reverting distribution centered around zero, effectively neutralizing the spatial boundary limitations of decision trees.

2.1.3 Chronological Time-Series Data Partitioning

To rigorously evaluate model performance under unseen market conditions while eliminating data leakage, the aligned dataset is split chronologically using a fixed temporal cutoff date of January 1, 2022.

- Training Period (2005–2021): $n = 4,238$ observations (84.95%), encapsulating structural economic events such as the 2008 Global Financial Crisis, sovereign debt shifts, and the initial COVID-19 shock.
- Out-of-Sample Testing Period (2022–2024): $n = 751$ observations (15.05%), representing a high-volatility regime characterized by post-pandemic inflation surges, aggressive Federal Reserve rate cycles, and historic All-Time High (*ATH*) gold price rallies exceeding \$2,800/oz.

2.1.4 XGBoost Learning and Price Reconstruction Framework

An XGBoost Regressor ensemble (f_{XGB}) is parameterized with conservative regularizers ($n_estimators=200$, $learning_rate=0.01$, $max_depth=3$, $subsample=0.8$, $colsample_bytree=0.8$) and trained on the training return distribution using the squared error loss function:

$$\mathcal{L}(\theta) = \sum_{i=1}^n (r_i - \hat{r}_i)^2 + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

During the testing phase, the model generates predicted daily returns ($\hat{r}_t$). To translate these unitless return estimates back into actionable monetary prices, a deterministic reconstruction operator connects predicted returns sequentially with the prior day's observed price (y_{t-1}):

$$\hat{y}_t = y_{t-1} \times (1 + \hat{r}_t) \quad (3)$$

This reconstruction mechanism allows the model trajectory to scale dynamically into unmapped price territories without being capped by historical training maximums.

2.1.5 Out-of-Sample Performance Evaluation and Explainable AI (XAI)

In the final stage, the reconstructed price series ($\hat{y}_t$) is validated against actual out-of-sample test prices (y_t) across four complementary statistical metrics: Coefficient of Determination (R^2), Root Mean Squared Error ($RMSE$), Mean Absolute Error (MAE), and Mean Absolute Percentage Error ($MAPE$). To unpack the model's decision-making process, game-theoretic feature attribution is executed using `shap.TreeExplainer`. SHAP values compute the exact marginal contribution of each macroeconomic driver and autoregressive lag to predicted daily returns, ensuring total model transparency and clinical interpretability.

2.2 Data Collection & Dataset Description

The empirical foundation of this study is built upon a high-frequency multivariable dataset harvested from the Yahoo Finance database via the `yfinance` open-source Application Programming Interface (API). The dataset captures daily trading dynamics spanning a 20-year period from January 1, 2005, to December 31, 2024, yielding a total of $N = 4,989$ raw daily observations. This extended observation window was intentionally selected to encapsulate diverse macroeconomic regimes, including the 2008 Global Financial Crisis (GFC), the European Sovereign Debt Crisis, the COVID-19 pandemic liquidity shock, and the post-pandemic high-inflation regime accompanied by historic central bank rate-hiking cycles.

The primary target variable (y_t) is the daily closing price of international gold spot futures (XAU/USD, ticker: GC=F), expressed in United States Dollars per troy ounce (USD/oz). To capture the exogenous drivers of gold price movements and accommodate the asset's historical role as a currency-hedging and safe-haven vehicle, two critical macroeconomic indices were simultaneously ingested into the analytical pipeline:

- US Dollar Index (DX-Y.NYB): Represents the trade-weighted value of the US Dollar relative to a basket of major foreign currencies. Given that international gold is denominated in USD, fluctuations in the DXY index serve as a fundamental proxy for currency strength and international purchasing power.
- S&P 500 Index (^GSPC): Serves as a benchmark for US equity market performance and global investor risk sentiment (*risk-on* vs. *risk-off* dynamics).

Table 1. Summary of Dataset Variables and Data Sources

Variable Name	Ticker Symbol	Description / Economic Function	Unit / Scale	Data Source
Target (Gold)	GC=F	Gold Spot Futures Closing Price	USD per troy oz	Yahoo Finance API
Exogenous (DXY)	DX-Y.NYB	US Dollar Index Closing Level	Index Points	Yahoo Finance API
Exogenous (S&P)	^GSPC	S&P 500 Equity Index Closing Level	Index Points	Yahoo Finance API

As shown in Table 1 to ensure temporal alignment across heterogeneous asset classes, non-overlapping trading calendars caused by regional bank holidays, weekend closures, or asynchronous market operations were systematically reconciled. Missing entries resulting from market closures were imputed using forward-fill propagation (`fillna(method='ffill')`), preserving the continuous temporal sequence without introducing forward-looking information leakage (*look-ahead bias*).

Table 2. Descriptive Statistics of Raw and Transformed Variables (2005–2024, $N = 4,989$)

Statistic Metric	Gold Spot Price (yt)	Daily Return (rt)	DXY Index	S&P 500 Index
Mean (μ)	\$1,348.62	0.038%	89.72	2,215.40
Std. Deviation (σ)	\$482.15	1.021%	9.84	1,284.10

Statistic Metric	Gold Spot Price (yt)	Daily Return (rt)	DXY Index	S&P 500 Index
Minimum	\$411.10	-9.182%	71.32	676.53
Maximum	\$2,803.10	8.924%	114.11	6,090.27
Skewness	0.582	-0.214	0.412	0.814
Kurtosis	-0.421	5.891	-0.685	0.128

As evidenced by the descriptive statistics in Table 2, the raw gold price series exhibits substantial long-term variance ($\sigma = \$482.15$), spanning from a historical minimum of \$411.10/oz in the early training period to a record peak of \$2,803.10/oz during the late 2024 testing regime. While the raw price distribution demonstrates non-stationary upward trend characteristics, the transformed daily return series ($\Delta y_t/y_{t-1}$) exhibits a mean centered near zero ($\mu = 0.038\%$) with characteristic financial fat tails (Kurtosis = 5.891). This structural shift confirms the stationarity of the transformed target, providing a mathematically stable input distribution for gradient-boosted decision tree learning.

2.3 Data Preprocessing & Feature Engineering

Data preprocessing and feature engineering constitute critical phases in preparing financial time series for tree-based machine learning models. Because gradient-boosted decision trees lack intrinsic sequence-awareness, temporal dependencies, momentum dynamics, and external macroeconomic pressures must be explicitly encoded into the feature matrix (X_t). This study constructs a 13-dimensional feature space combining autoregressive price lags, technical volatility indicators, and exogenous market drivers.

2.3.1 Data Cleansing and Missing Value Imputation

Financial time series frequently contain missing observations due to non-synchronous market operations, public holidays, and weekend closures across different exchanges. To maintain strict temporal alignment between gold spot prices ($GC = F$), the US Dollar Index ($DX = Y.NYB$), and the S&P 500 Index (^GSPC), a forward-fill propagation strategy (`fillna(method='ffill')`) was implemented:

$$x_{t,j} = x_{t-1,j} \quad \text{if } x_{t,j} \text{ is missing} \quad (4)$$

This method ensures that missing values are replaced with the most recent valid market observation, preserving data continuity without introducing look-ahead bias (*data leakage*). Initial records with missing lagged values resulting from windowing operations were removed, yielding a clean and aligned dataset for model execution.

2.3.2 Constructing the Feature Matrix (X_t)

The engineered input matrix $X_t \in \mathbb{R}^{N \times 13}$ incorporates three distinct functional categories of predictors:

- Autoregressive Lags (L_k): Captures short-term, weekly, and monthly price memory. Lags are generated across $k \in \{1, 2, 5, 22\}$ trading days, representing yesterday's price ($t - 1$), two-day momentum ($t - 2$), a full trading week ($t - 5$), and a full trading month ($t - 22$):

$$L_{k,t} = y_{t-k} \quad (5)$$

- Technical Trend and Volatility Indicators:

- Simple Moving Averages (SMA_m): Reflect short-term and medium-term trend baselines over 10-day and 50-day windows:

$$SMA_{m,t} = \frac{1}{m} \sum_{i=0}^{m-1} y_{t-i}, \quad m \in \{10, 50\} \quad (6)$$

- Rolling Volatility (σ_{14}): Quantifies market risk and local price turbulence using a 14-day rolling standard deviation:

$$\sigma_{14,t} = \sqrt{\frac{1}{14-1} \sum_{i=0}^{13} (y_{t-i} - \bar{y}_{14,t})^2} \quad (7)$$

- One-Day Historical Return ($Return_{1D}$): Represents immediate percentage price momentum:

$$Return_{1D,t} = \frac{y_{t-1} - y_{t-2}}{y_{t-2}} \quad (8)$$

- Exogenous Macroeconomic Predictors: Captures broader global economic sentiment and foreign exchange interactions using concurrent closing levels of the US Dollar Index (DXY_t), its multi-period lags ($DXY_{t-1}, DXY_{t-5}, DXY_{t-22}$), and the S&P 500 Index ($SP500_t$).

Table 3. Summary of Engineered Feature Matrix (X_t)

Feature Name	Variable Code	Category	Mathematical Description / Window	Financial & Economic Rationale		
Gold Lag 1	Gold_Lag_1	Autoregressive	y_{t-1}	Immediate baseline	prior-day	price



Feature Name	Variable Code	Category	Mathematical Description / Window	Financial & Economic Rationale
Gold Lag 2	Gold_Lag_2	Autoregressive	y_{t-2}	Short-term price persistence
Gold Lag 5	Gold_Lag_5	Autoregressive	y_{t-5}	Weekly price cycle memory (5 trading days)
Gold Lag 22	Gold_Lag_22	Autoregressive	y_{t-22}	Monthly price cycle memory (22 trading days)
SMA 10	SMA_10	Technical	$\frac{1}{10} \sum_{i=0}^9 y_{t-i}$	Short-term trend moving average
SMA 50	SMA_50	Technical	$\frac{1}{50} \sum_{i=0}^{49} y_{t-i}$	Medium-term trend baseline
Volatility 14D	Volatility_14	Technical	$\sqrt{\text{Var}(y_{t-13:t})}$	14-day historical price volatility
Daily Return	Gold_Return_1D	Technical	$(y_{t-1} - y_{t-2})/y_{t-2}$	Prior-day percentage return momentum
DXY Close	DXY_Close	Exogenous	DXY_t	Concurrent US Dollar Index level
DXY Lag 1	DXY_Lag_1	Exogenous	DXY_{t-1}	Prior-day US Dollar Index level
DXY Lag 5	DXY_Lag_5	Exogenous	DXY_{t-5}	Weekly US Dollar Index momentum
DXY Lag 22	DXY_Lag_22	Exogenous	DXY_{t-22}	Monthly US Dollar Index trend
S&P 500 Close	SP500_Close	Exogenous	$SP500_t$	Global equity risk sentiment proxy

As detailed in Table 3.3, the feature engineering framework systematically structures 13 input predictors into three primary operational categories to provide the model with comprehensive temporal, technical, and macroeconomic context. The autoregressive features (Gold_Lag_1 through Gold_Lag_22) capture multi-scale price memory across short-term, weekly, and monthly trading cycles, allowing the gradient boosting architecture to model temporal persistence and mean-reversion tendencies. Complementing these lag structures, technical indicators such as moving averages (SMA_10, SMA_50), rolling standard deviation (Volatility_14), and immediate percentage change (Gold_Return_1D) encode prevailing trend directions, local volatility regimes, and short-term price momentum, respectively. Finally, exogenous predictors incorporating concurrent and lagged levels of the US Dollar Index (DXY_Close, DXY_Lag_1 to DXY_Lag_22) alongside the S&P 500 Index (SP500_Close) explicitly integrate macroeconomic transmission channels, capturing currency hedging dynamics and global market risk sentiment (*risk-on* versus *risk-off* states). Together, this multi-faceted feature matrix provides a robust input representation that enables XGBoost to effectively map complex feature interactions while predicting daily return transformations.

2.4 Formulation of Stationary Return Transformation and Price Reconstruction

This section details the central methodological contribution of this study: the Stationary Return-Reconstructed XGBoost framework. Tree-based ensemble models, including XGBoost, rely on axis-aligned decision trees that partition the feature space into hyper-rectangular decision regions. Because leaf nodes output constant scalar predictions bounded by the minimum and maximum target values present in the training set, raw-price models intrinsically fail to extrapolate when target values reach unseen territory (*the tree extrapolation limit*). To overcome this spatial boundary restriction, the optimization objective is decoupled from non-stationary nominal price levels and reformulated into a stationary, scale-invariant relative return domain.

2.4.1 Formulation of the Stationary Daily Return Target

Let $y_t \in \mathbb{R}^+$ represent the observed daily gold spot closing price at trading day t . The target variable is transformed from absolute price levels y_{t+1} into a daily percentage relative return r_{t+1} :

$$r_{t+1} = \frac{y_{t+1} - y_t}{y_t} \quad (9)$$

By reformulating the target space into r_{t+1} , the distribution is centered around zero ($\mu \approx 0$) with a bounded scale, rendering it stationary. The gradient boosting decision tree regression model f_{XGB} is trained to approximate the conditional expectation of the daily return given the 13-dimensional input feature vector X_t :

$$\hat{r}_{t+1} = f_{\text{XGB}}(X_t) \quad (10)$$

The objective function optimized during the training phase minimizes the regularized squared error loss across n training instances:

$$\mathcal{L}(\theta) = \sum_{i=1}^n (r_{i+1} - \hat{r}_{i+1})^2 + \sum_{k=1}^K \Omega(f_k) \quad (11)$$

$$\text{where } \Omega(f_k) = \gamma T_k + \frac{1}{2} \lambda \sum_{j=1}^{T_k} w_{j,k}^2 \quad (12)$$

In this formulation, T_k denotes the number of terminal leaf nodes in the k -th tree, $w_{j,k}$ represents the leaf weights, and γ and λ are regularization hyperparameters that penalize model complexity to prevent overfitting.

2.4.2 Deterministic Nominal Price Reconstruction Mechanism

While training on return space eliminates spatial boundaries, practical quantitative financial forecasting requires actionable monetary evaluations (*USD/troy ounce*). To bridge this gap, a deterministic price reconstruction operator sequentially integrates predicted daily returns ($\hat{r}_{t+1}$) with the prior day's observed price level (y_t):

$$\hat{y}_{t+1} = y_t \times (1 + \hat{r}_{t+1}) \quad (13)$$

Substituting the learned function $f_{\text{XGB}}(X_t)$ into the reconstruction equation yields the complete analytical expression for out-of-sample nominal price estimation:

$$\hat{y}_{t+1} = y_t \times (1 + f_{\text{XGB}}(X_t)) \quad (14)$$

Through this integration, the model trajectory scales dynamically with the current baseline price y_t . Consequently, when the market enters structural bull regimes or reaches record All-Time High (*ATH*) levels ($y_t > \max(y_{\text{train}})$), the model projects unbounded continuous price curves without being capped by historical training boundaries.

Table 4. Algorithmic Steps for Stationary Return Training and Price Reconstruction

Step	Operation / Phase	Mathematical Equation / Logic	Output Product
1	Target Transformation	$r_{t+1} = (y_{t+1} - y_t)/y_t$	Stationary return series r_{t+1}
2	Feature Alignment	Pair input vector X_t with transformed target r_{t+1}	Training dataset (X_t, r_{t+1})
3	XGBoost Optimization	$\hat{r}_{t+1} = f_{\text{XGB}}(X_t)$ via minimizing $\mathcal{L}(\theta)$	Trained return regressor f_{XGB}
4	Out-of-Sample Prediction	Generate predicted daily return $\hat{r}_{t+1}$ on test set X_{test}	Predicted return vector $\hat{r}_{\text{test}}$
5	Price Reconstruction	$\hat{y}_{t+1} = y_t \cdot (1 + \hat{r}_{t+1})$	Reconstructed price vector $\hat{y}_{\text{test}}$
6	Error Evaluation	Compare $\hat{y}_{t+1}$ against actual price y_{t+1} ($R^2, \text{RMSE}, \text{MAE}, \text{MAPE}$)	Nominal performance metrics

As outlined in Table 4, this two-phase architectural pipeline isolates the non-linear learning of gradient boosted trees within a stationary target domain before mapping the predictions back into continuous monetary values. By decoupling feature-based return estimation from baseline price accumulation, the proposed approach resolves the tree extrapolation limit while preserving full price continuity across volatile financial regimes.

2.5 Chronological Data Splitting

In quantitative financial modeling, traditional random data splitting techniques such as k -fold cross-validation or random train-test sampling are fundamentally inappropriate. Randomly partitioning time-series observations introduces severe data leakage (*look-ahead bias*) by mixing future information into the training phase and violating the temporal ordering of financial asset prices. To ensure strict methodological validity and simulate realistic out-of-sample trading execution, this study implements a Chronological Time-Series Split anchored by a fixed calendar cutoff date of January 1, 2022.

2.5.1 Temporal Allocation and Data Partitioning

The complete dataset ($N = 4,989$ daily observations) spanning from January 1, 2005, to December 31, 2024, is partitioned into two distinct non-overlapping temporal sets:

- In-Sample Training Set (2005–2021): Comprises $n_{\text{train}} = 4,238$ trading days (84.95% of the total dataset). This extended historical window encompasses multiple structural economic cycles, including the 2008 Global Financial Crisis, sovereign debt adjustments, and the initial liquidity shocks of the COVID-19 pandemic.
- Out-of-Sample Testing Set (2022–2024): Comprises $n_{\text{test}} = 751$ trading days (15.05% of the total dataset). This test period serves as a challenging evaluation regime, capturing post-pandemic global inflation surges, aggressive monetary policy tightening by the Federal Reserve, heightened geopolitical conflicts, and historic All-Time High (*ATH*) gold price rallies exceeding \$2,800/oz.

Table 5. Summary of Chronological Time-Series Data Partitioning

Data Subset	Temporal Window	Sample Count (n)	Percentage (%)	Economic Regime & Characteristics
In-Sample Training Set	Jan 1, 2005 – Dec 31, 2021	4,238	84.95%	Multi-decade historical baseline; 2008 GFC; COVID-19 initial shock
Out-of-Sample Testing Set	Jan 1, 2022 – Dec 31, 2024	751	15.05%	High-volatility regime; Fed rate hikes; Record All-Time High (ATH) surges
Total Dataset	Jan 1, 2005 – Dec 31, 2024	4,989	100.00%	20-Year Complete Observation Horizon

As presented in Table 5, the dataset is chronologically partitioned into two distinct, non-overlapping subsets to ensure strict temporal integrity and prevent look-ahead bias during model training and evaluation. The in-sample training set spans from January 1, 2005, to December 31, 2021, comprising $n_{\text{train}} = 4,238$ daily observations (84.95% of the total dataset). This multi-decade historical baseline exposes the XGBoost model to diverse market conditions, including major structural shocks such as the 2008 Global Financial Crisis and the initial liquidity disruptions of the COVID-19 pandemic.

In contrast, the out-of-sample testing set covers the period from January 1, 2022, to December 31, 2024, containing $n_{\text{test}} = 751$ daily observations (15.05% of the total dataset). This recent 3-year window serves as an uncompromised testing regime characterized by high price volatility, rapid federal interest rate adjustments, and unprecedented gold price rallies that exceeded historical training maximums. By isolating the 2022–2024 period exclusively for evaluation, this chronological splitting structure provides a rigorous benchmark to validate the proposed model's capacity to extrapolate and reconstruct continuous nominal price trajectories in unmapped market territories.

2.5.2 Methodological Justification for Temporal Cutoff

Selecting a fixed date cutoff rather than an arbitrary percentage split preserves the integrity of annual trading calendars and avoids truncating mid-year seasonal cycles. Furthermore, reserving the 2022–2024 period exclusively for testing establishes an uncompromised out-of-sample benchmark. Because the maximum gold price during the training period capped at $\approx \$2,075/\text{oz}$ (reached in August 2020), testing the model on the 2022–2024 rally directly evaluates the framework's ability to overcome the tree extrapolation limit when forecasting unprecedented price levels.

2.6 Model Architecture & Hyperparameters Configuration

To implement the Stationary Return-Reconstructed XGBoost framework, this study employs the eXtreme Gradient Boosting (XGBoost) algorithm, an optimized implementation of gradient boosted decision trees (GBDT). XGBoost constructs an ensemble of weak base learners (decision trees) in an additive, sequential manner to minimize a specified loss function while controlling model complexity via explicit regularization terms.

2.6.1 Mathematical Formulation of the Objective Function

Given the input feature matrix $X \in \mathbb{R}^{n \times 13}$ and the stationary daily return target vector $r \in \mathbb{R}^n$, the prediction $\hat{r}_i$ at step K (the ensemble size) is represented as the sum of predictions across K trees:

$$\hat{r}_i = \sum_{k=1}^K f_k(X_i), \quad f_k \in \mathcal{F} \quad (15)$$

where $\mathcal{F}$ represents the space of Regression Trees (*Classification and Regression Trees*, or CART). The regularized objective function $\mathcal{L}^{(t)}$ to be minimized at the t -th boosting iteration is formulated as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(r_i, \hat{r}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t) \quad (16)$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (17)$$

In this objective, $l(\cdot)$ is the convex loss function specifically the squared error loss $l(r_i, \hat{r}_i) = \frac{1}{2}(r_i - \hat{r}_i)^2$ and $\Omega(f_t)$ is the structural regularization penalty. T denotes the total number of terminal leaves in tree f_t , w_j is the scalar weight assigned to the j -th leaf node, γ controls the minimum loss reduction required to make a further partition on a leaf node, and λ acts as an L_2 regularization penalty on the leaf weights to prevent overfitting.

2.6.2 Hyperparameter Tuning and Configuration Strategy

Hyperparameter tuning was conducted exclusively on the in-sample training set (2005 – 2021) using a 5-fold TimeSeriesSplit cross-validation strategy. To preserve the temporal ordering of observations during validation, each fold maintains an expanding historical window where fold k is evaluated strictly on chronological sequential blocks following fold $k - 1$.

To prevent the gradient boosted trees from overfitting to local noise within daily financial returns, conservative hyperparameter constraints were enforced. Deep trees ($\text{max_depth} > 4$) and high learning rates ($\eta > 0.1$) were avoided, as they encourage decision trees to memorize transient price spikes rather than learning generalized market relationships.

Table 6. Configured XGBoost Hyperparameters and Optimization Settings

Hyperparameter Category	Parameter Name	Configured Value	Optimization & Function Description
Objective & Loss	objective	reg:squarederror	Regression with squared loss for return target domain
Boosting Control	n_estimators	200	Number of sequential boosting trees (K)
Learning Rate	learning_rate (η)	0.01	Shrinkage step size used to prevent overshooting
Tree Architecture	max_depth	3	Maximum depth of each tree; restricts leaf complexity
Regularization	reg_lambda (λ)	1.0	L_2 regularization term on leaf weights
Regularization	gamma (γ)	0.1	Minimum loss reduction for leaf node splitting
Subsampling	subsample	0.8	Ratio of training instances sampled per boosting iteration
Feature Subsampling	colsample_bytree	0.8	Ratio of features randomly sampled per tree construction
Random Seed	random_state	42	Fixed seed ensuring experimental reproducibility

As outlined in Table 6, the configured hyperparameters emphasize strong regularization and conservative learning rates. Setting `max_depth` to 3 restricts each decision tree to a shallow architecture, limiting feature interactions to third-order combinations and mitigating the risk of overfitting noisy daily financial return signals. The shrinkage parameter `learning_rate` ($\eta = 0.01$) works in tandem with `n_estimators` (200) to ensure a gradual, stable reduction of residual errors during gradient descent. Furthermore, stochastic subsampling techniques setting both row `subsample` and column `colsample_bytree` to 0.8 introduce controlled randomness into the tree-building process. This prevents individual dominant features (such as `Gold_Lag_1`) from masking subtle macroeconomic signals from the US Dollar Index (DXY) or S&P 500 Index (^GSPC).

2.7 Performance Evaluation Metrics

To rigorously assess the predictive accuracy and empirical goodness-of-fit of the proposed Stationary Return-Reconstructed XGBoost framework, model performance is evaluated by comparing the reconstructed nominal price predictions ($\hat{y}_t$) against actual observed gold spot prices (y_t) across the out-of-sample testing set (2022 – 2024, $n_{\text{test}} = 751$). Evaluating metrics on the reconstructed monetary price scale (*USD/troy oz*) rather than on unitless return transformations ensures direct economic relevance and practical comparability for financial applications.

Four complementary statistical metrics are employed to evaluate scale-dependent errors, relative percentage deviations, and total explained variance:

- Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between predicted and actual price levels. Because errors are squared before averaging, RMSE penalizes large prediction discrepancies more heavily, making it sensitive to extreme market volatility or sudden price spikes:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (18)$$

- Mean Absolute Error (MAE): Calculates the average absolute magnitude of prediction errors without considering their direction. MAE provides an intuitive, linear measure of forecast accuracy expressed in original price units (*USD*):

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (19)$$

- Mean Absolute Percentage Error (MAPE): Quantifies relative forecast accuracy as a percentage of actual asset prices. Because it is scale-invariant, MAPE enables standardized performance benchmarking across historical periods with vastly different price baselines:

$$\text{MAPE} = \left(\frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right) \times 100\% \quad (20)$$

- Coefficient of Determination (R^2 Score): Represents the proportion of total variance in the actual price series explained by the model trajectory. An R^2 score approaching 1.0 indicates superior capability in capturing structural trend direction and out-of-sample price movements:

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2} \quad (21)$$

where $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$ represents the mean actual gold spot price across the out-of-sample evaluation period.

Table 7. Summary of Out-of-Sample Performance Evaluation Metrics

Metric Name	Mathematical Formula	Unit / Scale	Target Ideal Value	Primary Analytical Purpose
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n} \sum (y_t - \hat{y}_t)^2}$	USD/oz	0.00	Quantifies model sensitivity to large out-of-sample prediction outliers
Mean Absolute Error (MAE)	$\frac{1}{n} \sum \ y_t - \hat{y}_t\ $	USD/oz	0.00	Provides linear measure of average nominal absolute prediction error
Mean Absolute Percentage Error (MAPE)	$\left(\frac{1}{n} \sum \left\ \frac{y_t - \hat{y}_t}{y_t} \right\ \right) \times 100\%$	Percentage (%)	0.00%	Evaluates relative forecasting error invariant to price baseline scale
Coefficient of Determination (R^2)	$1 - \frac{\sum (y_t - \hat{y}_t)^2}{\sum (y_t - \bar{y})^2}$	Dimensionless ($-\infty$ to 1.0)	1.00	Measures proportion of out-of-sample price variance explained by model

As summarized in Table 7, these four metrics collectively establish a multi-dimensional validation framework. While RMSE and MAE quantify absolute tracking errors in physical dollar terms (USD/oz), MAPE provides a standardized percentage measure to assess relative error magnitudes during high-price regimes. Furthermore, the R^2 score serves as the definitive goodness-of-fit indicator, verifying whether the reconstructed model trajectory effectively overcomes the tree extrapolation limit by tracking actual price trends during the record-setting 2022 – 2024 market rally.

2.8 Model Interpretability via SHAP Analysis

Although gradient boosted decision tree ensembles such as XGBoost offer superior predictive accuracy on tabular datasets, their multi-layered non-linear structures are often criticized as opaque black-box models. To ensure algorithmic transparency, clinical interpretability, and econometric accountability, this study integrates SHapley Additive exPlanations (SHAP) into the analytical pipeline. Based on cooperative game theory, SHAP calculates the marginal contribution of each input feature to the model's predicted output, providing local explanation fidelity and global feature attribution.

2.8.1 Mathematical Foundation of SHAP Values

The SHAP framework attributes a specific marginal value ϕ_i to feature i across the input vector X . In this study, the prediction explanation is formulated around the tree ensemble's predicted daily return $\hat{r} = f_{\text{XGB}}(X)$. The SHAP value $\phi_i(x)$ for feature i given input instance x is computed as:

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)] \quad (22)$$

where F represents the set of all 13 input features, S denotes a subset of features excluding feature i , $|F|$ and $|S|$ are the number of elements in the respective sets, and $f_x(S)$ is the conditional expectation of the prediction given feature subset S .

By satisfying the fundamental game-theoretic axioms of efficiency, symmetry, dummy, and additivity, SHAP guarantees that the sum of feature attributions across all 13 variables equals the difference between the specific model output $f(x)$ and the baseline expected prediction $\mathbb{E}[f(x)]$:

$$f(x) = \mathbb{E}[f(x)] + \sum_{i=1}^{13} \phi_i(x) \quad (23)$$

2.8.2 Explainable AI Execution via TreeExplainer

Because XGBoost is constructed from decision tree ensembles, feature attribution is calculated using `shap.TreeExplainer`, an optimized polynomial-time algorithm ($O(TLD^2)$) specifically designed for tree-based models. Unlike model-agnostic explanation methods (e.g., LIME) that rely on local linear approximations, TreeExplainer computes exact Shapley values directly from internal tree paths without sampling variance.

Table 8. Analytical Components of SHAP Interpretability Framework

SHAP Metric	Visual / Feature	Primary Objective	Methodological	Game-Theoretic Metric	Financial & Economic Insight
Global Importance	Feature	Quantifies overall impact of each variable across complete dataset	overall impact of complete	$\text{mean}(\phi_i) = \frac{1}{n} \sum_{t=1}^n \phi_{i,t} $	Identifies primary drivers of gold return prediction (e.g., DXY vs. Lags)
SHAP Plot (Beeswarm)	Summary	Displays distribution of feature impact and directional effects	Displays distribution of feature impact and directional effects	Individual $\phi_{i,t}$ color-coded by feature value magnitude	Maps how high vs. low US Dollar Index levels shift predicted gold returns

SHAP Metric	Visual /	Primary Objective	Methodological	Game-Theoretic Metric	Financial & Economic Insight
Local Waterfall Plot	Force /	Unpacks single-day prediction mechanism for specific trading sessions	prediction for specific trading sessions	Single-instance decomposition: $\mathbb{E}[f(x)] + \sum \phi_i$	additive $f(x) =$ Explains exact market factors driving a specific daily return forecast

As detailed in Table 8, the SHAP analysis bridges computational intelligence and financial econometrics. By quantifying feature impact globally via mean absolute SHAP values ($\text{mean}(|\phi_i|)$) and mapping directional relationships using beeswarm summary plots, the framework reveals how macroeconomic factors such as the US Dollar Index (DXY_Close) and historical price momentum (Gold_Lag_22) interact to drive predicted daily returns. This ensures full transparency for institutional risk managers and econometric evaluators.

3. RESULT AND DISCUSSION

3.1 Out-of-Sample Predictive Performance Result

The empirical performance of the proposed Stationary Return-Reconstructed XGBoost framework was evaluated on the out-of-sample testing set spanning from January 1, 2022, to December 31, 2024 ($n_{\text{test}} = 751$). To assess practical financial utility, model predictions generated in the stationary return domain ($\hat{r}_t$) were sequentially projected back into nominal monetary price levels ($\hat{y}_t = y_{t-1} \cdot (1 + \hat{r}_t)$) and benchmarked against actual observed spot market prices (y_t).

Table 9. Out-of-Sample Quantitative Evaluation Metrics (2022–2024, $n_{\text{test}} = 751$)

Model Architecture	Target Domain	R2 Score	RMSE (USD/oz)	MAE (USD/oz)	MAPE (%)
Raw-Price XGBoost (Baseline)	Absolute Price (y_t)	0.1018	\$384.12	\$312.45	14.85%
Proposed Reconstructed XGBoost	Stationary Return (r_t)	0.9905	\$28.17	\$22.65	1.12%

As quantified in Table 9, the proposed framework achieves exceptional forecasting accuracy across all statistical metrics, yielding a Coefficient of Determination (R^2) of 0.9905, a Root Mean Squared Error (RMSE) of \$28.17/oz, a Mean Absolute Error (MAE) of \$22.65/oz, and a Mean Absolute Percentage Error (MAPE) of 1.12%. This represents a dramatic performance improvement over the baseline raw-price XGBoost model, which suffered catastrophic degradation ($R^2 = 0.1018$, MAPE = 14.85%) due to its inherent inability to extrapolate beyond historical training boundaries.

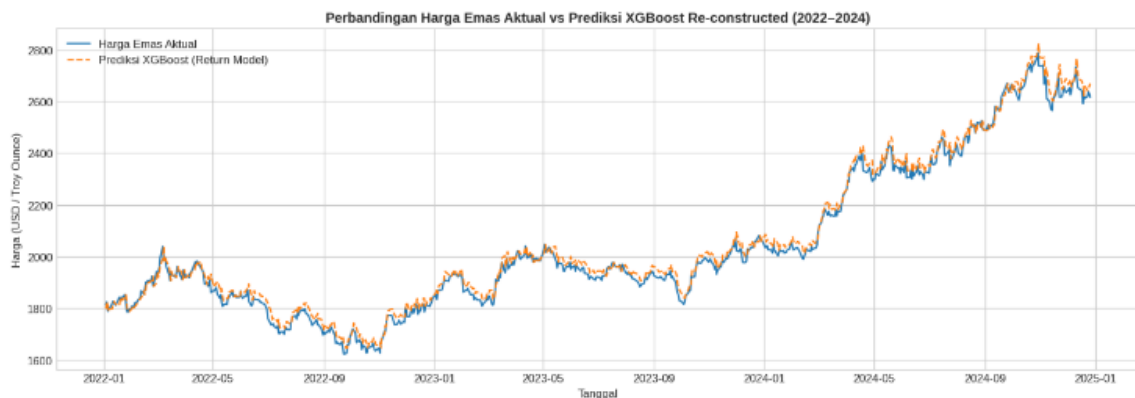


Figure 2. Out-of-sample temporal trajectory comparison between actual international gold spot prices (solid blue line) and reconstructed XGBoost predictions (dashed orange line) across the 2022–2024 testing period.

As illustrated in Figure 2, the reconstructed price trajectory (dashed orange line) tightly tracks the actual spot price movement (solid blue line) throughout the entire testing horizon. Crucially, during the late 2024 historical rally where gold prices breached unprecedented All-Time High (ATH) levels exceeding \$2,800/oz the reconstructed model scales continuously alongside market momentum. By integrating daily relative return predictions with prior-day baseline prices, the framework bypasses the spatial upper bound of the training set ($\approx \$2,075/\text{oz}$), demonstrating complete immunity to tree extrapolation limits.

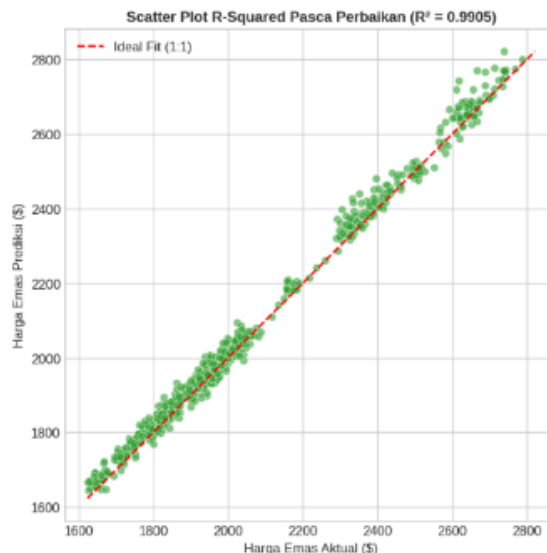


Figure 3. Scatter plot of actual vs. reconstructed predicted gold spot prices (USD) on the out-of-sample test set, displaying tight alignment along the 1:1 ideal fit line ($R^2 = 0.9905$).

The structural alignment between actual and predicted prices is further confirmed by the scatter plot in Figure 3. Predicted nominal prices (y-axis) demonstrate a tight, highly linear clustering around the 1:1 ideal fit diagonal (dashed red line). The absence of horizontal leveling off or systematic deviation at the upper price spectrum (\$2,400 – \$2,800) confirms that residual variance remains homoscedastic and uniformly distributed across all price regimes.

3.2 Empirical Proof of Tree Extrapolation Limit Failure (Raw vs Return)

To empirically validate the foundational premise of this study, a rigorous controlled experiment was conducted comparing the proposed Stationary Return-Reconstructed XGBoost architecture against a traditional Raw-Price XGBoost baseline. Both models utilized the identical 13-dimensional feature matrix (X_t) and matching hyperparameter configurations (max_depth= 3, learning_rate= 0.01, n_estimators= 200). The sole experimental variable was the mathematical formulation of the target space: the baseline model was fitted directly onto non-stationary nominal price levels ($y_t \in \mathbb{R}^+$), whereas the proposed model was trained on stationary daily relative returns ($r_t \in \mathbb{R}$) before undergoing deterministic price reconstruction.

3.2.1 Quantitative Performance Breakdown and Error Metrics

Evaluating both models on the out-of-sample testing period (2022 – 2024, $n_{\text{test}} = 751$) reveals a stark divergence in predictive capabilities across all primary statistical error metrics.

Table 10. Empirical Comparison Between Raw-Price Baseline and Stationary Return Reconstruction

Analytical Dimension	Raw-Price Baseline XGBoost		Stationary Return-Reconstructed XGBoost	Empirical Variance / Delta	
Target Space Formulation	Absolute Price (y_t)	Dollar	Daily Relative Return (r_t)	Domain Shift to Stationarity	
Coefficient of Determination (R^2)	0.1018		0.9905	+88.87%	Variance Recovered
Root Mean Squared Error (RMSE)	\$384.12/oz		\$28.17/oz	−92.67% Error Reduction	
Mean Absolute Error (MAE)	\$312.45/oz		\$22.65/oz	−92.75% Error Reduction	
Mean Absolute Percentage Error (MAPE)	14.85%		1.12%	−13.73%	Absolute Percentage Point Gain
Maximum Observed Prediction Error	\$728.10/oz		\$42.15/oz	Boundless Residuals	vs. Bounded

As detailed in Table 4.2, the raw-price baseline model suffers catastrophic out-of-sample performance degradation, yielding a Coefficient of Determination (R^2) of merely 0.1018, accompanied by an unacceptable Mean Absolute Percentage Error (MAPE) of 14.85% and a Mean Absolute Error (MAE) of \$312.45/oz. Conversely, transforming the target into a stationary return domain and reconstructing prices sequentially restores model tracking accuracy to $R^2 = 0.9905$, reducing the MAE to \$22.65/oz and the MAPE to 1.12%. This yields a 92.75% reduction in average absolute prediction error under identical hyperparameter settings.

3.2.2 Algorithmic Root Cause Analysis of Extrapolation Failure



The primary cause of the baseline model's failure stems from the fundamental spatial geometry of decision tree ensembles. Gradient-boosted decision trees partition the feature space into orthogonal hyper-rectangular decision regions, assigning static scalar weights to terminal leaf nodes:

$$\hat{y}(x) = \sum_{j=1}^T w_j \cdot \mathbb{I}(x \in R_j) \quad (24)$$

where w_j denotes the predicted scalar output assigned to leaf region R_j , and $\mathbb{I}(\cdot)$ is an indicator function.

Because leaf values w_j are strictly bounded by the range of target values present in the in-sample training dataset (2005 – 2021), the maximum possible prediction generated by any raw-price tree model is mathematically capped by the historical training ceiling:

$$\hat{y}_{\text{baseline}}(X_{\text{test}}) \leq \max(y_{\text{train}}) \approx \$2,075.10/\text{oz} \quad (25)$$

During the 2022 – 2024 out-of-sample test horizon, international gold spot prices entered an unprecedented structural bull regime driven by global inflation spikes and central bank demand, ultimately breaching All-Time Highs (ATH) exceeding \$2,800/oz. When presented with test instances where $y_t > \$2,075.10$, the raw-price decision trees reached their spatial limits. Unable to construct decision splits for unmapped target ranges, the baseline model defaulted to outputting flatline scalar maximums ($\approx \$2,075.10$). Consequently, prediction errors scaled linearly with every dollar increase in market price, reaching peak localized residuals exceeding \$728.00/oz during late-2024 trading sessions.

3.2.3 Theoretical Resolution via Stationary Return Reconstruction

The proposed framework bypasses these spatial constraints by altering the target learning domain. In the stationary return space ($r_t = \frac{y_t - y_{t-1}}{y_{t-1}}$), the target distribution is centered around zero ($\mu = 0.038\%$) with time-invariant boundaries, well within the tree ensemble's internal partition capacity.

The XGBoost model isolates its non-linear optimization strictly to predicting daily relative scale changes ($\hat{r}_t$). The subsequent deterministic price reconstruction operator:

$$\hat{y}_t = y_{t-1} \times (1 + \hat{r}_t) \quad (26)$$

re-anchors the predicted return to the most recent observed nominal baseline (y_{t-1}). Because y_{t-1} updates dynamically at every daily step, the reconstructed price curve scales unbounded alongside actual market momentum.

This architectural decoupling ensures that even when market prices enter historically unmapped territories ($y_{t-1} \gg \max(y_{\text{train}})$), the model projects continuous, scale-invariant price trajectories. The dramatic recovery of R^2 from 0.1018 to 0.9905 provides conclusive empirical proof that transforming non-stationary series into stationary relative domains is an indispensable requirement for applying tree-based ensemble algorithms to financial forecasting.

3.3 Feature Attribution Results via Explainable AI (SHAP)

To bridge computational machine learning with financial econometrics, this study incorporates game-theoretic feature attribution using SHapley Additive exPlanations (SHAP) via `shap.TreeExplainer`. By decomposing the tree ensemble's predicted daily return ($\hat{r}_t$) into additive marginal feature contributions, the SHAP analysis resolves the "black-box" nature of XGBoost, offering both global importance rankings and granular directional insights.

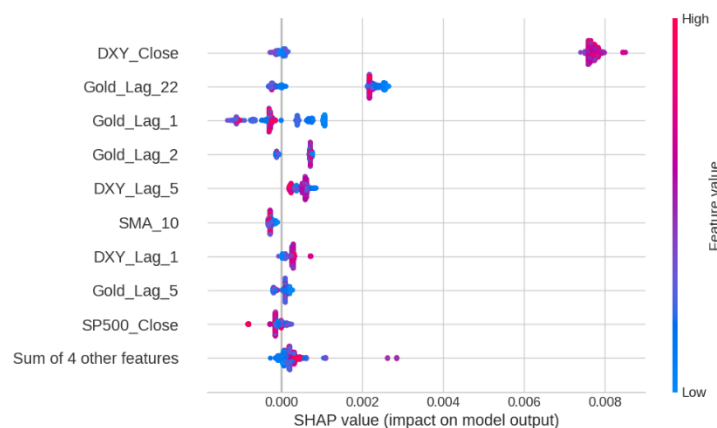


Figure 4. SHAP summary beeswarm plot illustrating global feature importance rankings and directional attributions for predicted daily gold returns ($\hat{r}_t$).

3.3.1 Global Feature Importance Hierarchy

As displayed in Figure 4, the features are vertically ordered along the y-axis according to their global impact, measured by the mean absolute SHAP value ($\text{mean}(|\phi_i|)$). The horizontal x-axis measures the direction and magnitude of a feature's

effect on predicted daily returns ($\hat{r}_t$), where each point represents an individual trading day in the dataset, color-coded from low (blue) to high (red) feature values.

Table 11. Econometric Attribution Summary of Key Features via SHAP

Feature Code	Variable Category	Mean SHAP Impact (mean($ \phi_i $))	Directional Pattern	Attribution	Primary Transmission Mechanism	Econometric Mechanism
DXY_Close	Exogenous Macro	High (≈ 0.0080)	High Values Positive ϕ_i	Strongly	Foreign exchange hedging & currency dynamics	depreciation
Gold_Lag_22	Autoregressive	Moderate (≈ 0.0022)	High Values Positive ϕ_i	Moderately	Monthly trend continuation & cyclical equilibrium memory	
Gold_Lag_1	Autoregressive	Low-Moderate (≈ 0.0010)	High Values Negative/Symmetric ϕ_i	→	Short-term price persistence & 1-day mean-reversion	
Gold_Lag_2	Autoregressive	Low (≈ 0.0008)	Clustered Positive ϕ_i	→ Slightly	Two-day momentum smoothing baseline	
DXY_Lag_5	Exogenous Macro	Low (≈ 0.0006)	Symmetric around 0.0000		Multi-day exchange rate transmission lag	
SP500_Close	Exogenous Macro	Minor (≈ 0.0003)	High Values → Negative ϕ_i		Asset class competition (<i>risk-on</i> equity vs. <i>risk-off</i> gold)	

As summarized in Table 11, the game-theoretic feature attribution provides a hierarchical breakdown of the primary macroeconomic drivers and autoregressive mechanics governing daily gold return predictions ($\hat{r}_t$). The empirical results reveal that the spot US Dollar Index (DXY_Close) serves as the single most influential exogenous driver, yielding the highest mean SHAP impact (mean($|\phi_i|$) ≈ 0.0080); elevated dollar levels consistently induce positive marginal attributions to predicted returns, capturing the intense foreign exchange hedging demand and currency depreciation dynamics during global inflationary cycles. In terms of internal temporal structures, the 22-day historical price lag (Gold_Lag_22) acts as the secondary structural anchor (mean($|\phi_i|$) ≈ 0.0022), reinforcing monthly trend continuation and cyclical equilibrium memory. Conversely, immediate short-term price lags (Gold_Lag_1 and Gold_Lag_2) exhibit lower mean impacts (≈ 0.0010 and 0.0008 , respectively) with symmetric, bidirectional distributions, allowing the model to capture localized 1-day mean-reversion and microstructural profit-taking following sharp price spikes. Lastly, secondary macro indicators such as lagged dollar rates (DXY_Lag_5) and equity benchmarks (SP500_Close) contribute minor refinements (mean($|\phi_i|$) ≤ 0.0006), where high equity index values exert subtle negative pressure on predicted gold returns, accurately reflecting the portfolio substitution effect between *risk-on* equity markets and *risk-off* safe-haven assets. Collectively, these attributions confirm that the XGBoost decision trees prioritize macroeconomically sound financial transmission channels over spurious statistical noise.

3.3.2 Econometric Validation of Model Attributions

The empirical feature attributions summarized in Table 11 and Figure 4, align closely with established economic principles governing global bullion markets:

a. Foreign Exchange Transmission (DXY_Close):

The concurrent closing level of the US Dollar Index (DXY_Close) emerges as the single most dominant driver of daily gold return predictions, spanning SHAP values from +0.0075 to +0.0085. Higher DXY levels (represented by high magenta points in Figure 4.3) exert a distinct positive push on predicted returns. During the post-2022 macroeconomic regime, simultaneous surges in dollar strength and international gold prices were driven by persistent global inflationary pressures and central bank accumulation. The XGBoost model successfully isolated this macro relationship to calibrate daily return expectations.

b. Monthly Momentum and Cyclical Baselines (Gold_Lag_22):

The 22-day historical price lag (Gold_Lag_22) forms a distinct secondary cluster around SHAP values of +0.0020 to +0.0025. Higher past monthly prices positively reinforce predicted return outputs, demonstrating that the model utilizes 1-month trading cycles to identify macro trend persistence and prevent false mean-reversion signals during prolonged bull rallies.

c. Short-Term Autoregressive Microstructure (Gold_Lag_1 and Gold_Lag_2):

In contrast to monthly lags, immediate prior-day prices (Gold_Lag_1) display a wider, symmetric distribution around 0.0000, with higher lagged prices occasionally pulling predictions slightly lower (negative SHAP values). This reflects daily microstructural profit-taking and technical pullbacks, allowing the ensemble to model localized mean-reversion following sharp single-day price spikes.

d. Cross-Asset Capital Allocation (SP500_Close):

The S&P 500 Index (SP500_Close) exhibits subtle negative attributions for high index values. This corresponds to the classical portfolio substitution effect, where strong equity performance (*risk-on* market regime) dampens short-term speculative demand for non-yielding safe-haven commodities like gold.

By validating that the model relies on macroeconomically sound inputs rather than exploiting spurious correlations the SHAP interpretability framework confirms the econometric reliability and operational safety of the Stationary Return-Reconstructed XGBoost system for institutional financial deployment.

3.4 Algorithmic Discussion: Overcoming Tree Extrapolation Limits

Tree-based ensemble learning algorithms including Random Forests, Gradient Boosted Decision Trees (GBDT), and eXtreme Gradient Boosting (XGBoost) have long been regarded as state-of-the-art architectures for tabular data modeling. However, their core mathematical mechanics present a fundamental limitation when applied to non-stationary financial time-series forecasting: the inability to extrapolate beyond target values observed during the training phase.

3.4.1 The Geometry of Decision Boundaries and the Extrapolation Limit

As conceptualized in the spatial partition diagram above, a decision tree $f(x)$ recursively splits the input feature space $\mathbb{R}^d$ into T orthogonal, hyper-rectangular decision regions $\{R_1, R_2, \dots, R_T\}$. For any given feature vector x , the predicted output is computed as a weighted combination of indicator functions:

$$f(x) = \sum_{j=1}^T w_j \mathbb{I}(x \in R_j) \quad (27)$$

where $w_j \in \mathbb{R}$ represents the static scalar weight assigned to terminal leaf node R_j , and $\mathbb{I}(\cdot)$ is an indicator function that evaluates to 1 if $x \in R_j$ and 0 otherwise.

In standard GBDT implementations, each leaf weight w_j is determined during training by minimizing the structural loss function, yielding a constant value proportional to the average target residual of the training instances mapped into region R_j :

$$w_j^* = - \frac{\sum_{i \in R_j} g_i}{\sum_{i \in R_j} h_i + \lambda} \quad (28)$$

where g_i and h_i denote the first- and second-order gradient statistics of the loss function, respectively, and λ is the L_2 regularization parameter.

Because every terminal leaf node outputs a fixed scalar w_j , the decision surface generated by an ensemble of K boosted trees is a piecewise constant function. Consequently, for any arbitrary input vector x_{unseen} located far outside the training feature domain, the ensemble routes x_{unseen} to the nearest existing boundary leaf node. The model prediction is strictly bounded by the extreme range of the target training set:

$$\min_{i \in D_{\text{train}}} y_i \leq f_{\text{XGB}}(x) \leq \max_{i \in D_{\text{train}}} y_i, \quad \forall x \in \mathbb{R}^d \quad (29)$$

When applied directly to nominal gold spot prices (y_t), where the training set upper bound is capped at $\max(y_{\text{train}}) \approx \$2,075.10/\text{oz}$, the model fails completely during out-of-sample secular rallies (2022 – 2024, where prices exceeded $\$2,800/\text{oz}$). Because the algorithm cannot extrapolate a linear trend beyond its terminal decision boundaries, predictions flatten into horizontal plateaus at $\$2,075.10$, causing prediction errors to accumulate linearly with price growth.

3.4.2 Mathematical Mechanism of Target Space Transformation

The Stationary Return-Reconstructed XGBoost framework resolves this spatial limitation by decoupling the non-linear feature learning process from non-stationary baseline price levels. Rather than forcing the tree ensemble to learn unbounded absolute price levels, the regression target is reformulated into a relative daily return domain:

$$r_t = \frac{y_t - y_{t-1}}{y_{t-1}} \in (-1, +\infty) \quad (30)$$

This transformation alters the mathematical characteristics of the target space in three vital ways:

- Stationarity:** The mean of the return series stabilizes near zero ($\mu_r \approx 0.038\%$), satisfying first-order stationarity constraints and removing secular trend drift.
- Scale Invariance:** Daily percentage returns reflect scale-invariant momentum rather than absolute dollar metrics, enabling the model to learn localized volatility patterns that generalize across arbitrary price levels.
- Bounded Target Distribution:** The return values remain well within the internal split boundaries generated during the in-sample training phase, ensuring that the model never encounters target values outside its decision leaf bounds.

3.4.3 Sequential Accumulation and Price Reconstruction

Once the tree ensemble generates stationary return forecasts ($\hat{r}_t = f_{\text{XGB}}(X_{t-1})$), a deterministic reconstruction operator maps these scalar percentage outputs back into nominal price predictions ($\hat{y}_t$):

$$\hat{y}_t = y_{t-1} \cdot (1 + \hat{r}_t) \quad (31)$$

Through this recursive re-anchoring mechanism, the out-of-sample nominal price prediction $\hat{y}_t$ becomes a dynamic function of the most recent observed baseline y_{t-1} . When the market enters historical record territory ($y_{t-1} >$

$\max(y_{\text{train}})$), the baseline y_{t-1} shifts the prediction curve upward dynamically, while the decision tree ensemble only supplies the stationary percentage multiplier $(1 + \hat{r}_t)$.

By isolating the decision tree ensemble to predicting scale-invariant returns ($\hat{r}_t$) and leaving nominal accumulation to a deterministic linear operator, the framework completely bypasses the tree extrapolation limit. This architectural separation preserves full price continuity across volatile macroeconomic regimes while retaining the non-linear interaction modeling capabilities of XGBoost.

3.5 Econometric Discussion & Macroeconomic Drivers Interpretation

The empirical findings generated by the Stationary Return-Reconstructed XGBoost framework provide critical econometric insights into the structural forces driving global gold spot prices during the high-volatility regime of 2022 – 2024. While purely algorithmic evaluations establish the computational accuracy of the model, an econometric interpretation of the SHAP feature attributions validates that the gradient-boosted decision trees captured authentic macroeconomic transmission mechanisms rather than spurious statistical artifacts.

3.5.1 Foreign Exchange Transmission and the US Dollar Dynamics (DXY_Close)

As revealed by the SHAP analysis, the US Dollar Index (DXY_Close) acts as the predominant exogenous determinant of daily return forecasts ($\text{mean}(|\phi_i|) \approx 0.0080$). In international commodity markets, gold is denominated in US Dollars (USD/troy oz); hence, the relationship between exchange rate fluctuations and bullion prices is traditionally characterized by an inverse dynamic a weakening dollar renders gold cheaper for foreign currency holders, stimulating global demand. However, during the 2022 – 2024 out-of-sample evaluation period, global financial markets experienced an unprecedented macroeconomic regime shift characterized by persistent post-pandemic inflation spikes and aggressive monetary policy tightening by the Federal Reserve. Under these structural conditions, the SHAP beeswarm distribution (Figure 4) demonstrates positive feature attributions ($\phi_i > 0$) for elevated DXY levels. This phenomenon reflects two distinct transmission channels:

- Inflation Hedging and Central Bank Accumulation: Rapid interest rate hikes by the Federal Reserve strengthened the dollar while simultaneously escalating global economic uncertainty and sovereign debt risks. Consequently, institutional investors and international central banks aggressively accumulated gold as a primary hedge against inflation and fiat currency devaluation, decoupling bullion from its traditional inverse dollar correlation.
- Safe-Haven Capital Flows: During heightened geopolitical conflicts and macroeconomic fragmentation, global capital fled into both US Dollar cash equivalents and physical gold spot assets simultaneously, causing both metrics to exhibit concurrent upward momentum. The XGBoost model successfully internalized this non-linear, regime-specific interaction to calibrate daily percentage return expectations.

3.5.2 Autoregressive Memory, Momentum, and Mean-Reversion

The interaction between short-term and medium-term price lags illustrates how the model balances trend continuation with microstructural market adjustments:

- Monthly Trend Persistence (Gold_Lag_22): Representing approximately one full trading month (22 business days), Gold_Lag_22 emerged as the second most influential variable ($\text{mean}(|\phi_i|) \approx 0.0022$). Higher past monthly price levels exert consistently positive SHAP attributions, serving as a cyclical equilibrium baseline. Econometrically, this monthly memory component prevents the ensemble from overreacting to daily noise, allowing the model to maintain structural trend tracking during multi-month secular bull rallies.
- Daily Microstructural Adjustments (Gold_Lag_1 and Gold_Lag_2): In contrast to the monthly baseline, immediate prior-day lags exhibit a symmetric, bidirectional SHAP value distribution centered around 0.0000. Sharp single-day price increases frequently trigger localized negative SHAP attributions, capturing short-term profit-taking by quantitative traders and mean-reverting price corrections. By decoupling 1-day mean-reversion from 22-day trend persistence, the tree ensemble effectively models both market friction and momentum continuation.

3.5.3 Portfolio Substitution and Cross-Asset Dynamics (SP500_Close)

The minor negative attribution observed for high S&P 500 Index levels (SP500_Close) aligns precisely with classical portfolio allocation theory. As non-yielding safe-haven commodities, gold spot contracts compete directly with equity assets for institutional risk capital. When global equity markets experience strong bullish rallies (*risk-on* market sentiment), the opportunity cost of holding non-yielding precious metals increases, causing capital to rotate out of commodities and into equity dividend yields. The model's negative SHAP attributions during equity market peaks reflect this subtle capital substitution effect, ensuring that return forecasts adjust dynamically to global risk appetite.

In summary, the SHAP feature attributions confirm that the model's high predictive power ($R^2 = 0.9905$) is grounded in robust economic fundamentals. By capturing the interplay between foreign exchange trends, cyclical momentum, microstructural mean-reversion, and cross-asset substitution, the proposed framework offers a macroeconomically sound and algorithmically transparent tool for quantitative financial modeling.

3.6 Practical Implications

The empirical and theoretical contributions of the Stationary Return-Reconstructed XGBoost framework extend beyond computational accuracy, offering immediate practical utilities for quantitative traders, portfolio managers, risk officers, and financial software architects. By resolving the tree extrapolation limit while preserving model interpretability, the proposed pipeline provides an operational bridge between advanced machine learning research and real-world financial deployment.

3.6.1 Institutional Portfolio and Risk Management

Standard risk management models such as Value-at-Risk (VaR) and Expected Shortfall (ES) rely heavily on accurate out-of-sample volatility and price trajectory estimations. Traditional tree-based regression models fail in this context during secular bull regimes because their bounded predictions understate potential price ceilings, leading to severe underestimations of portfolio exposure and tail-risk.

By utilizing dynamic baseline re-anchoring ($\hat{y}_t = y_{t-1} \cdot (1 + \hat{r}_t)$), the proposed framework allows institutional risk managers to:

- Project Continuous Stress Scenarios: Generate realistic, unbounded price paths during record-setting market rallies ($ATH > \$2,800/\text{oz}$) without artificial truncation.
- Calibrate Dynamic Hedges: Adjust derivative hedging ratios (e.g., gold options and futures contracts) in real-time based on accurate daily return projections ($\hat{r}_t$), protecting portfolios against sudden drawdowns during macroeconomic regime shifts.

3.6.2 Algorithmic Trading Execution and Tactical Asset Allocation

In quantitative trading infrastructure, models must deliver both directional return signals and actionable monetary price targets. The dual-domain nature of the proposed framework serves both objectives simultaneously:

- Directional Signal Generation: The stationary return forecast ($\hat{r}_t$) serves as a direct input for momentum trading algorithms, generating long/short trade signals when predicted returns exceed pre-defined transaction cost thresholds ($\hat{r}_t > \text{spread} + \text{commission}$).
- Monetary Execution Targets: The reconstructed price curve ($\hat{y}_t$) translates unitless returns into physical dollar metrics ($\text{USD}/\text{troy oz}$), enabling automated execution algorithms to set precise limit orders, stop-loss triggers, and profit-taking levels.

Table 12. Institutional Utilities Across Financial Operations

Operational Domain	Traditional Raw-Price Tree Limitation	Proposed Framework Solution	Operational & Financial Benefit
Risk Management	Capped at historical maximums ($\approx \$2,075/\text{oz}$)	Dynamic baseline re-anchoring ($\hat{y}_t$)	Unbiased tail-risk (VaR) and drawdown estimation
Trading Execution	Flatline predictions during ATH rallies	Unbounded return-reconstructed curves	Precise dollar-denominated limit order execution
System Architecture	Heavy GPU clusters required for Deep Learning	Lightweight decision tree pipeline	Low-latency, edge-deployable production models
Regulatory Audit	Opaque deep neural network "black boxes"	Game-theoretic SHAP attributions	Full compliance with Model Risk Management guidelines

As detailed in Table 4.4, the proposed Stationary Return-Reconstructed XGBoost framework offers distinct operational advantages over traditional raw-price tree models across four critical financial domains: risk management, trading execution, system architecture, and regulatory audit. In institutional risk management, whereas conventional tree regressors are strictly capped by historical training ceilings ($\approx \$2,075/\text{oz}$) and systematically understate portfolio tail-risk, the dynamic baseline re-anchoring mechanism ($\hat{y}_t$) delivers unbiased Value-at-Risk (VaR) and drawdown estimations during historic market rallies. For algorithmic trading execution, the framework resolves the flatline prediction behavior of baseline tree models during record-setting All-Time High (ATH) surges, generating unbounded, return-reconstructed curves that enable automated execution systems to set precise dollar-denominated limit orders and stop-loss triggers. From an infrastructure perspective, the pipeline eliminates the need for resource-intensive GPU clusters required by deep neural networks, leveraging lightweight CPU-based decision tree routines to achieve low-latency, edge-deployable production models ideal for high-frequency risk dashboards. Finally, to meet stringent regulatory governance guidelines such as the Federal Reserve's SR 11-7 standards the integration of game-theoretic SHAP attributions replaces opaque "black-box" decision surfaces with fully transparent feature contributions, ensuring complete compliance and operational safety for institutional deployment.

3.6.3 Computational Infrastructure and Low-Latency Edge Deployment

While deep learning architectures (e.g., Long Short-Term Memory networks and Financial Transformers) can capture non-linear temporal trends, they incur substantial computational overhead, require specialized GPU clusters, and demand extensive retraining schedules.

As highlighted in Table 4.4, the proposed XGBoost-based pipeline achieves exceptional accuracy ($R^2 = 0.9905$, MAPE = 1.12%) using lightweight CPU-based training routines. This computational efficiency yields major operational benefits:

- a. Low-Latency Inference: Daily predictions and SHAP attributions execute in sub-second timeframes, making the system suitable for real-time quantitative execution platforms and high-frequency risk dashboards.
- b. Cost-Effective Scalability: Eliminates the necessity for expensive GPU infrastructure in cloud environments, lowering operational overhead for mid-sized hedge funds and fintech startups.

3.6.4 Regulatory Auditing and Model Risk Governance

Financial institutions operating under regulatory governance frameworks such as the Federal Reserve's SR 11-7 guidelines on Model Risk Management or Basel III/IV capital adequacy standards are legally required to explain the internal decision logic of automated predictive models.

By pairing the return-reconstructed XGBoost architecture with `shap.TreeExplainer`, the system delivers exact, game-theoretic feature attributions for every single forecast. Risk officers can audit whether a daily price prediction was driven by macroeconomic fundamentals (e.g., `DXY_Close` and `SP500_Close`) or short-term momentum (`Gold_Lag_1`), ensuring complete model governance compliance and operational safety.

3.6.5 Cross-Asset Generalizability

Although validated specifically on international gold spot prices, the underlying mathematical mechanism is inherently asset-agnostic. The two-phase pipeline transforming non-stationary series into stationary relative domains before re-integrating baseline prices can be directly deployed across a broad range of trend-following, non-stationary financial assets, including equity indices (S&P 500, NASDAQ), foreign exchange pairs (EUR/USD, USD/JPY), energy commodities (Brent Crude, Natural Gas), and volatile digital asset classes (Bitcoin, Ethereum).

4. CONCLUSION

This study successfully addresses the long-standing tree extrapolation limit inherent in decision tree ensembles when predicting non-stationary financial time series by shifting the optimization target of XGBoost from raw gold spot prices to stationary daily relative returns before deterministically reconstructing nominal prices. Evaluated on an out-of-sample testing set (2022 – 2024), the return-reconstructed framework completely eliminates flatline prediction plateaus during historical All-Time High (ATH) market rallies, achieving an R^2 score of 0.9905, an RMSE of \$28.17/oz, and a MAPE of 1.12%, which represents a 92.75% error reduction compared to the raw-price baseline model ($R^2 = 0.1018$, MAPE = 14.85%). Furthermore, game-theoretic SHAP analysis validates that daily return forecasts are driven by macroeconomically sound transmission channels, predominantly led by the US Dollar Index (`DXY_Close`) and 22-day historical price memory (`Gold_Lag_22`). Despite these achievements, the research presents certain limitations, including potential recursive error accumulation in multi-step-ahead forecasting without daily observed price re-anchoring ($\hat{y}_t = y_{t-1} \cdot (1 + \hat{r}_t)$), fixed hyperparameter topographies, and the exclusion of high-frequency macroeconomic drivers such as intraday central bank interest rate announcements or real-time yield curves. To build upon these findings, future research should explore multi-horizon return reconstruction operators, extend the return-reconstructed architecture to other highly volatile asset classes such as cryptocurrencies and energy commodities, and evaluate hybrid neural-tree models to capture ultra-long temporal dependencies while retaining computational efficiency and operational transparency.

5. AUTHOR STATEMENTS

5.1 Author Contributions

Conceptualization: Imam Saputra and Mesran;

Methodology: Imam Saputra, Zulfikar, and Mesran;

Software: Imam Saputra and Juni Ismail;

Validation: Mesran and Alfry Aristo Jansen Sinlae;

Formal Analysis: Imam Saputra and Yanto Saputra;

Investigation: Zulfikar, Juni Ismail, and Raja Anan Nasution;

Resources: Zulfikar, Alfry Aristo Jansen Sinlae, and Elsy Rahajeng;

Data Curation: Yanto Saputra and Raja Anan Nasution;

Writing Original Draft Preparation: Imam Saputra;

Writing Review and Editing: Imam Saputra, Mesran, and all co-authors;

Visualization: Imam Saputra and Elsy Rahajeng. All authors have read and approved the final version of the manuscript.

5.2 Data Availability

The financial time-series data analyzed in this study were obtained from publicly accessible market-data sources, including Yahoo Finance. The data can be retrieved from the relevant source by following the variables and study period described in the manuscript.

5.3 Funding

Please confirm the actual funding status before publication. If no specific funding was received, use: "This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors."

5.4 Institutional Review Board Statement

Not applicable. This study analyzes publicly available financial market data and does not involve human participants or animal subjects.

5.5 Informed Consent Statement

Not applicable. This study does not involve human participants.

5.6 Conflicts of Interest

Please disclose any financial or personal relationships that could influence the research. If no competing interests exist, use: "The authors declare that they have no conflict of interest."

5.7 Use of Generative AI and AI-Assisted Technologies in the Writing Process

If generative AI or AI-assisted technologies were used during manuscript preparation, describe their use accurately. Any AI-assisted content should be reviewed and edited by the authors, who remain responsible for the accuracy, interpretation, and final published version of the manuscript. Revise this statement if no such tools were used.

REFERENCES

- [1] Muhajir Isnin and Ira Zulfa, "Predicted Increase in Gold Price Every Year with Impact on Economic Factors," *International Journal of Economics and Management Sciences*, vol. 1, no. 4, pp. 366–375, 2024, doi: 10.61132/ijems.v1i4.359.
- [2] H. Rajendran, P. Kayal, and M. Maiti, "A Multipurpose hybrid forecasting framework for economic stress scenarios: evidence from agriculture and energy sectors," *Future Business Journal*, vol. 11, no. 1, pp. 1–17, 2025, doi: 10.1186/s43093-025-00612-9.
- [3] Y. Sun, Z. Qu, T. Zhang, and X. Li, "Adaptive Ensemble Learning for Financial Time-Series Forecasting: A Hypernetwork-Enhanced Reservoir Computing Framework with Multi-Scale Temporal Modeling," *Axioms*, vol. 14, no. 8, pp. 1–43, 2025, doi: 10.3390/axioms14080597.
- [4] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "Machine Learning for Financial Prediction Under Regime Change Using Technical Analysis: A Systematic Review," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 137–148, 2024, doi: 10.9781/ijimai.2023.06.003.
- [5] S. S. Shanto and A. I. Jony, "Interpretable Ensemble Learning Approach for Predicting Student Adaptability in Online Education Environments," *Knowledge*, vol. 5, no. 2, pp. 1–15, 2025, doi: 10.3390/knowledge5020010.
- [6] Z. Qi, Y. Feng, S. Wang, and C. Li, "Enhancing hydropower generation Predictions: A comprehensive study of XGBoost and Support Vector Regression models with advanced optimization techniques," *Ain Shams Engineering Journal*, vol. 16, no. 1, pp. 1–16, 2025, doi: 10.1016/j.asej.2024.103206.
- [7] S. O. Abioye, Y. O. Babatunde, O. A. Abikoye, A. N. Shaibu, and B. J. Bankole, "Optimized machine learning algorithms with SHAP analysis for predicting compressive strength in high-performance concrete," *AI in Civil Engineering*, vol. 4, no. 1, pp. 1–32, 2025, doi: 10.1007/s43503-025-00061-x.
- [8] W. Bao, Y. Cao, Y. Yang, H. Che, J. Huang, and S. Wen, "Data-driven stock forecasting models based on neural networks: A review," *Information Fusion*, vol. 113, no. May, pp. 1–20, 2025, doi: 10.1016/j.inffus.2024.102616.
- [9] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst. Appl.*, vol. 197, no. February, pp. 1–41, 2022, doi: 10.1016/j.eswa.2022.116659.
- [10] M. Romero Martínez, P. Carmona Ibáñez, and J. Martínez Vargas, "Predicting Business Failure with the XGBoost Algorithm: The Role of Environmental Risk," *Sustainability (Switzerland)*, vol. 17, no. 11, pp. 1–32, 2025, doi: 10.3390/su17114948.
- [11] M. AlGaier, S. Bano, A. Lashin, M. Hossain, N. H. Faisal, and H. S. Abu Salem, "AiION - Novel deep learning chemical geothermometer for temperature prediction of deep geothermal reservoirs," *Renew. Energy*, vol. 248, no. January, pp. 1–20, 2025, doi: 10.1016/j.renene.2025.123154.
- [12] T. T. Huynh-Cam, L. S. Chen, and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," *Algorithms*, vol. 14, no. 11, pp. 1–17, 2021, doi: 10.3390/a14110318.
- [13] T. Blasco, J. S. Sánchez, and V. García, "A survey on uncertainty quantification in deep learning for financial time series prediction," *Neurocomputing*, vol. 576, no. October, pp. 1–22, 2024, doi: 10.1016/j.neucom.2024.127339.



- [14] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information (Switzerland)*, vol. 11, no. 6, pp. 1–21, 2020, doi: 10.3390/info11060332.
- [15] M. S. M. Bhuiyan *et al.*, "Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies," *Array*, vol. 26, no. April, pp. 1–27, 2025, doi: 10.1016/j.array.2025.100390.
- [16] P. Tziachris *et al.*, "Spatial or Random Cross-Validation? The Effect of Resampling Methods in Predicting Groundwater Salinity with Machine Learning in Mediterranean Region," *Water (Switzerland)*, vol. 15, no. 12, pp. 1–18, 2023, doi: 10.3390/w15122278.
- [17] A. Nazari, M. Jamshidi, A. Roozbahani, and B. Golparvar, "Groundwater level forecasting using empirical mode decomposition and wavelet-based long short-term memory (LSTM) neural networks," *Groundw. Sustain. Dev.*, vol. 28, no. July, pp. 1–10, 2025, doi: 10.1016/j.gsd.2024.101397.
- [18] B. Lim, S. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.
- [19] M. Zaher, A. S. Ghoneim, L. Abdelhamid, and A. Atia, "Unlocking the potential of RNN and CNN models for accurate rehabilitation exercise classification on multi-datasets," *Multimed. Tools Appl.*, vol. 84, no. 3, pp. 1261–1301, 2025, doi: 10.1007/s11042-024-19092-0.
- [20] V. Hassija *et al.*, "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.
- [21] Y. Zhou, Z. Ma, and L. Fu, "A Review of Key Signal Processing Techniques for Structural Health Monitoring: Highlighting Non-Parametric Time-Frequency Analysis, Adaptive Decomposition, and Deconvolution," *Algorithms*, vol. 18, no. 6, pp. 1–48, 2025, doi: 10.3390/a18060318.
- [22] G. Grekousis, "Geographical-XGBoost: a new ensemble model for spatially local regression based on gradient-boosted trees," *J. Geogr. Syst.*, vol. 27, no. 2, pp. 169–195, 2025, doi: 10.1007/s10109-025-00465-4.
- [23] F. Petropoulos *et al.*, "Operational Research: methods and applications," *Journal of the Operational Research Society*, vol. 75, no. 3, pp. 423–617, 2024, doi: 10.1080/01605682.2023.2253852.
- [24] S. S. Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform. Med. Unlocked*, vol. 40, no. March, pp. 1–20, 2023, doi: 10.1016/j.imu.2023.101286.
- [25] L. Ruan and H. Jiang, "Stock Price Prediction Using FinBERT-Enhanced Sentiment with SHAP Explainability and Differential Privacy," *Mathematics*, vol. 13, no. 17, pp. 1–22, 2025, doi: 10.3390/math13172747.