Modelling the Price of Gold

Using GARCH, machine learning and deep learning models

Tom O'Connor

Antonin Meudic

Adrien Gibert

Thomas Goumont

L'Ecole Nationale de la Statistique et de l'Analyse de l'Information
Risk Management / Advanced statistical engineering Specialisation
Supervisor: Youssef ESSTAFA

2025-2026

This page intentionally left blank.

ABSTRACT

This study investigates the stochastic dynamics and predictive modeling of gold prices using a comprehensive dataset of the SPDR Gold Shares ETF (GLD) spanning two decades (2004–2025). This period encompasses critical structural shifts, from the post-2008 bull market to the inflation-driven surge of 2024. Preliminary diagnostics confirm that gold returns adhere to canonical stylized facts of financial time series: while the conditional mean approximates a white noise process, rendering standard linear forecasting (ARMA) ineffective, the series exhibits pronounced volatility clustering, heavy tails, and regime-dependent behavior. Consequently, we demonstrate that persistence is concentrated in the conditional variance, efficiently captured by a GARCH(1,1) specification with Student-t innovations, which outperforms long-memory alternatives (FIGARCH).

Beyond traditional econometrics, this research evaluates the efficacy of hybrid machine learning frameworks in extracting signal from noise. We propose two advanced architectures: a hybrid ARIMA-XGBoost model, designed to capture non-linear residual dynamics, and a deep learning LSTM-GARCH framework. Our results indicate that while pointwise return predictability remains low, hybrid models significantly enhance directional accuracy and trend reconstruction. Furthermore, when deployed in a volatility-targeted portfolio strategy, the LSTM-GARCH approach delivers superior risk-adjusted returns and reduced drawdowns compared to passive benchmarks. These findings suggest that integrating deep learning with rigorous econometric risk modeling offers a robust edge for asset allocation in highly efficient markets.

CONTENTS

List of Figures iv

List of Tables vi

1 Data Preprocessing and Analysis **2**

1.1 The dataset 2

1.1.1 Data source 2

1.1.2 How to model gold 3

1.1.3 Return definition 4

1.2 Cyclical and Regime Nature of Gold Prices 5

1.2.1 Macroeconomic Regimes and Secular Cycles in Gold Prices 5

1.3 Seasonality Analysis 6

1.3.1 Monthly Aggregation 6

1.4 Outlier Detection 8

1.4.1 Robust MAD-based z-scores 8

1.4.2 Outlier results 8

1.5 Temporal Dependence and Volatility Clustering 9

1.5.1 ACF and PACF 9

1.5.2 Ljung–Box test 10

1.5.3 Augmented Dickey–Fuller test 10

1.5.4 Mean Predictability 11

1.6 Long–Memory Diagnostics and FARIMA Assessment 11

1.6.1 FARIMA Estimation on Returns 11

1.6.2 Long Memory in Volatility Proxies 11

1.6.3 Implications for Modeling : choosing the correct GARCH 12

2 LSTM model **13**

2.1 LSTM–GARCH model 13

2.1.1 Data and Target Construction 13

2.1.2 Train/Test Split and Strict Out-of-Sample Design 13

2.1.3 Risk-Targeted Long/Short Strategy 14

2.1.4 Benchmark Strategies and Evaluation 15

3 Gradient Boosting **18**

3.1 Gradient Boosting method for nonlinear modeling of gold returns 18

3.2 XGBoost model 19

3.2.1 Prediction target and forecast horizon 19

3.2.2 Modeling Framework and Walk-Forward Prediction 19

3.3 ARIMA-XGBoost 21

4 Portfolio Management **25**

Bibliography 29

Appendices

A Appendix 33
A.1 Seasonality . . . 33
A.2 Formal calendar-month seasonality test (ANOVA on monthly returns) . . . 33
A.3 Performance metrics . . . 34
A.4 Features used in the XGBoost models and feature importance . . . 34

B Hyperparameter tuning 36

C How the LSTM trading strategy works? 38
C.1 Trading Strategy LSTM . . . 38

LIST OF FIGURES

1.1 Daily gold proxy price $P_t = 10 \times$ GLD Close (Yahoo Finance), 2005–2025. 2

1.2 Daily log-returns r_t (percent). 4

1.3 Histogram of daily returns. Heavy tails are visible via frequent large-magnitude observations. 5

1.4 QQ-plot of daily returns. Deviations in the tails indicate non-Gaussian behavior (fat tails). 5

1.5 Trend visualization: daily price with MA(50) and MA(200). 6

1.6 Seasonal plot of monthly gold prices. 7

1.7 Additive decomposition of monthly gold prices. 7

1.8 Daily returns with MAD-based outliers highlighted ($|z_t| > 4$). 9

1.9 ACF of gold returns. No significant autocorrelation is observed, indicating weak linear dependence in the conditional mean. 9

1.10 PACF of gold returns. The absence of significant partial autocorrelation confirms that returns behave approximately as white noise. 9

1.11 ACF of squared returns. Persistent positive autocorrelation indicates volatility clustering (ARCH/GARCH effects). 10

1.12 PACF of squared returns. Significant partial autocorrelation confirms strong volatility persistence. 10

2.1 Normalized equity curves comparing the LSTM–GARCH trading strategy with a buy-and-hold benchmark over the out-of-sample period. The active strategy exhibits smoother growth and reduced drawdowns relative to the passive benchmark. 15

2.2 Next-day log returns: comparison between LSTM-predicted returns and realized returns. Predicted returns are smoother and primarily capture short-term directional signals rather than exact magnitudes. 15

2.3 One-step-ahead conditional volatility forecasts obtained from the GARCH model. Volatility clustering and gradual mean reversion are clearly visible. 15

2.4 Executed portfolio weights over time after combining return forecasts with volatility targeting. Position sizes adjust dynamically across market regimes, including periods of short exposure. 16

2.5 Daily transaction costs incurred by the trading strategy. Spikes correspond to major portfolio rebalancing events, while average costs remain limited. 16

2.6 Normalized equity curves of the LSTM–GARCH strategy and a buy-and-hold benchmark over the period 2021–2025. 16

2.7 One-step-ahead conditional volatility forecasts $\hat{\sigma}_{t+1}$ from the GARCH model. The series exhibits pronounced volatility clustering, with sharp spikes during periods of market stress, particularly toward the end of the sample. 17

2.8 Next-day log-return forecasts compared with realized returns. While predicted returns remain close to zero, realized returns display large fluctuations, highlighting the difficulty of short-horizon return prediction despite time-varying volatility. 17

3.1 Cumulative log-return predictions at horizon $h = 10$ obtained through the XGBoost procedure . . . 20
3.2 Out-of-Sample Gold Price Trajectory (USD): Actual 2025 Bull Market vs. XGBoost Conservative Trend Forecast . . . 20
3.3 Comparison of log returns predictions ARIMA vs Hybrid . . . 22
3.4 Comparison of Gold Price predictions ARIMA vs Hybrid . . . 22
3.5 Comparison of ARIMA vs Hybrid from 2021 . . . 23
3.6 Feature Importance Ranking (Short-term Test: 2025) . . . 24
3.7 Feature Importance Ranking (Long-term Test: 2021-2026) . . . 24

4.1 Backtest Walk-Forward & Benchmarks . . . 26
4.2 . . . 26

A.1 Shap plot of the XGBoost model . . . 35
A.2 Covariance matrix between features and target . . . 35

B.1 Heatmap of test RMSE (in basis points) as a function of the learning rate and look-back window size. Larger window sizes consistently reduce prediction error, while excessively large learning rates degrade performance. . . . 36
B.2 Average test RMSE as a function of the look-back window size. Prediction accuracy improves monotonically with longer sequences, suggesting that long-term temporal dependencies are relevant for return dynamics. . . . 36
B.3 Mean test RMSE as a function of the learning rate (log scale). Lower learning rates yield more stable convergence and improved generalization, while overly aggressive learning rates increase forecasting error. . . . 36
B.4 Mean test RMSE as a function of the number of hidden units. Increasing model capacity initially improves performance, but excessive width leads to mild overfitting. . . . 37
B.5 Heatmap of test RMSE as a function of window size and number of hidden units. The best performance is achieved by combining long look-back windows with moderate network width. . . . 37

LIST OF TABLES

1.1 Price level summary statistics 4
1.2 Daily return summary statistics 4
1.3 Comparison of GARCH-family volatility models 12

2.1 Five-year (2021-2025) average performance of the LSTM+GARCH strategy across rolling test windows. 16

3.1 Performance Comparison: Static Model vs. Walk-Forward vs. Naive Benchmark (Horizon $h = 10$) 20
3.2 Performance comparison of the Hybrid Model (ARIMA + XGBoost) across different testing periods. 23

4.1 Performance Summary (2021-2025). 26

A.1 Features used in the XGBoost model 34

INTRODUCTION

Gold occupies a unique position in the global economy. Unlike traditional financial assets whose value is derived from future cash flows, gold is a ”debt-free” currency and a millennia-old store of value. In the contemporary context, marked by increasing geopolitical instability, fluctuating monetary policies, and unpredictable inflation cycles, understanding the drivers of gold prices has become an imperative for institutional investors and risk managers alike.

This report focuses on the analysis of the SPDR Gold Shares (GLD), the largest exchange-traded fund backed by physical gold bullion, providing a representative sample of market liquidity and volatility over the last two decades from 2004 to 2025. This period is of particular interest as it encompasses several major structural regime shifts: the post-2008 bull market, the 2013 price collapse driven by U.S. monetary policy normalization, and the recent inflation-driven surge of 2024–2025.

From a statistical perspective, forecasting daily gold price movements constitutes a real challenge. The asset is characterized by a notoriously low signal-to-noise ratio, where daily returns often approximate a random walk behavior that defies classical linear forecasting. We are acutely aware of the theoretical difficulties in predicting such a highly liquid and efficient market, where price changes are predominantly driven by stochastic news flow rather than deterministic patterns. However, rather than viewing this inherent unpredictability as a deterrent, we approach it as a rigorous testing ground. Recognizing that the series exhibits complex ”stylized facts”, such as volatility clustering and non-Gaussian distributions, this study deliberately embraces the challenge to investigate whether modern analytical frameworks can succeed where traditional intuition suggests no edge can be found.

To address these complexities, this research adopts a progressive three-step methodology:

- **Data preprocessing and analysis:** A rigorous study of seasonality, stationarity, and return distributions to identify the stylized facts of the series, confirming that gold does not exhibit exploitable calendar seasonality.
- **Classical modeling:** The application of ARMA(0,0) specifications for the conditional mean and GARCH(1,1) models with Student-t innovations to capture the dynamics of conditional variance and heavy-tailed innovations.
- **Advanced predictive approaches:** A comparison between deep learning, via Long Short-Term Memory (LSTM) networks capable of capturing long-term dependencies, and a hybrid ARIMA-XGBoost framework designed to isolate non-linear signals from statistical noise.

While grounded in established research on gold price prediction (Thilak et al., 2025), this report embraces the specific challenge of modeling returns, a rigorous approach designed to isolate genuine market dynamics from non-stationary noise. The ultimate goal of this study is to determine whether the integration of machine learning techniques can outperform traditional econometric models in forecasting an asset as sensitive to macroeconomic shocks as gold.

1

DATA PREPROCESSING AND ANALYSIS

1.1 The dataset

1.1.1 Data source

We retrieve the daily time series for the ticker `GLD` from Yahoo Finance using the `quantmod::getSymbols` function. The closing price series, extracted via `Cl()`, spans the period from 2004-11-18 to 2025-12-31, yielding approximately twenty years of data. The final sample consists approximately of 5,300 daily observations.
The SPDR Gold Shares ETF (GLD) is an exchange-traded fund designed to closely track the spot price of physical gold bullion, as defined by the LBMA Gold Price PM (London afternoon gold fixing). The fund is physically backed by allocated gold bars stored in London, and each share represents approximately one-tenth of a troy ounce of gold. Consequently, fluctuations in GLD prices closely follow movements in the international spot gold market, aside from minor deviations arising from management fees, tracking error, and market microstructure effects. GLD is therefore widely used as a proxy for spot gold in empirical financial studies, including Dao et al. (2017) and Hood et al. (2013).

Let P_t^{GLD} denote the GLD daily closing price, since it's one-tenth the price of gold, we define a scaled proxy series as :

$$P_t = 10 \times P_t^{\text{GLD}}, \tag{1.1}$$

Here's a visualisation of the gold price :

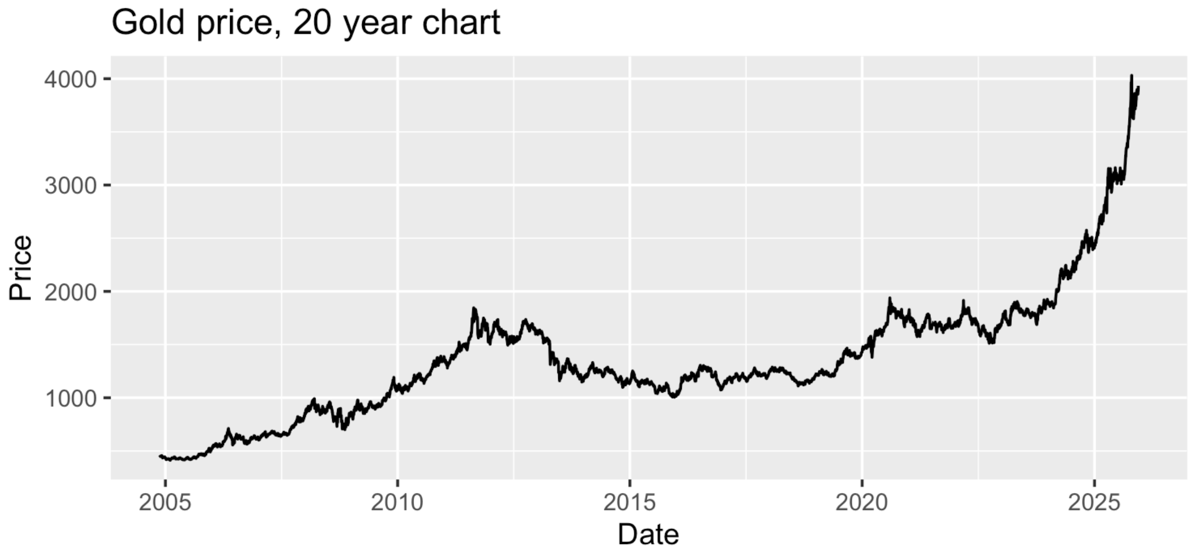


Figure 1.1: *Daily gold proxy price $P_t = 10 \times$ GLD Close (Yahoo Finance), 2005–2025.*

1.1.2 How to model gold

The chart above 1.1 shows that the gold price follows a long-run upward trend, with a few notable temporary corrections. Gold prices collapsed in 2013 and remained weak through 2014 mainly because the macro-financial environment that had supported the long post-2008 gold bull market abruptly reversed. A key driver was the **shift in investor expectations, especially about U.S. monetary policy**, as the Federal Reserve began signalling in 2013 that it would taper its asset-purchase programs and move gradually toward monetary policy normalization. This change in expectations translated into **stronger U.S. dollar and rising real interest rates**, which sharply increased the opportunity cost of holding gold (an asset that generates no yield). While simultaneously reducing international demand by making gold more expensive for non-U.S. investors. At the same time, fears of systemic financial collapse, deflation, and runaway inflation faded as the global economy stabilized, leading investors to rotate out of safe-haven assets such as gold and into equities and other risk assets delivering higher returns. This fundamental shift in sentiment triggered massive liquidations of gold ETFs and speculative futures positions, particularly during the April 2013 crash, amplifying price declines through technical selling and stop-loss cascades. Weakening physical demand from key consuming countries, notably China and India, where import restrictions and higher tariffs were introduced, further reduced price support. As these forces persisted into 2014, with inflation expectations remaining subdued and monetary tightening gradually approaching, gold failed to recover, marking a structural regime change after a decade-long bull market driven by crisis-era monetary expansion.

Gold prices surged again in 2024–2025 mainly due to **heightened global uncertainty and a renewed decline in real interest rates**. Persistent geopolitical tensions and global economic uncertainty strengthened gold's role as a safe-haven asset, while expectations of monetary easing by major central banks led to lower real interest rates, reducing the opportunity cost of holding a non-yielding asset such as gold. At the same time, strong central-bank purchases and rising ETF inflows significantly increased investment demand, as many institutions sought to diversify reserves and hedge against currency and financial risks, pushing gold to new historical highs.

So, we can model the gold price by :

$$P_t^{\text{Gold}} = f\left(-r_t^{\text{real}}, -\text{USD}_t, +\mathcal{L}_t, +\mathcal{R}_t, +\pi_t^e\right)$$

where:

- P_t^{Gold} denotes the price of gold at time t.
- r_t^{real} is the real interest rate at time t, proxied by the yield on inflation-indexed government bonds (e.g. U.S. 10-year TIPS yield).
- USD_t denotes the strength of the U.S. dollar at time t, typically measured by the Dollar Index (DXY).
- $\mathcal{L}_t$ represents global liquidity conditions at time t, proxied by broad money aggregates (e.g. global M2), central-bank balance sheet size, or monetary policy stance indicators.
- $\mathcal{R}_t$ denotes financial and geopolitical risk at time t, proxied by market volatility and stress indicators such as the VIX, credit spreads, or financial stress indices.
- π_t^e denotes expected inflation at time t, proxied by breakeven inflation rates or inflation-swap measures.

1.1.3 Return definition

Instead of simple returns, we work with log-returns defined by

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t denotes the asset price at time t. Log-returns are preferred because they are time-additive, i.e.

$$\sum_{t=1}^{T} r_t = \ln\left(\frac{P_T}{P_0}\right),$$

which greatly simplifies multi-period aggregation and econometric modeling. For small price variations, log-returns closely approximate simple percentage returns, ensuring interpretability while retaining superior theoretical properties. Below is a chart of the log-returns :

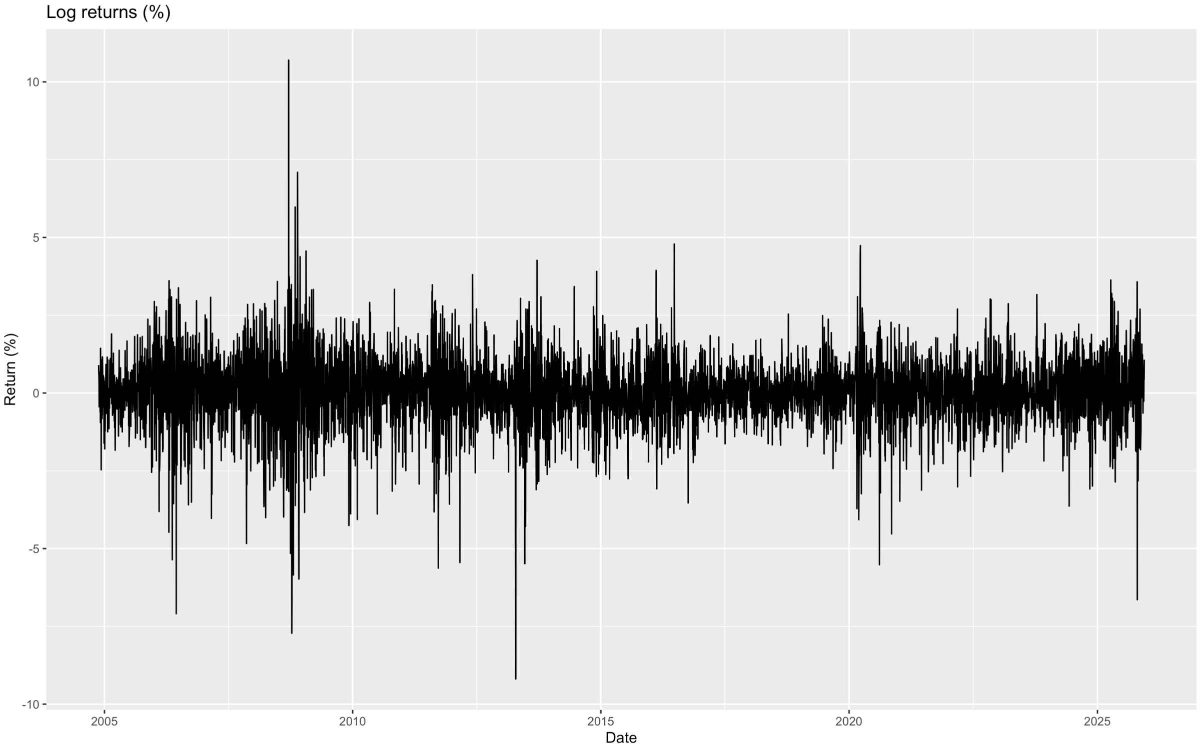


Figure 1.2: *Daily log-returns r_t (percent).*

The gold proxy price exhibits the following summary statistics:

Table 1.1: *Price level summary statistics*

Statistic	Value
Minimum	412.6
1st Quartile	1061.2
Median	1248.8
Mean	1355.5
3rd Quartile	1662.6
Maximum	4031.5
Std. deviation	583.1

Table 1.2: *Daily return summary statistics*

Statistic	Value (%)
Minimum	-9.1905
1st Quartile	-0.5115
Median	0.0576
Mean	0.0412
3rd Quartile	0.6213
Maximum	10.6974
Std. deviation	1.1121

Higher-order moment diagnostics of the returns; notably Skewness ≈ -0.327 (mild negative skew), large negative returns are more extreme than large positive ones, so downside risk dominates upside extremes, big crashes are sharper than big one-day increases in the price. The Kurtosis is ≈ 8.865 so the distribution is fat-tailed, meaning extreme moves happen far more often than a standard gaussian distribution, motivating a us to look at an outlier detection and non-normal error distributions in volatilty modeling.

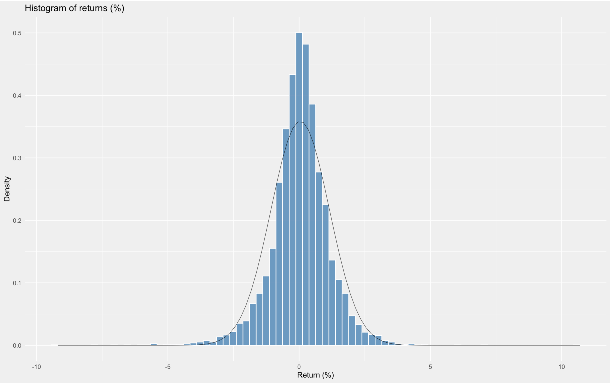


Figure 1.3: *Histogram of daily returns. Heavy tails are visible via frequent large-magnitude observations.*

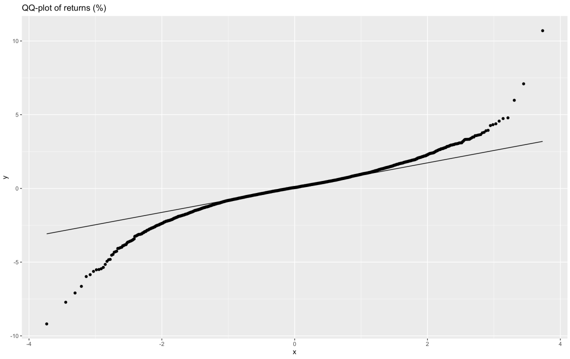


Figure 1.4: *QQ-plot of daily returns. Deviations in the tails indicate non-Gaussian behavior (fat tails).*

The Kolmogorov-Smirnov test has a p-value of 2.2e-16 and the Anderson-Darling normality test has the same p-value so the distribution is statistically not normal.

1.2 Cyclical and Regime Nature of Gold Prices

1.2.1 Macroeconomic Regimes and Secular Cycles in Gold Prices

Gold prices are strongly influenced by exogenous macroeconomic and geopolitical factors. Rather than evolving as a stationary process with constant statistical properties, the gold market exhibits distinct cyclical regimes characterised by persistent trends and regime-dependent dynamics.

In particular, a prolonged secular bull market can be identified over the period 2005–2012, following the burst of the dot-com bubble and the subsequent accommodative monetary policies implemented by major central banks. During this phase, declining real interest rates and rising inflation expectations supported a sustained upward trajectory in gold prices.

From 2012 onward, expectations of monetary tightening by the Federal Reserve induced a regime shift, initiating a prolonged bear cycle. This downward phase culminated in a trough around 2016. The subsequent increase in geopolitical uncertainty following the 2016 United States presidential election marked the beginning of a new bull regime, characterised by renewed upward momentum in gold prices.

These regime transitions can generate long-memory effects, volatility clustering, and asymmetric tail risk in returns, thereby violating the assumptions of independent and identically distributed Gaussian innovations. Consequently, models that explicitly account for regime persistence and long-memory volatility dynamics, such as LSTM-based predictors coupled with GARCH volatility processes, are particularly well suited to capture the structural behaviour of gold returns.

We can visualise rolling moving averages in order to highlight persistent trends in gold prices. Formally, let P_t denote the daily gold price at time t. The k-day moving average is defined as:

$$\mathrm{MA}_k(t) = \frac{1}{k}\sum_{i=0}^{k-1} P_{t-i}$$

These rolling averages smooth short-term fluctuations and reveal medium- and long-run price dynamics, thereby facilitating the identification of persistent trends and potential regime turning points, the 200 day moving average captures the long-term tendency.

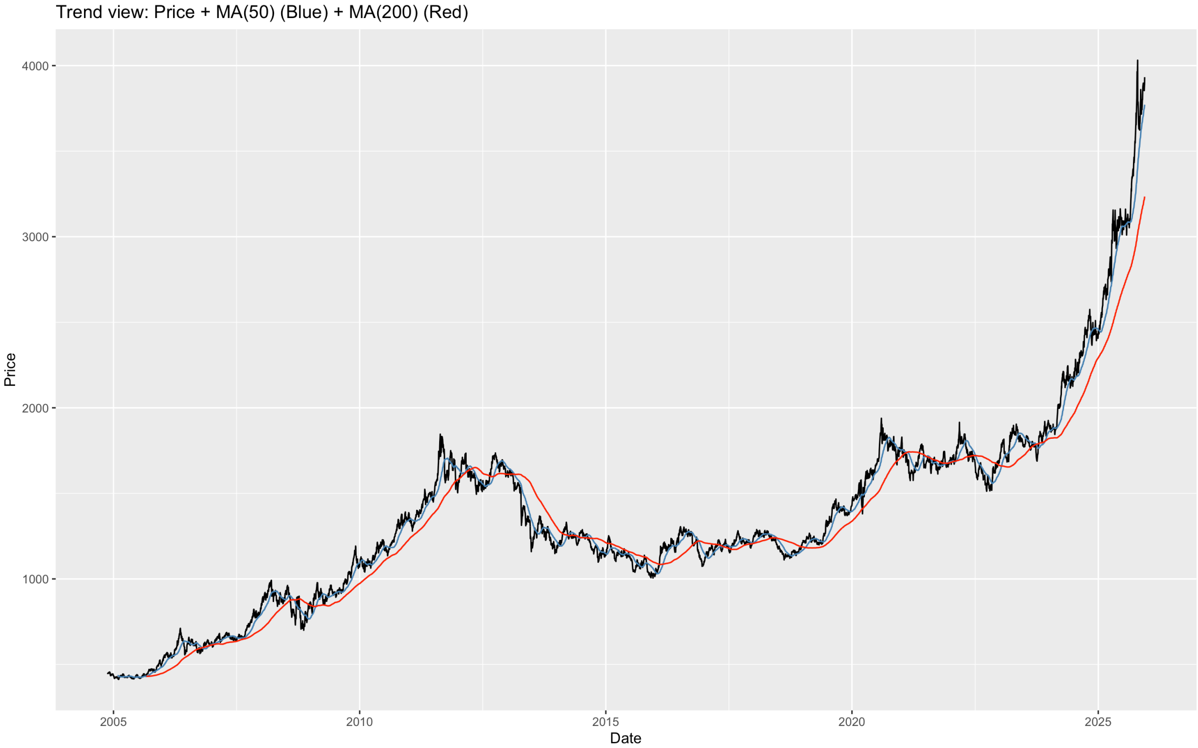


Figure 1.5: *Trend visualization: daily price with $MA(50)$ and $MA(200)$.*

We conclude that gold exhibits a pronounced cyclical regime structure. Based on the observed long-run dynamics, the following major regimes can be identified:

- **2004–2011: Bull market regime** — characterised by declining real interest rates and accommodative global monetary conditions.
- **2012–2015: Bear market regime** — associated with expectations of monetary tightening and a reversal in inflation expectations.
- **2016–2019: Stagnation regime** — marked by range-bound price dynamics and subdued macroeconomic momentum.
- **2020–2025: Inflation-driven regime** — characterised by elevated inflation expectations, expansionary fiscal policies, and heightened geopolitical uncertainty.

1.3 Seasonality Analysis

1.3.1 Monthly Aggregation

Our previous analysis reveals that gold prices exhibit pronounced cyclical behaviour driven by macroeconomic regimes. A natural question is whether, in addition to these long-run cycles, gold also displays systematic calendar seasonality at the monthly frequency.

To address this question, we first aggregate daily spot prices to an end-of-month monthly series and construct corresponding log-returns. Figure 1.6 presents a seasonal plot of monthly prices across years, allowing for a visual inspection of recurring calendar patterns.

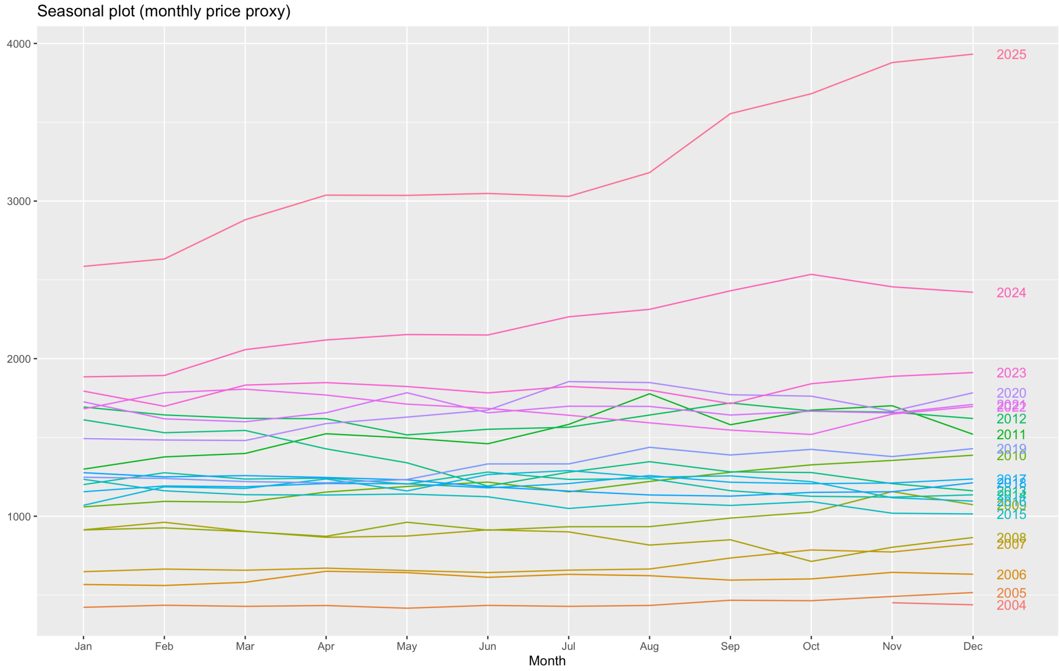


Figure 1.6: *Seasonal plot of monthly gold prices.*

The seasonal overlay does not reveal any stable repeating month-of-year pattern. The dispersion across years largely dominates any potential calendar regularity, suggesting that gold prices are not characterised by persistent monthly seasonality.

To further investigate this, Figures 1.7 report the classical additive decomposition of the monthly series.

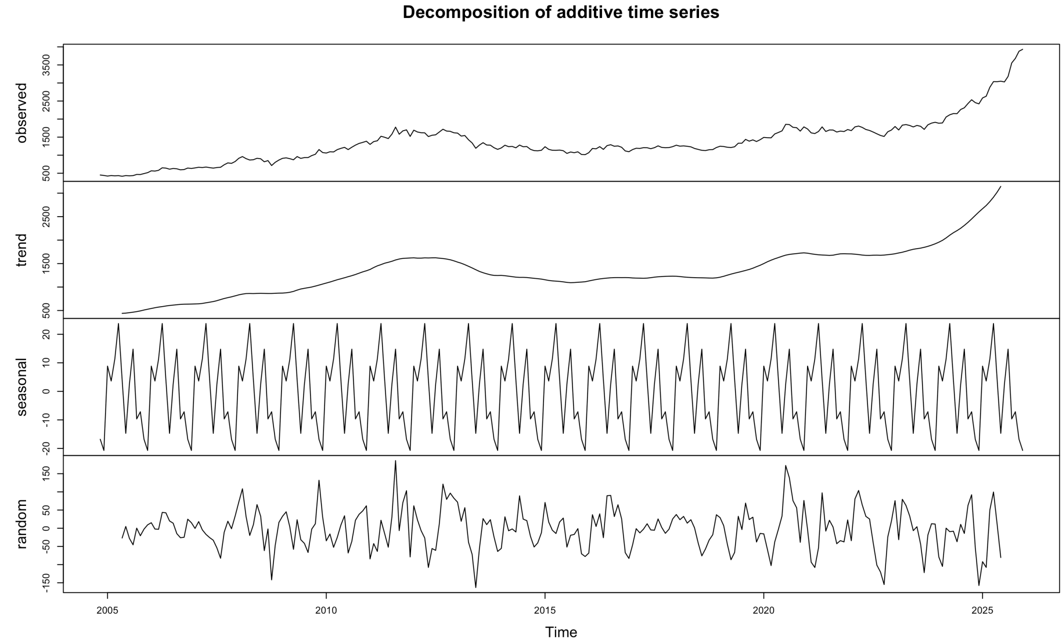


Figure 1.7: *Additive decomposition of monthly gold prices.*

In the decomposition, the estimated seasonal component is economically negligible (range bounded by 20 dollars) when compared to the scale of the trend component. Seasonal fluctuations remain small relative to the overall price level, indicating that potential calendar effects play a minor role in the long-run dynamics of gold prices.

Finally, we formally test for monthly seasonality in returns by applying an analysis of variance (ANOVA)[1] to monthly log-returns grouped by calendar month. The null hypothesis of equal mean returns across months cannot be rejected (p-value = 0.59), providing no statistical evidence of monthly seasonality.

Overall, these results indicate that gold does not exhibit exploitable calendar seasonality. Its dynamics are instead dominated by macroeconomic regime shifts and long-run cyclical behaviour rather than recurring month-of-year effects.

1.4 Outlier Detection

1.4.1 Robust MAD-based z-scores

Given the heavy-tailed nature of returns, outliers are detected using a robust z-score based on the median absolute deviation (MAD):

$$z_t = \frac{r_t - \text{median}(r)}{1.4826 \cdot \text{MAD}(r)}, \tag{1.2}$$

where 1.4826 scales MAD to be comparable to the standard deviation under normality. Observations with $|z_t| > 4$ are flagged as extreme.

1.4.2 Outlier results

Using the threshold ($|z_t| > 4$), 63 outliers are detected. The most extreme observations include:

- 2008–09–17: $+10.697\%$ (RobustZ ≈ 12.67)
- 2013–04–15: -9.191% (RobustZ ≈ -11.01)
- 2008–10–10: -7.721% (RobustZ ≈ -9.26)

These dates coincide with known stress regimes (e.g., the Global Financial Crisis and major gold selloffs), supporting the interpretation that outliers are associated with macro/market dislocations rather than isolated measurement errors. Figure 1.8 shows the the outliers of the returns.

[1] See the results in Appendix A.2

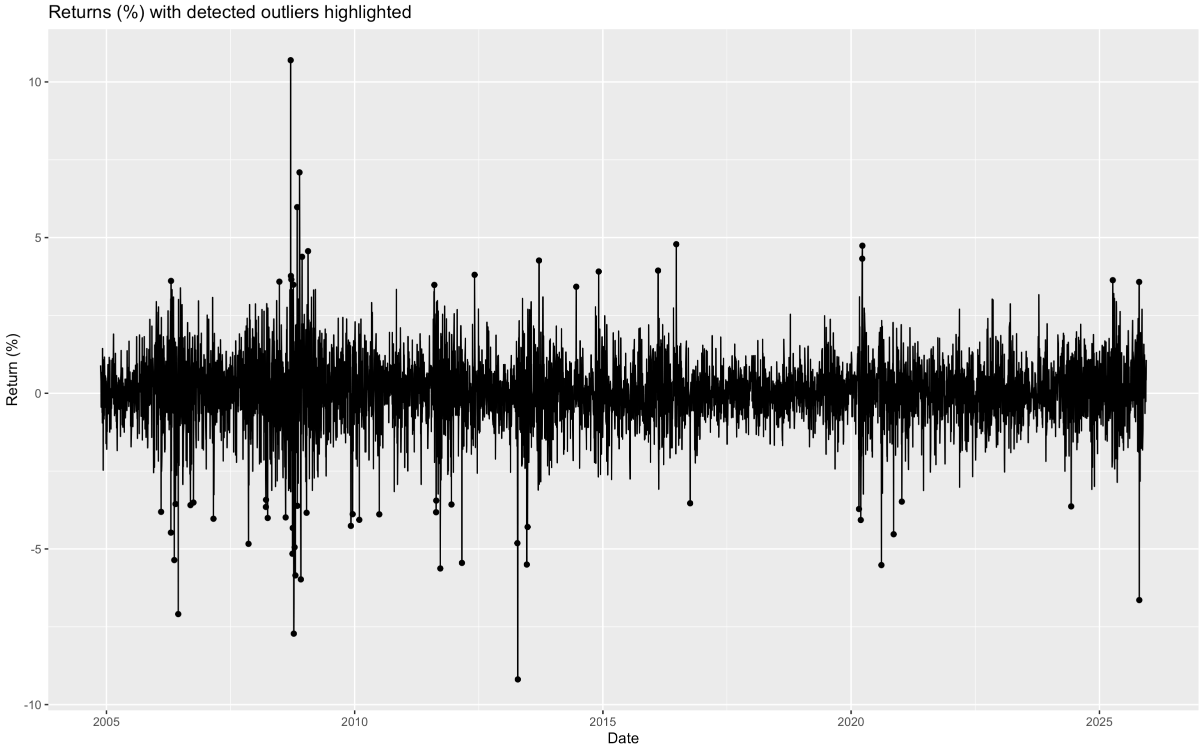


Figure 1.8: *Daily returns with MAD-based outliers highlighted* ($|z_t| > 4$).

1.5 Temporal Dependence and Volatility Clustering

1.5.1 ACF and PACF

The literature suggests that gold returns exhibit weak serial correlation (Cont, 2001), whereas squared returns display persistent positive autocorrelation, reflecting pronounced volatility clustering (Bhuyan, 2021; Çankaya, 2023). We observe these results in our ACF and PACF graphs.

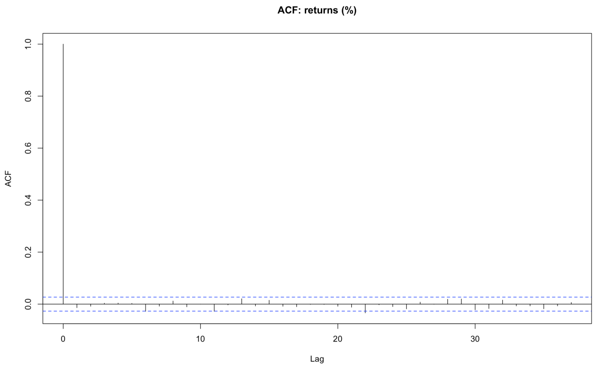


Figure 1.9: *ACF of gold returns. No significant autocorrelation is observed, indicating weak linear dependence in the conditional mean.*

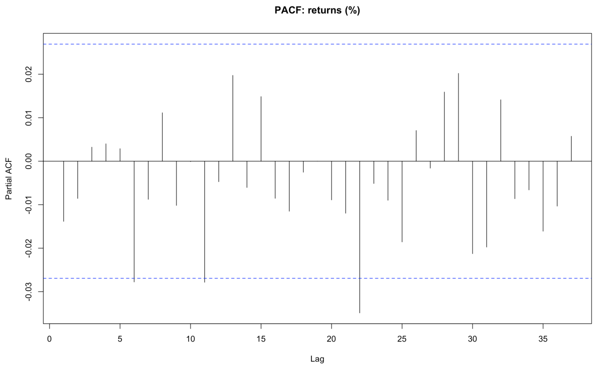


Figure 1.10: *PACF of gold returns. The absence of significant partial autocorrelation confirms that returns behave approximately as white noise.*

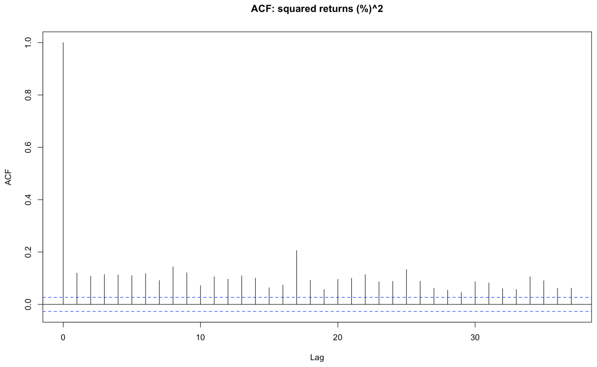


Figure 1.11: *ACF of squared returns. Persistent positive autocorrelation indicates volatility clustering (ARCH/-GARCH effects).*

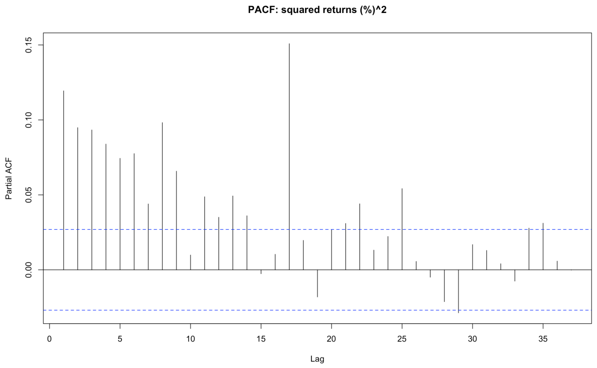


Figure 1.12: *PACF of squared returns. Significant partial autocorrelation confirms strong volatility persistence.*

The correlograms of gold returns reveal no significant serial dependence in the conditional mean, as both the ACF and PACF of returns lie almost entirely within the confidence bounds. This indicates that gold returns are approximately serially uncorrelated and can be modelled as a white-noise process. In contrast, the ACF and PACF of squared returns display strong and persistent positive autocorrelation over many lags, providing clear evidence of pronounced volatility clustering and long-memory behaviour. These features justify the use of ARCH/GARCH-type models for modelling conditional variance.

We conducted an exhaustive grid search over all integer pairs $(p, q) \in \{0, \dots, 9\}^2$ and selected the model minimizing the Akaike Information Criterion (AIC). As expected, the ARMA$(0, 0)$ model was identified as having the best fit.

1.5.2 Ljung–Box test

We assess the presence of temporal dependence in gold log-returns using Ljung–Box portmanteau tests applied at 20 lags. When applied to raw returns r_t, the test fails to reject the null hypothesis of no serial correlation ($p = 0.68$), indicating that daily gold returns exhibit weak linear dependence and behave approximately as white noise in the conditional mean. This suggests that only limited predictability is present in return directions.

In contrast, the Ljung–Box test applied to squared returns r_t^2 strongly rejects the null hypothesis ($p < 2.2 \times 10^{-16}$), revealing substantial positive autocorrelation in second moments. This provides clear evidence of pronounced volatility clustering, whereby periods of high (low) volatility tend to persist over time. These results motivate the use of conditional heteroskedasticity models, such as GARCH-type specifications, to capture the dynamic behavior of gold price volatility.

1.5.3 Augmented Dickey–Fuller test

An ADF test is run on r_t with automatic lag selection (reported lag order 17). The test statistic is:

- Dickey–Fuller statistic: -17.626
- Reported p-value: 0.01 (with warning indicating the true p-value is smaller than the printed bound)

We reject the unit root null, concluding that daily returns are stationary, consistent with standard empirical stylized facts.

1.5.4 Mean Predictability

Model selection based on information criteria indicates that the optimal linear specification for daily gold returns is ARMA$(0, 0)$, i.e.,

$$r_t = \mu + \varepsilon_t,$$

which corresponds to a constant mean with no autoregressive or moving–average dynamics. This result implies that past returns contain essentially no linear predictive information for the conditional mean of gold returns at the daily frequency.

Consequently, further improvements in mean forecasting accuracy are unlikely to be achieved by increasing the order of ARMA or FARIMA models. Instead, potential gains must rely on alternative modeling structures, such as (i) the inclusion of economically motivated exogenous predictors (ARMAX–type models), (ii) regime-switching specifications allowing for state-dependent means.

1.6 Long–Memory Diagnostics and FARIMA Assessment

1.6.1 FARIMA Estimation on Returns

FARIMA as been proved to be useful for gold price forecasting price (Sahed et al., 2020). Therefore, we first investigate whether gold returns exhibit long memory in the conditional mean by estimating a FARIMA$(0, d, 0)$ specification. Using a direct nonlinear least squares implementation of the FARIMA recursion, we obtain

$$\hat{d} = 0.001,$$

which is statistically indistinguishable from zero. Residual diagnostics confirm the adequacy of this short-memory specification: the Ljung–Box test on residuals up to lag 20 yields $p = 0.66$, so the null of no remaining serial correlation cannot be rejected.

To benchmark these results, we also estimate an ARFIMA$(0, d, 0)$ model using the `arfima` package. The corresponding estimate of the fractional parameter is $\hat{d} = -0.015$ (standard error 0.011, $p = 0.15$), again statistically indistinguishable from zero. Both approaches therefore consistently reject the presence of long memory in the conditional mean of gold returns.

1.6.2 Long Memory in Volatility Proxies

We next examine long-range dependence in volatility proxies, namely absolute returns $|r_t|$ and squared returns r_t^2. ARFIMA$(0, d, 0)$ estimations yield:

Series	$\hat{d}$	p–value
$\lvert r_t \rvert$	0.144	$< 2 \times 10^{-16}$
r_t^2	0.132	$< 2 \times 10^{-16}$

In both cases, the fractional differencing parameter is positive and highly significant, providing strong evidence of long memory in volatility dynamics. This pattern of short memory in returns but persistent fractional integration in volatility is a well-documented stylized fact of financial and commodity markets.

1.6.3 Implications for Modeling : choosing the correct GARCH

The absence of long memory in returns combined with pronounced fractional integration in volatility implies that FARIMA dynamics are not appropriate for the conditional mean of gold returns. Instead, persistence is concentrated in the conditional variance. Consequently, the natural modeling framework for gold prices is an ARMA mean specification coupled with a FIGARCH-type volatility process, which explicitly accommodates fractional integration in volatility while preserving short-memory mean dynamics.

Table 1.3 reports the in-sample fit criteria, residual diagnostics, and out-of-sample variance forecast performance for several GARCH-family specifications. All models adequately whiten the standardized residuals, as indicated by Ljung–Box and Engle ARCH-LM tests failing to reject the null hypotheses of no remaining serial correlation and no remaining conditional heteroskedasticity. Model selection is therefore driven by parsimony and forecasting accuracy.

Table 1.3: *Comparison of GARCH-family volatility models*

Model	AIC	BIC	$\text{QLIKE}_{\text{OOS}}$	MSE_{OOS}	$\text{LB}_{10}(\text{res})$	$\text{LB}_{10}(\text{res}^2)$	ARCH-LM_{12}
GARCH(1,1)-t	5342.03	5369.22	**1.2018**	**9.607**	0.518	0.703	0.821
GJR-GARCH(1,1)-t	5331.14	5363.77	1.2206	9.806	0.641	0.568	0.695
EGARCH(1,1)-t	5335.87	5368.50	1.2184	9.904	0.625	0.479	0.610
FIGARCH(1,1)-t	5343.39	5376.02	1.2680	10.200	0.487	0.850	0.932

Among the candidate models, the symmetric GARCH$(1,1)$ specification with Student-t innovations achieves the lowest out-of-sample QLIKE loss, indicating superior variance forecasting performance. Asymmetric specifications (GJR-GARCH and EGARCH) do not improve predictive accuracy, suggesting the absence of a pronounced leverage effect in gold returns. Furthermore, the FIGARCH model, which allows for long-memory dynamics in volatility, performs substantially worse.

The selected GARCH$(1,1)$-t model exhibits a highly persistent but stationary volatility process, with $\alpha + \beta \approx 0.995$, and heavy-tailed innovations $(\nu \approx 6)$. These results indicate that gold volatility is characterized by strong short-memory persistence and fat tails.

GARCH-X specifications, incorporating macroeconomic variables into the variance equation, were also estimated. However, they did not lead to any improvement in model fit. In fact, the Akaike Information Criterion slightly increased when exogenous regressors were included (AIC = 5442.83 for GARCH-X versus AIC = 5342.03 for the standard GARCH(1,1)). We therefore retained the simpler GARCH(1,1) representation of the volatility.

2

LSTM MODEL

2.1 LSTM–GARCH model

2.1.1 Data and Target Construction

To address the limitations of linear models in capturing complex temporal dependencies, this section introduces a Deep Learning approach based on Long Short-Term Memory (LSTM) networks, specifically designed to extract non-linear predictive signals from historical return sequences. Recent literature on gold price forecasting demonstrates a growing paradigm shift toward Deep Learning (Amini et al., 2024), driven by its superior ability to capture the non-linear complexities inherent in precious metal markets.

To try and model the return on a given day we are going use an LSTM model. Like aforementioned we will incoporate exogenous factors in the LSTM model
At each date t, the feature vector $\mathbf{x}_t$ combines the current gold return and macro-financial shocks:

$$\mathbf{x}_t = \big(r_t,\ \Delta \mathrm{Real10y}_t,\ \Delta \mathrm{BE10y}_t,\ \Delta \mathrm{VIX}_t,\ \Delta \log(\mathrm{DXY}_t)\big),$$

where:

- $\mathrm{Real10y}_t$ is the 10-year U.S. real yield (FRED series `DFII10`);
- $\mathrm{BE10y}_t$ is 10-year breakeven inflation (FRED series `T10YIE`);
- VIX_t is the implied volatility index (FRED series `VIXCLS`);
- DXY_t is the U.S. Dollar Index (ticker `DX-Y.NYB`).

The macro predictors enter as daily differences (or log-differences for DXY), which is consistent with using *innovations* rather than levels. The LSTM receives a rolling window of the most recent L observations (here $L = 250$, a hyperparameter determined from the validation set):

$$\mathbf{X}_t = [\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t] \in \mathbb{R}^{L\times d},$$

with d denoting the number of features. Each training example pairs $\mathbf{X}_t$ with the corresponding target $y_t = r_{t+1}$.

2.1.2 Train/Test Split and Strict Out-of-Sample Design

To avoid look-ahead bias, the dataset is divided chronologically into a training segment and a test segment. The last `test_size` observations are reserved for evaluation, and the LSTM sequences in the test set are built only from information available within the test window.

Feature standardisation is performed using statistics estimated **only on the training set**. That is, if μ_X and σ_X are the training mean and standard deviation (computed feature-wise), then

$$\tilde{\mathbf{x}}_t = \frac{\mathbf{x}_t - \mu_X}{\sigma_X},$$

and the same transformation is applied to the test set.

Daily returns are small in magnitude, which can lead to near-zero predictions if the learning signal is weak (a common issue in return forecasting). To stabilise optimisation, the target is normalised using **training-only** moments:

$$\tilde{y}_t = \frac{y_t - \mu_y}{\sigma_y},$$

where μ_y and σ_y are computed on the training targets. The model outputs $\widehat{\tilde{y}}_t$, which is converted back to returns via

$$\hat{y}_t = \widehat{\tilde{y}}_t \sigma_y + \mu_y.$$

This scaling (and inverse transform) is essential to prevent "collapse-to-zero" behaviour and to ensure predictions have a realistic dispersion.

The LSTM is used to approximate the conditional expectation of next-day returns:

$$\hat{y}_t \approx \mathbb{E}[r_{t+1} \mid \mathbf{X}_t] .$$

The architecture consists of two stacked LSTM layers followed by a small feed-forward head :

- Layer Normalisation on inputs;
- LSTM with 64 units returning sequences;
- LSTM with 32 units returning the final hidden state;
- Dense layer (ReLU) and Dropout;
- Linear output layer producing $\widehat{\tilde{y}}_t$.

Regularisation is kept light (small L^2 penalties and modest dropout) to avoid underfitting and overly-shrunk forecasts. Training uses the Adam optimizer with gradient clipping (clipnorm) to improve stability. See Appendix B, for more details on tuning the hyper-parameters.

A time-series-safe validation scheme is adopted: the final portion of the training sequences (approximately one trading year) is held out as validation, preserving temporal order.

2.1.3 Risk-Targeted Long/Short Strategy

The LSTM forecast provides a directional signal, while the GARCH forecast provides a scale (risk) estimate. We define a smoothed predictive signal using a short rolling mean (rebalanced every `HOLD_DAYS` = 5 days):

$$s_t = \mathrm{sign}\left(\frac{1}{H} \sum_{j=0}^{H-1} \hat{y}_{t-j} \right),$$

with H = `HOLD_DAYS`. The target portfolio weight is then chosen to match a desired daily volatility $\sigma^\star$:

$$w_t^\star = \mathrm{clip}\left(\frac{\sigma^\star}{\hat{\sigma}_t} s_t, \; -w_{\max}, \; w_{\max} \right),$$

where $w_{\max}$ is a leverage cap.
To avoid using contemporaneous information, execution uses a one-day lag:

$$w_t^{\text{exec}} = w_{t-1}^{\star}.$$

Portfolio equity evolves according to arithmetic returns:

$$\text{Equity}_t = \text{Equity}_{t-1} + \left(w_t^{\text{exec}}\text{Equity}_{t-1}\right) R_t - \text{Cost}_t.$$

Transaction costs are modelled as proportional to dollar turnover:

$$\text{Cost}_t = c\,|w_t^{\text{exec}} - w_{t-1}^{\text{exec}}|\,\text{Equity}_{t-1},$$

where c is the per-dollar cost rate (e.g., 1 basis point).

2.1.4 Benchmark Strategies and Evaluation

Performance is compared against a standard baseline *Buy & Hold*, for a more detailed explanation of the trading strategy see Appendix C.

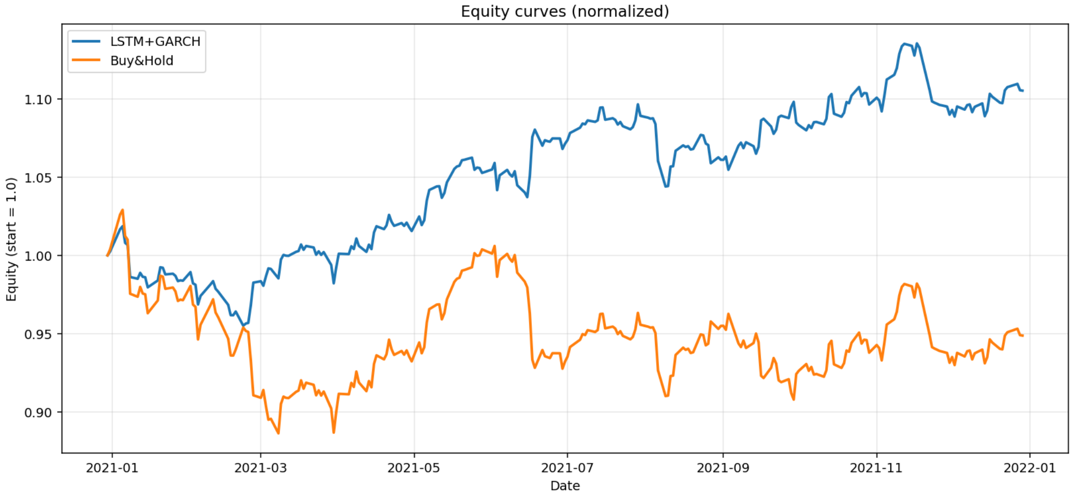


Figure 2.1: *Normalized equity curves comparing the LSTM–GARCH trading strategy with a buy-and-hold benchmark over the out-of-sample period. The active strategy exhibits smoother growth and reduced drawdowns relative to the passive benchmark.*

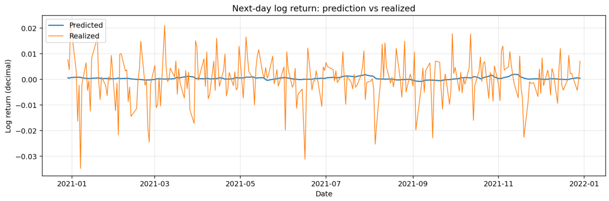


Figure 2.2: *Next-day log returns: comparison between LSTM-predicted returns and realized returns. Predicted returns are smoother and primarily capture short-term directional signals rather than exact magnitudes.*

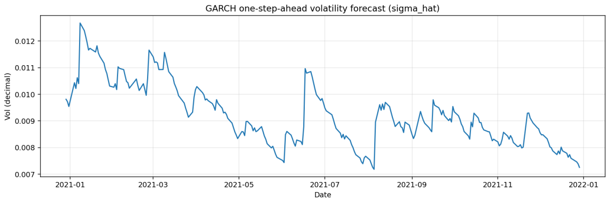


Figure 2.3: *One-step-ahead conditional volatility forecasts obtained from the GARCH model. Volatility clustering and gradual mean reversion are clearly visible.*

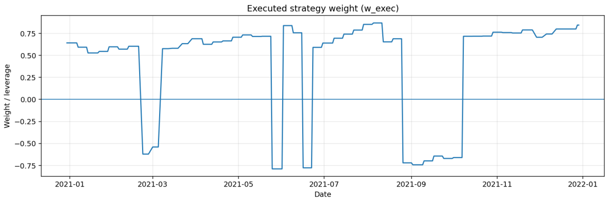


Figure 2.4: *Executed portfolio weights over time after combining return forecasts with volatility targeting. Position sizes adjust dynamically across market regimes, including periods of short exposure.*

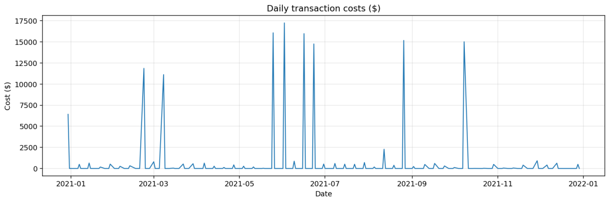


Figure 2.5: *Daily transaction costs incurred by the trading strategy. Spikes correspond to major portfolio rebalancing events, while average costs remain limited.*

To assess the predictive performance of the LSTM model, we first compute the root mean squared error (RMSE)[1] of the daily log-return forecasts. The resulting RMSE is equal to 0.01839, corresponding to an average forecast error of approximately 2% of the underlying daily return error. While this metric provides a useful indication of pointwise prediction accuracy, it remains difficult to evaluate the economic relevance of the model based solely on RMSE. In particular, a low forecasting error does not necessarily translate into superior trading performance. For this reason, we complement the RMSE analysis with more economically meaningful performance metrics, reported in Table 2.1. These indicators allow us to assess the risk-adjusted returns, downside risk, and implementation costs of the proposed strategy over rolling out-of-sample test windows.

Table 2.1: *Five-year (2021-2025) average performance of the LSTM+GARCH strategy across rolling test windows.*

Metric	**CAGR (%)**	**Ann Vol (%)**	**Sharpe**	**Max DD (%)**	**Total Tcost ($)**
5-year average	11.25	10.09	1.09	-7.22	192,305

Note: Each test window spans approximately one trading year; hence CAGR coincides with annual return in the reported averages.

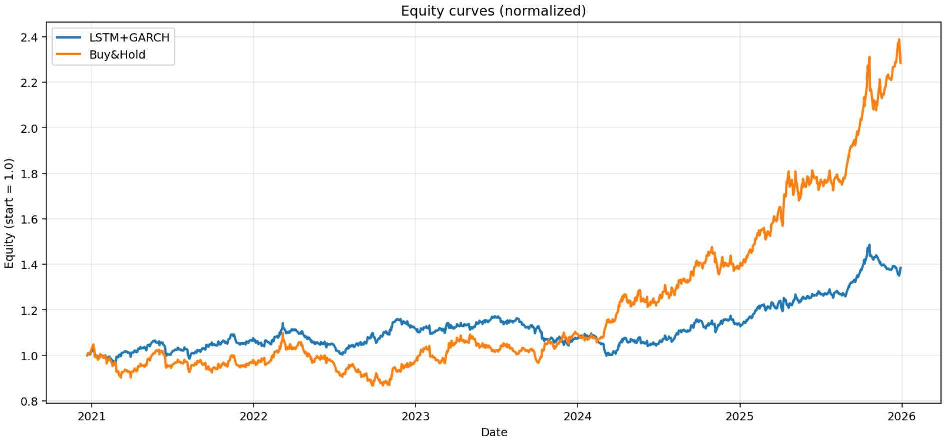


Figure 2.6: *Normalized equity curves of the LSTM–GARCH strategy and a buy-and-hold benchmark over the period 2021–2025.*

[1] More details on the metrics are provided in Appendix A.3

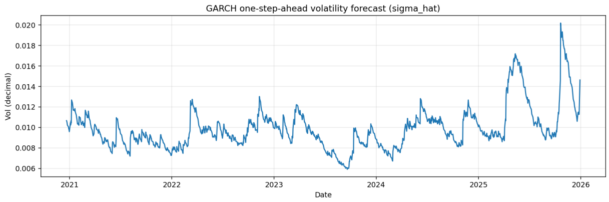


Figure 2.7: *One-step-ahead conditional volatility forecasts $\hat{\sigma}_{t+1}$ from the GARCH model. The series exhibits pronounced volatility clustering, with sharp spikes during periods of market stress, particularly toward the end of the sample.*

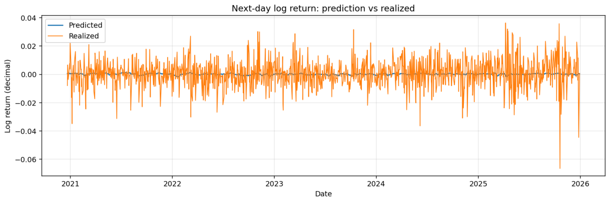


Figure 2.8: *Next-day log-return forecasts compared with realized returns. While predicted returns remain close to zero, realized returns display large fluctuations, highlighting the difficulty of short-horizon return prediction despite time-varying volatility.*

The performance metrics reported in Table 2.1 summarize the average behavior of the LSTM–GARCH strategy across rolling out-of-sample windows. The strategy delivers an average annual return of 11.25% with an annualized volatility close to 10%, resulting in a Sharpe ratio above one. In addition, the maximum drawdown remains limited to approximately 7%, highlighting the strategy's effective downside risk control, albeit at the cost of non-negligible transaction costs due to active rebalancing.

Figure 2.6 provides a visual comparison between the normalized equity curves of the LSTM–GARCH strategy and a passive buy-and-hold benchmark. Over the first part of the sample (2021–2023), the LSTM–GARCH strategy exhibits a relatively smooth and stable evolution, with limited drawdowns and moderate growth. In contrast, the buy-and-hold strategy experiences larger fluctuations and prolonged stagnation, particularly during the 2021–2022 period, indicating weaker downside protection in adverse market conditions.

From 2024 onward, the buy-and-hold benchmark enters a pronounced bullish phase, leading to a sharp acceleration in cumulative performance and a substantial outperformance in terms of raw returns by the end of the sample. During this period, the LSTM–GARCH strategy continues to grow at a steadier pace, reflecting its conservative, volatility-targeted design and its reduced exposure during high-risk regimes.

Overall, the figure illustrates the trade-off between absolute performance and risk control. While the buy-and-hold strategy benefits fully from strong upward market trends, the LSTM–GARCH strategy prioritizes risk-adjusted performance and capital preservation, delivering smoother equity dynamics and lower drawdowns across varying market regimes.

3

GRADIENT BOOSTING

3.1 Gradient Boosting method for nonlinear modeling of gold returns

Machine learning algorithms have recently emerged as powerful alternatives to traditional time series forecasting methods, particularly for data exhibiting high volatility and nonlinear patterns. In the context of gold price modeling, gradient boosting methods have demonstrated superior performance in capturing irregular dynamics, as explained by Jabeur et al. (2021).

From a statistical learning perspective, gradient boosting is an ensemble method that aims at estimating the regression function $f(x) = \mathbb{E}[Y \mid X = x]$, by minimizing iteratively an empirical risk:

$$\hat{f}(x) = \sum_{m=1}^{M} \beta_m h_m(x), \tag{3.1}$$

where $h_m(x)$ are weak learners and β_m are their respective weights. At each iteration m, a new weak learner is fitted to the negative gradient (pseudo-residuals) of the loss function:

$$r_i^{(m)} = -\left[\frac{\partial \mathcal{L}(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=\hat{f}^{(m-1)}}. \tag{3.2}$$

For squared loss $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$, the pseudo-residuals simplify to $r_i^{(m)} = y_i - \hat{f}^{(m-1)}(x_i)$, yielding a direct residual-correction interpretation. While smooth weak learners can be considered (e.g., splines or kernel estimators), regression trees offer a flexible alternative for capturing nonlinear structures particularly relevant in financial contexts, where data-generating processes rarely follow globally linear patterns.

The eXtreme Gradient Boosting (XGBoost) algorithm extends classical gradient boosting through:

- **Second-order approximation**: The loss function is approximated using both first and second derivatives (Newton-Raphson step), improving convergence:

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{n} \left[g_i h_m(x_i) + \frac{1}{2} h_i h_m^2(x_i)\right] + \Omega(h_m), \tag{3.3}$$

  where g_i and h_i denote the first and second derivatives of the loss with respect to $\hat{f}^{(m-1)}(x_i)$.
- **Explicit regularization**: Tree complexity is penalized through:

$$\Omega(h_m) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2, \tag{3.4}$$

where T is the number of leaves and w_j are leaf weights, effectively controlling overfitting.

In this section, we investigate whether XGBoost can extract weak predictive signals from historical information for gold returns. While exploratory analysis suggests limited unconditional predictability in the mean, we examine whether conditional information—embedded in lagged returns and macro-financial variables[1] can be exploited to improve out-of-sample forecasting accuracy.

3.2 XGBoost model

3.2.1 Prediction target and forecast horizon

The objective of this part is not to forecast daily gold returns, but to detect short- to medium-term price movements consistent with a swing trading horizon. We focus on predicting the cumulative gold log-return over a 10-day horizon and evaluate the model on an out-of-sample test period covering the year 2025. Indeed, predicting one-day-ahead returns $(h = 1)$ for highly liquid assets such as gold is notoriously difficult, as daily returns are largely dominated by noise and exhibit weak conditional mean predictability. Preliminary experiments at the daily horizon with XGBoost reveal almost no exploitable predictive signal, especially when macro-financial variables are included. By contrast, extending the forecast horizon to $h = 10$ days increases the signal-to-noise ratio by aggregating returns over time. The target variable is defined as the cumulative log-return:

$$y_t^{(10)} = \sum_{i=1}^{10} r_{t+i} = \log(P_{t+10}) - \log(P_t), \tag{3.5}$$

which represents the total return accumulated over the ten trading days following time t. This should not be interpreted as the return *on* day $t + 10$ itself, but rather the aggregate movement *between* t and $t + 10$. This choice improves the signal-to-noise ratio through return aggregation and aligns with a short- to medium-term investment horizon suitable for liquid ETF trading.

3.2.2 Modeling Framework and Walk-Forward Prediction

Let $\mathcal{F}_t$ denote the information set available at time t, represented by the feature vector $x_t \in \mathbb{R}^d$. To prevent look-ahead bias, the target variable is defined as the strictly forward-looking h-day cumulative return:

$$y_t^{(h)} = \sum_{k=1}^{h} r_{t+k}, \quad \text{with } h = 10. \tag{3.6}$$

The XGBoost framework estimates the conditional expectation $\hat{y}_t = \mathbb{E}[y_t^{(h)} \mid x_t]$ via gradient boosting. To address the non-stationarity inherent in financial time series and mitigate the effects of concept drift, we implement an expanding window Walk-Forward validation. At each time step t, the model is re-calibrated on the historical set $\{(x_i, y_i)\}_{i=1}^{t-h}$ to generate the out-of-sample prediction $\hat{y}_t$. This dynamic updating procedure mimics a realistic trading environment, allowing the model to adapt to evolving market regimes.

[1] An exhaustive description of the features can be found in Appendix A.4.

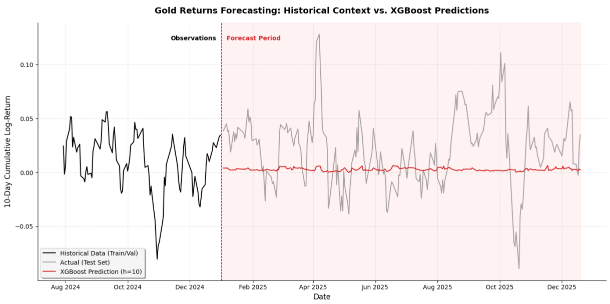


Figure 3.1: *Cumulative log-return predictions at horizon $h = 10$ obtained through the XGBoost procedure*

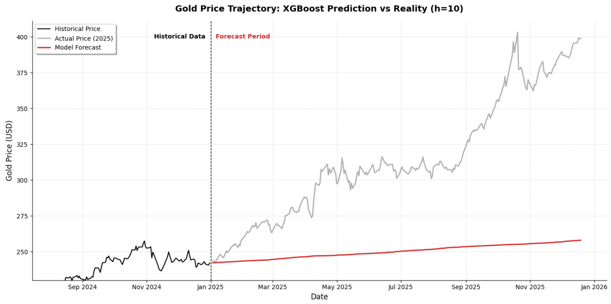


Figure 3.2: *Out-of-Sample Gold Price Trajectory (USD): Actual 2025 Bull Market vs. XGBoost Conservative Trend Forecast*

The visual analysis reveals the model's inherently conservative behavior and its focus on trend extraction over volatility tracking. In the returns graph, the XGBoost predictions (red line) exhibit significantly lower variance than the actual market data, effectively filtering out high-frequency noise to focus on a stable mean tendency. This 'smoothing effect' is confirmed by the price trajectory: while the model correctly identified the directional regime, maintaining a steady upward slope throughout 2025. It significantly underestimated the magnitude of the bull run. This indicates that the model functions as a trend follower, prioritizing reliable directional accuracy and risk minimization over capturing the full amplitude of extreme market momentum.

To maximize out-of-sample generalization, model hyperparameters (learning rate, tree depth, regularization) were optimized[2] on the validation set. The analysis of feature importance reveals[3] that the model relies primarily on trend and volatility regime indicators. This confirms that, for this horizon, the model acts as a momentum trader that filters signals based on market risk perception, rather than reacting to daily macro-economic noise.

Table 3.1: *Performance Comparison: Static Model vs. Walk-Forward vs. Naive Benchmark* (*Horizon $h = 10$*)

Metric	**Naive Benchmark**	**Static XGBoost**	**Walk-Forward XGBoost**
RMSE	0.039251	0.037582	**0.037265**
MAE	0.031389	0.029864	**0.029637**
Directional Accuracy	75.22% (Theoretical)	76.11%	**76.11%**

Note: The best results are indicated in bold. The Walk-Forward approach slightly outperforms the static approach and beats the naive benchmark across all error metrics.

Comparing the static evaluation with the Walk-Forward approach, we observe a slight improvement in robustness (RMSE decreasing from 0.0376 to 0.0373). However, despite a strong directional accuracy (76.11%), the standalone XGBoost model's regression performance remains underwhelming. The model only marginally outperforms the naive benchmark (0.0393), suggesting it struggles to distinguish signal from noise and fails to capture the full amplitude of market volatility.

To address this, we propose exploring a hybrid ARIMA-XGBoost approach, aiming to combine ARIMA's strength in handling linear time-series dynamics with XGBoost's ability to capture non-linear macroeconomic effects on the residuals.

[2] Optimization was done using the *Optuna* framework
[3] See Appendix A.4

3.3 ARIMA-XGBoost

Following the poor predictive performance of complex linear specifications—specifically the inability to capture extreme movements, often resulting in flat predictions—we shifted our focus toward a model capable of capturing non-linear patterns. Furthermore, we utilized weekly data values to naturally smooth out the noise.

We observed that the optimal linear model determined by the AIC criterion was an ARIMA$(0, 0, 0)$. This implies that the linear component of the series is essentially a white noise around a constant mean, confirming that traditional econometric models fail to find significant autocorrelation. Conversely, we observed that applying a standalone XGBoost model directly to the log returns yielded unsatisfactory results due to the noise.

Consequently, we concluded that a hybrid approach was necessary, as several studies have already combined ARIMA with machine learning methods for gold price prediction (Makala et al., 2021). Here, we utilize the ARIMA$(0, 0, 0)$ model as a baseline to capture the global mean (centering the data), and an XGBoost model to predict the residuals (the deviations from this mean), thereby focusing the machine learning effort on the volatility and non-linear shocks.

The methodology is as follows: regarding the training set prior to January 2025, we found that using the entire history was suboptimal; instead, restricting the training window to the last 3 years yielded the best predictive performance. On this dataset, the optimization algorithm converged to an ARIMA$(0, 0, 0)$ structure. We extracted the training residuals (which correspond to the demeaned returns). Subsequently, our XGBoost model (using features detailed below) was trained to predict these specific residuals. To prevent overfitting and maintain a robust generalization capability, we deliberately employed very simple trees, setting the maximum depth to 2 and the number of estimators to 20. This hybrid approach yielded significantly better results by isolating the non-linear signal from the statistical noise.

We can clearly distinguish the difference between the standalone ARIMA prediction, which remains a constant log return (representing the mean), and the hybrid XGBoost approach, which effectively captures specific movements in the log returns. Nevertheless, the volatility of the predicted log returns is significantly lower than that of the observed test data, which suggests that the model tends to smooth out extreme variations and underestimates the magnitude of market shocks.

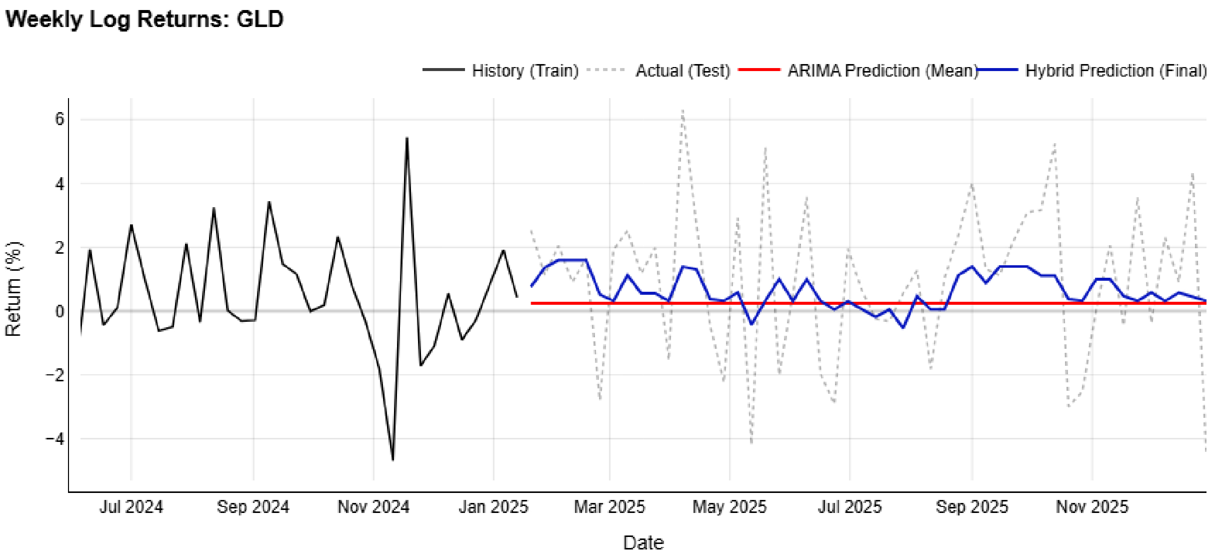


Figure 3.3: *Comparison of log returns predictions ARIMA vs Hybrid*

To rigorously assess the model's robustness, we performed a pure out-of-sample simulation starting in January 2025. The chart below displays a recursive price reconstruction, where the projected path relies exclusively on the cumulative predicted log returns applied to the last observation of 2024, assuming no knowledge of future prices.

Despite the conservative bias observed in the log returns, where the model tends to smooth out extreme volatility. The reconstructed price path demonstrates remarkable alignment with the actual market trend. This ability to capture the long-term drift is particularly significant considering the exceptional dynamics of 2025, which exhibited a sustained, near-exponential growth pattern that typically challenges standard regression models.

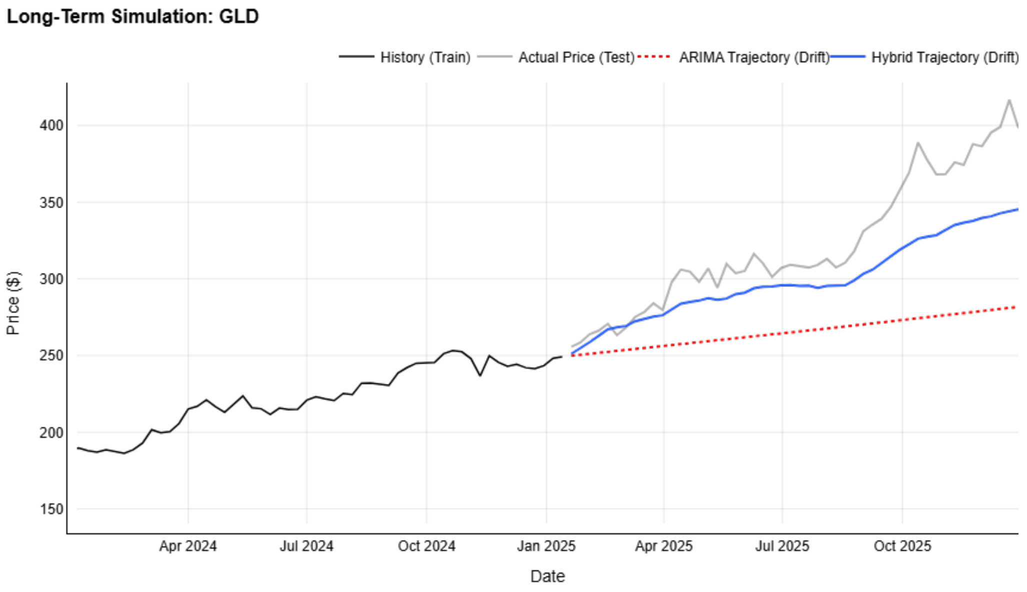
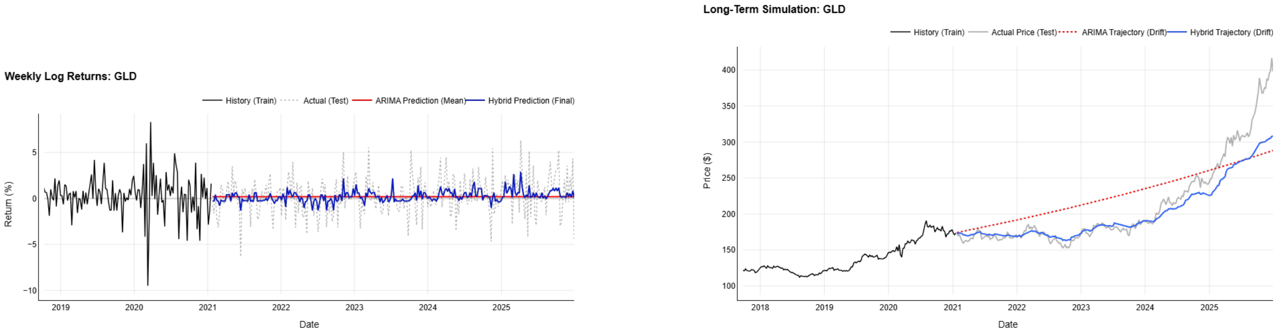


Figure 3.4: *Comparison of Gold Price predictions ARIMA vs Hybrid*

Even more impressive is the model's performance when tested on a rigorous out-of-sample dataset starting from January 2021, rather than 2025. This extended testing period encompasses diverse market trends, significantly increasing the complexity of the forecasting task. Despite this, the model performs exceptionally well and successfully captures the data's dynamics. Notably, for this 2021 start date, the optimal results were achieved by restricting the training set to the last 5 years prior to 2021, rather than using the entire history.



(**a**) *Log returns* (**b**) *Gold Price*

Figure 3.5: *Comparison of ARIMA vs Hybrid from 2021*

In conclusion, the table below summarizes the performance metrics—specifically RMSE, MAE, and Directional Accuracy—evaluating the Hybrid Model across the two tested horizons.

Table 3.2: *Performance comparison of the Hybrid Model (ARIMA + XGBoost) across different testing periods.*

Test Period	**Training**	**RMSE**		**MAE**		**Directional**	**Improvement**
(Start Date)	(Lookback)	Log Ret.	Price ($)	Log Ret.	Price ($)	**Accuracy**	**vs ARIMA**
Jan 2021	5 Years	0.0174	$24.70	0.0134	$14.46	68.99%	+14.2%
Jan 2025	3 Years	0.0224	$29.60	0.0180	$23.40	70.00%	+11.6%

Note: RMSE and MAE for Log Returns are expressed in decimal form. Improvement vs ARIMA is based on the reduction of the RMSE on log returns.

Consistent with the methodology applied to the LSTM architecture, the XGBoost model was trained using a similar set of input variables. This feature set combines technical indicators. Specifically the Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Lagged Returns, and Bollinger Bands, with key macroeconomic drivers, including the returns of the VIX, Interest Rates, and the US Dollar Index.

To interpret the model's decision-making process, we analyzed the *Feature Importance*. This metric is calculated by aggregating the improvement in accuracy (Gain) contributed by each feature across all decision trees in the ensemble. It effectively quantifies the relative influence of each variable in reducing the prediction error.

Figure 3.6 illustrates the hierarchy of the most influential features for both the short-term test (2025) and the structural long-term test (2021-2026).

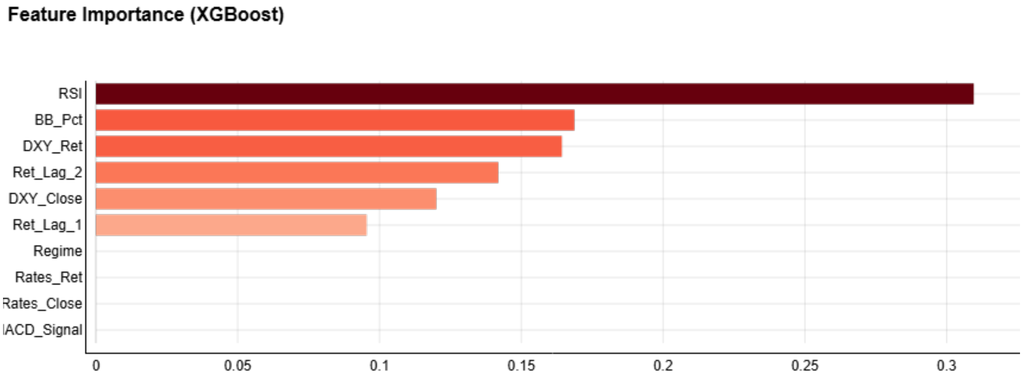


Figure 3.6: *Feature Importance Ranking (Short-term Test: 2025)*

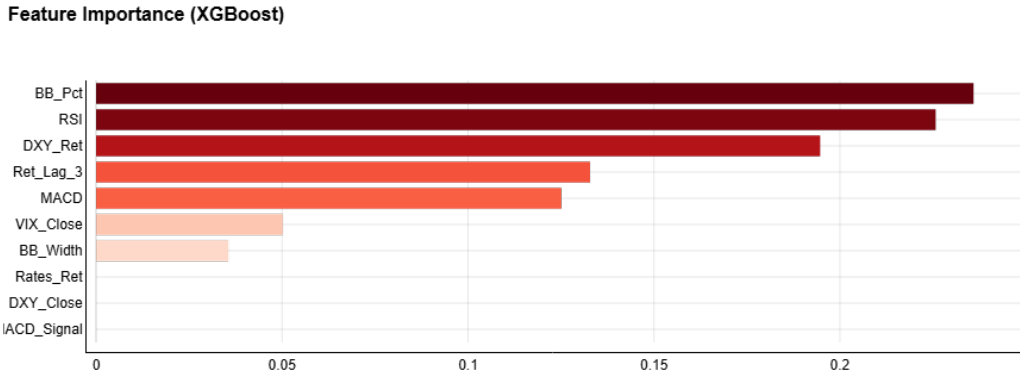


Figure 3.7: *Feature Importance Ranking (Long-term Test: 2021-2026)*

4

PORTFOLIO MANAGEMENT

Having established that predicting Gold log returns is a complex task—and observing that standalone predictions failed to outperform a classic Buy & Hold strategy—we expanded our scope to a multi-asset approach. We constructed a diversified portfolio comprising Indices, Commodities, and Forex to optimize risk-adjusted returns over time.

Our portfolio consists of the S&P 500, EUR/USD, Gold, US Treasury Bonds (20Y), CAC 40, and Crude Oil. We implemented a directional strategy based on structural trends using 50-week and 200-week Moving Averages (MA). A bullish position is taken when the Price $> MA(50) > MA(200)$, and a bearish position when the Price $< MA(50) < MA(200)$. If neither condition is met (neutral trend), we abstain from taking a position on the asset.

The portfolio is re-optimized monthly based on the Markowitz Mean-Variance framework to maximize the Sharpe Ratio. The expected returns are derived from our hybrid ARIMA-XGBoost model, while volatility is estimated using GARCH processes.

For the simulation, we employed a Walk-Forward backtest that rigorously accounts for transaction costs. To mitigate excessive trading fees, we implemented a turnover buffer mechanism.

The figures below illustrate the strategy's performance over a 5-year period (2021-2026), including cumulative returns, drawdowns, and the evolution of asset allocation and transaction costs.

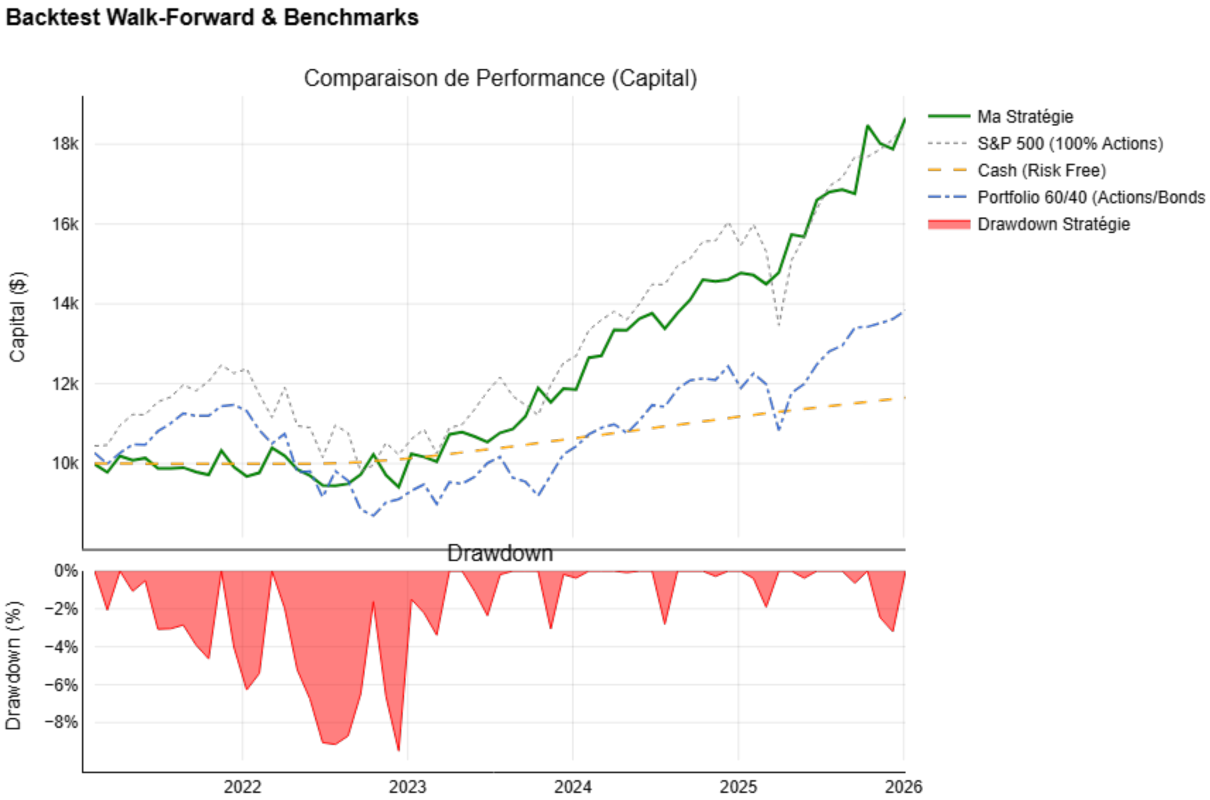
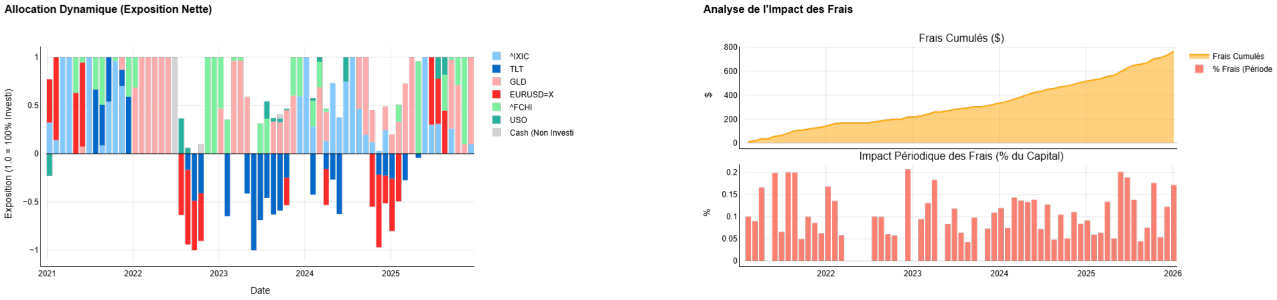


Figure 4.1: *Backtest Walk-Forward & Benchmarks*



(a) *Evolution of the allocation*

(b) *Evolution of the fees*

Figure 4.2

Table 4.1: *Performance Summary (2021-2025).*

CAGR	Vol.	Sharpe	Max DD	Win Rate	Alpha	Beta
13.55%	11.97%	0.97	-9.50%	53.85%	9.90%	0.12

CONCLUSION

This study investigated the dynamics of gold returns over the period 2004–2025 using a comprehensive modeling framework combining econometric approach, machine/deep learning and portfolio-level applications. The analysis confirms that gold returns exhibit the canonical stylized facts of financial time series, with weak serial dependence in the conditional mean, pronounced volatility clustering, heavy tails, and regime-dependent dynamics. While daily returns behave approximately as white noise, higher-order moments display strong persistence, justifying the use of conditional heteroskedasticity models.

From an econometric perspective, linear models provide limited predictive power for the mean of gold returns. Model selection procedures consistently identify an ARMA$(0, 0)$ specification as optimal, indicating the absence of exploitable linear autocorrelation. In contrast, volatility dynamics are highly persistent and non-Gaussian. Among GARCH-family models, a symmetric GARCH$(1, 1)$ delivers the best out-of-sample variance forecasts, outperforming asymmetric and long-memory alternatives. These results highlight that persistence in gold prices is concentrated in the conditional variance rather than in the mean.

Building on these diagnostics, the LSTM–GARCH framework demonstrates that deep learning, when paired with volatility targeting, extracts meaningful economic value despite noisy pointwise forecasts. The resulting strategy delivers superior risk-adjusted returns and limited drawdowns, confirming that model efficacy is best assessed through portfolio metrics rather than raw predictive accuracy alone. In parallel, the hybrid ARIMA–XGBoost model validates the synergy between statistical structure and nonlinear learning. By isolating residual dynamics via XGBoost after an ARIMA baseline, this approach captures complex market shocks, yielding substantial out-of-sample improvements in directional accuracy and trend reconstruction compared to purely linear specifications.

Finally, extending the analysis to a multi-asset portfolio context demonstrates the practical relevance of the proposed modeling framework. By combining trend-following signals and machine-learning-driven expected returns within a walk-forward optimization scheme, the resulting portfolio achieves robust performance and improved downside protection relative to standard benchmarks.

Overall, this work confirms that while gold returns are difficult to predict at short horizons, meaningful structure exists in volatility dynamics and nonlinear components. Hybrid modeling strategies that integrate machine learning flexibility and explicit risk management offer a promising avenue for both forecasting and portfolio construction. Future research could extend this framework by incorporating higher-frequency data, or by integrating unstructured data sources, such as real-time sentiment analysis, as proposed in Varela et al., 2025, which offers a promising avenue for capturing market psychology beyond traditional technical and macroeconomic indicators.

This page intentionally left blank.

BIBLIOGRAPHY

Amini, Amirhossein and Robab Kalantari (Mar. 2024). “Gold Price Prediction by a CNN-Bi-LSTM Model Along with Automatic Parameter Tuning”. In: *PLOS ONE* 19.3, e0298426. DOI: 10.1371/journal.pone.0298426.

Bhuyan, Bimal (2021). “Do LBMA gold prices follow a random walk? Evidence based on serial correlation and ARCH family models”. In: *Journal of Economic and Administrative Sciences* 37.4, pp. 451–472. DOI: 10.1108/JEAS-03-2020-0040.

Çankaya, Selim (2023). “Factors Impacting the Price of Gold: An Empirical Study of EGARCH Model”. In: *PressAcademia Procedia* 17, pp. 48–55. DOI: 10.17261/Pressacademia.2023.1715.

Cont, Rama (2001). “Empirical properties of asset returns: stylized facts and statistical issues”. In: *Quantitative Finance* 1.2, pp. 223–236.

Dao, Thong, Frank McGroarty, and Andrew Urquhart (June 2017). “Ultra-High-Frequency Pairs Trading in Gold ETFs”. In: SSRN Working Paper. URL: https://ssrn.com/abstract=2981717.

Hood, Matthew and Farooq Malik (2013). “Is gold the best hedge and a safe haven under changing stock market volatility?” In: *Review of Financial Economics* 22.2, pp. 47–52. DOI: 10.1016/j.rfe.2013.03.001.

Jabeur, S.B., S. Mefteh-Wali, and J.L. Viviani (2021). “Forecasting gold price with the XGBoost algorithm and SHAP interaction values”. In: *Annals of Operations Research* 334.2, pp. 679–699. DOI: 10.1007/s10479-021-04187-w.

Makala, D. and Z. Li (2021). “Prediction of gold price with ARIMA and SVM”. In: 1767, p. 012022. DOI: 10.1088/1742-6596/1767/1/012022.

Sahed, Abdelkader, Mohammeed Mekidiche, and Hacen Kahoui (2020). “Fuzzy Auto-Regressive Integrated Moving Average (FARIMA) Model for Forecasting the Gold Prices”. In: *Journal name* 5.1. ISSN: 2537-141X. URL: https://jseg.ro/index.php/jseg/article/view/90 (visited on 2025-01-16).

Thilak et al. (2025). “Gold Price Analysis and Forecasting Using a Machine Learning Approach”. In: *Proceedings of the 1st Engineering Data Analytics and Management Conference (EAMCON 2025)*. Available Online: 31 December 2025. Udupi, India: Atlantis Press, p. 33. DOI: 10.2991/978-94-6463-978-0_33.

Varela, Angel et al. (Sept. 2025). “An Analytical Framework for Real-Time Gold Trading Using Sentiment and Time-Series Forecasting”. In: *Decision Analytics Journal* 17. Received: 30 May 2025; Revised: 15 August 2025; Accepted: 9 September 2025; Available online: 12 September 2025, p. 100633. ISSN: 2772-6622. DOI: 10.1016/j.dajour.2025.100633.

This page intentionally left blank.

Appendices

This page intentionally left blank.

A APPENDIX

A.1 Seasonality

Additive decomposition. Let $\{X_t\}_{t=1}^{T}$ denote the monthly gold price series. We assume that X_t admits the classical additive representation

$$X_t = T_t + S_t + R_t,$$

where T_t denotes the trend component, S_t the seasonal component and R_t the irregular (random) component.

The trend T_t is extracted by applying a centered moving average filter of order 12, corresponding to the monthly frequency of the data. This operation removes short-term fluctuations and seasonal effects, thereby isolating the low-frequency, long-run evolution of gold prices.

The seasonal component S_t is then obtained by averaging the detrended series $(X_t - T_t)$ over each calendar month across all years. This yields a set of twelve seasonal coefficients that capture systematic intra-year patterns. These coefficients are normalized to sum to zero over a full year, ensuring identifiability in the additive decomposition.

Finally, the irregular component is computed as the residual

$$R_t = X_t - T_t - S_t,$$

which captures unpredictable short-term shocks and market-specific noise.

A.2 Formal calendar-month seasonality test (ANOVA on monthly returns)

To test whether average monthly returns differ across months, we estimate:

$$r_t^{(m)} = \mu + \sum_{j=1}^{11} \gamma_j \mathbb{1}\{\text{Month}_t = j\} + u_t, \tag{A.1}$$

and perform an ANOVA F-test for equality of monthly means. The ANOVA table is:

Source	Df	Sum Sq	F value	p-value
Month	11	216	0.849	0.591
Residuals	241	5564		

The p-value (0.591) implies we do not reject the null hypothesis of equal mean returns across months. Thus, within this sample and specification, there is no strong evidence of stable calendar-month seasonality in monthly returns.

A.3 Performance metrics

To evaluate the predictive performance of our models on Gold returns, we employ four standard statistical metrics. These indicators quantify the discrepancy between the actual observed values (Y_h) and the values predicted by the model ($\hat{Y}_h$) over a test horizon of size N.

1. **Root Mean Squared Error (RMSE):** Derived from the MSE, the RMSE represents its square root. Its primary advantage is that it is expressed in the same unit as the target variable (in this case, log returns), making it more intuitively interpretable than the MSE for assessing the average magnitude of the error. Like the MSE, it remains sensitive to outliers.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{h=1}^{N}(Y_h - \hat{Y}_h)^2} \tag{A.2}$$

2. **Mean Absolute Error (MAE):** The MAE calculates the average of the absolute deviations. Unlike the RMSE, it assigns linear weight to all errors, making it more robust (less sensitive) to outliers. A significant divergence where the RMSE is much larger than the MAE indicates that the model is making a few large, isolated errors.

$$\text{MAE} = \frac{1}{N}\sum_{h=1}^{N}|Y_h - \hat{Y}_h| \tag{A.3}$$

A.4 Features used in the XGBoost models and feature importance

Table A.1: *Features used in the XGBoost model*

Category	Feature	Description
	r_lag_1 to r_lag_10	Lagged gold returns (1 to 10 days)
Technical indicators	r_ma_5	5-day moving average of returns
	r_ma_20	20-day moving average of returns
	vol_5	5-day rolling volatility
Lagged returns	vol_20	20-day rolling volatility
Macroeconomic variables	d_dxy	Change in US Dollar Index (DXY)
	d_real	Change in 10-year real interest rate
	d_vix	Change in VIX volatility index
	d_silver	Log return of silver prices
	d_oil	Log return of oil prices
	d_breakeven	Change in 10-year breakeven inflation
	RSI	Relative Strength Index

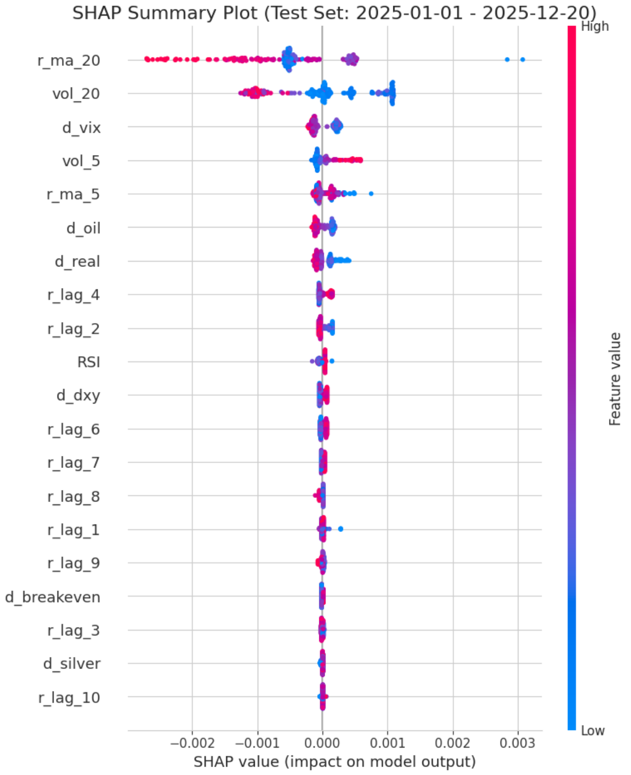
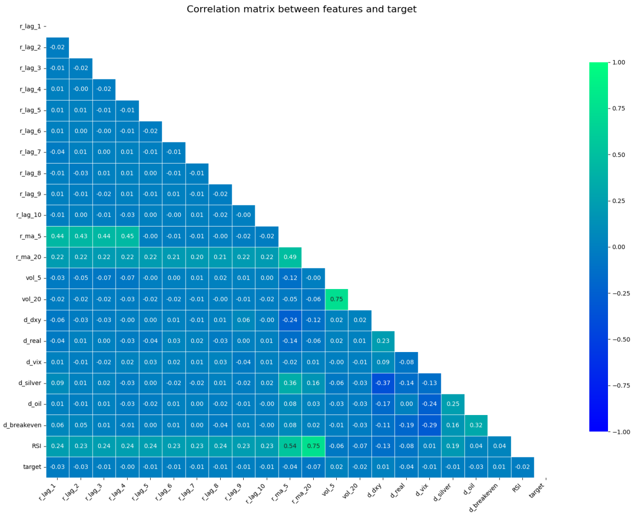


Figure A.1: *Shap plot of the XGBoost model*

The SHAP summary plot provides deeper insight into the model's decision-making process. Derived from cooperative game theory, Shapley values quantify the marginal contribution of each feature to the prediction's deviation from the average baseline. The SHAP summary plot aggregates these local contributions, offering a unified view of both global feature importance and the direction of their impact. As expected, momentum variables liker`_ma_20` show a positive correlation with the forecast (red dots on the right). However, interestingly, high volatility (`vol_20`) generally pushes predictions to the left (negative impact). This indicates that the XGBoost algorithm has intrinsically learned a risk-aversion behavior: during periods of high market stress, it tends to lower its return expectations, effectively acting as a risk-adjusted trend follower



	r_lag_1	r_lag_2	r_lag_3	r_lag_4	r_lag_5	r_lag_6	r_lag_7	r_lag_8	r_lag_9	r_lag_10	r_ma_5	r_ma_20	vol_5	vol_20	d_dxy	d_real	d_vix	d_silver	d_oil	d_breakeven	RSI	target
r_lag_1																						
r_lag_2	-0.02																					
r_lag_3	-0.01	-0.02																				
r_lag_4	0.01	-0.00	-0.02																			
r_lag_5	0.01	0.01	-0.01	-0.01																		
r_lag_6	0.01	0.00	-0.00	-0.01	-0.02																	
r_lag_7	-0.04	0.01	0.00	0.01	-0.01	-0.01																
r_lag_8	-0.01	-0.03	0.01	0.01	0.00	-0.01	-0.01															
r_lag_9	0.01	-0.01	-0.02	0.01	-0.01	0.01	-0.01	-0.02														
r_lag_10	-0.01	0.00	-0.01	-0.03	0.00	-0.00	0.01	-0.02	-0.00													
r_ma_5	0.44	0.43	0.44	0.45	-0.00	-0.01	-0.01	-0.00	-0.02	-0.02												
r_ma_20	0.22	0.22	0.22	0.22	0.22	0.21	0.20	0.21	0.22	0.22	0.49											
vol_5	-0.03	-0.05	-0.07	-0.07	-0.00	0.00	0.01	0.02	0.01	0.00	-0.12	-0.00										
vol_20	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.01	-0.00	-0.01	-0.02	-0.05	-0.06	0.75									
d_dxy	-0.06	-0.03	-0.03	-0.00	0.00	0.01	-0.01	0.01	0.06	-0.00	-0.24	-0.12	0.02	0.02								
d_real	-0.04	0.01	0.00	-0.03	-0.04	0.03	0.02	-0.03	0.00	0.01	-0.14	-0.06	0.02	0.01	0.23							
d_vix	0.01	-0.01	-0.02	0.02	0.03	0.02	0.01	0.03	-0.04	0.01	-0.02	0.01	-0.00	-0.01	0.09	-0.08						
d_silver	0.09	0.01	0.02	-0.03	0.00	-0.02	-0.02	0.01	-0.02	0.02	0.36	0.16	-0.06	-0.03	-0.37	-0.14	-0.13					
d_oil	0.01	-0.01	0.01	-0.03	-0.02	0.01	0.00	-0.02	-0.01	-0.00	0.08	0.03	-0.03	-0.03	-0.17	0.00	-0.24	0.25				
d_breakeven	0.06	0.05	0.01	-0.01	0.00	0.01	0.00	-0.04	0.01	-0.00	0.08	0.02	-0.01	-0.03	-0.11	-0.19	-0.29	0.16	0.32			
RSI	0.24	0.23	0.24	0.24	0.24	0.23	0.23	0.24	0.23	0.23	0.54	0.75	-0.06	-0.07	-0.13	-0.08	0.01	0.19	0.04	0.04		
target	-0.03	-0.03	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.04	-0.07	0.02	-0.02	0.01	-0.04	-0.01	-0.01	-0.03	0.01	-0.02	

Figure A.2: *Covariance matrix between features and target*

B

HYPERPARAMETER TUNING

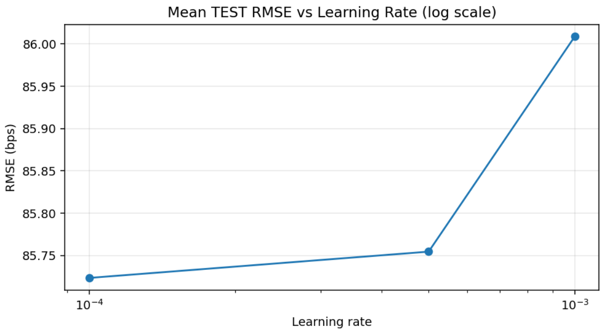


Figure B.1: *Heatmap of test RMSE (in basis points) as a function of the learning rate and look-back window size. Larger window sizes consistently reduce prediction error, while excessively large learning rates degrade performance.*

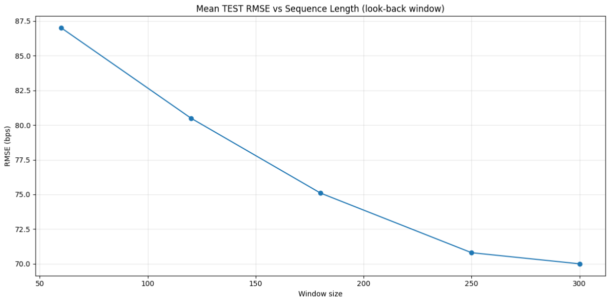


Figure B.2: *Average test RMSE as a function of the look-back window size. Prediction accuracy improves monotonically with longer sequences, suggesting that long-term temporal dependencies are relevant for return dynamics.*

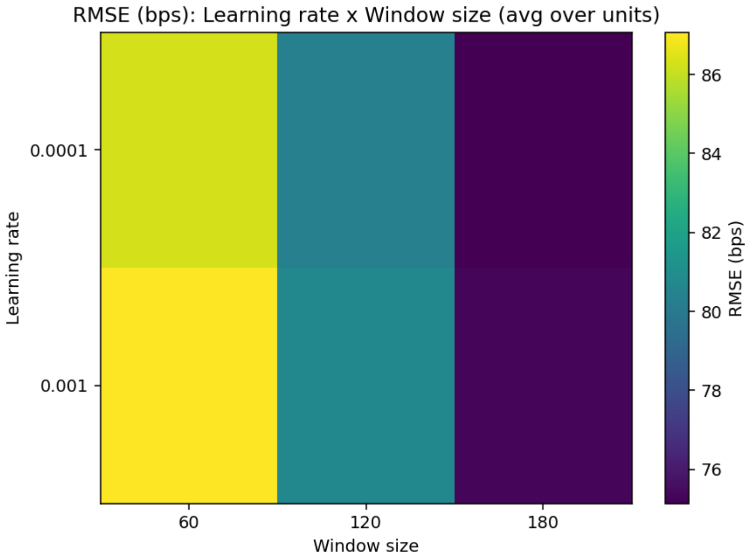


Figure B.3: *Mean test RMSE as a function of the learning rate (log scale). Lower learning rates yield more stable convergence and improved generalization, while overly aggressive learning rates increase forecasting error.*

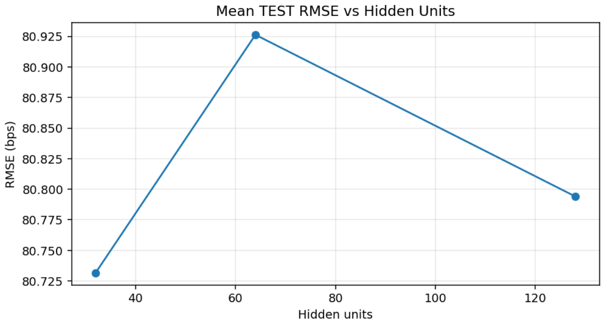


Figure B.4: *Mean test RMSE as a function of the number of hidden units. Increasing model capacity initially improves performance, but excessive width leads to mild overfitting.*

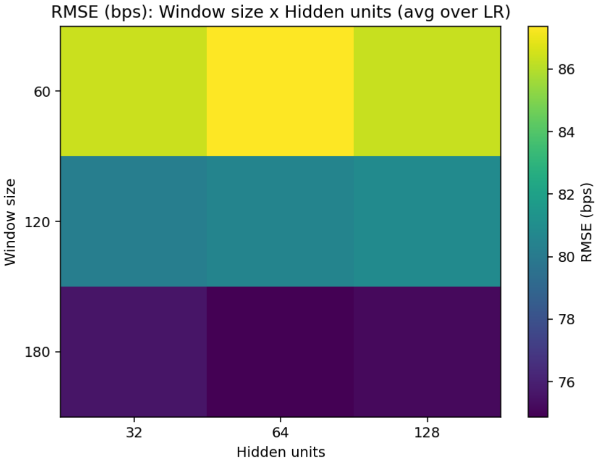


Figure B.5: *Heatmap of test RMSE as a function of window size and number of hidden units. The best performance is achieved by combining long look-back windows with moderate network width.*

C

HOW THE LSTM TRADING STRATEGY WORKS?

C.1 Trading Strategy LSTM

The proposed trading strategy is a systematic long–short approach applied to the SPDR Gold Shares ETF (GLD), combining return forecasts produced by a Long Short-Term Memory (LSTM) network with conditional volatility forecasts obtained from a GARCH model. The methodology explicitly separates the estimation of the conditional mean from that of the conditional variance, using each component for a distinct role within the trading process. While the LSTM is designed to capture weak, potentially nonlinear predictive information in gold returns, the GARCH model is used exclusively for risk measurement and position sizing.

Let P_t denote the adjusted closing price of GLD at date t. Daily log-returns are defined as

$$r_t = \log(P_t) - \log(P_{t-1}).$$

The prediction target of the LSTM is the one-step-ahead return r_{t+1}. At each date t, the input to the network consists of a rolling window of length L (here $L = 252$) of past observations,

$$X_t = \left[x_{t-L+1}, \ldots, x_t\right] \in \mathbb{R}^{L\times d},$$

where the feature vector x_t includes the current gold return and macro-financial innovations,

$$x_t = \left(r_t,\ \Delta r_t^{\text{real}},\ \Delta\pi_t^{\text{be}},\ \Delta\text{VIX}_t,\ \Delta\log(\text{DXY}_t)\right).$$

The LSTM produces an estimate of the conditional expectation of the next-day return,

$$\widehat{r}_{t+1} \approx \mathbb{E}[r_{t+1} \mid X_t].$$

To stabilise training, the target variable is standardised using moments estimated on the training sample,

$$\tilde{r}_{t+1} = \frac{r_{t+1} - \mu_r}{\sigma_r},$$

and the network outputs a prediction $\widehat{\tilde{r}}_{t+1}$, which is transformed back to the return scale via

$$\widehat{r}_{t+1} = \widehat{\tilde{r}}_{t+1}\sigma_r + \mu_r.$$

Because daily return forecasts are extremely noisy, the raw LSTM predictions are not used directly to determine trading positions. Instead, a smoothed predictive signal is constructed using a rolling average over H days (with $H = 5$),

$$\bar{r}_{t+1} = \frac{1}{H}\sum_{j=0}^{H-1}\widehat{r}_{t+1-j}.$$

The trading direction is then defined as

$$s_t = \operatorname{sign}(\bar{r}_{t+1}),$$

so that $s_t = +1$ corresponds to a long position, $s_t = -1$ to a short position, and $s_t = 0$ to a neutral stance.

Risk is controlled through volatility targeting based on conditional variance forecasts from a GARCH(1, 1) model with Student-t innovations. Let the return process be written as

$$r_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t,$$

where z_t follows a standardized Student-t distribution. The conditional variance evolves according to

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2.$$

The GARCH model is estimated on the training sample only, and one-step-ahead volatility forecasts $\widehat{\sigma}_t$ are generated for the test period.

The target portfolio weight is determined by combining the directional signal with volatility scaling,

$$w_t^\star = s_t\frac{\sigma^\star}{\widehat{\sigma}_t},$$

where $\sigma^\star$ denotes the target daily volatility, obtained by rescaling an annual volatility target. To prevent excessive leverage, weights are constrained such that

$$w_t = \max\bigl(-w_{\max},\ \min(w_t^\star,\ w_{\max})\bigr).$$

To reduce transaction costs and portfolio turnover, rebalancing is performed only every H trading days. Between rebalancing dates, the portfolio weight is held constant. Moreover, to avoid look-ahead bias, the executed weight is lagged by one day,

$$w_t^{\text{exec}} = w_{t-1}.$$

Let $R_t = (P_t - P_{t-1})/P_{t-1}$ denote the realised arithmetic return. Portfolio equity evolves according to

$$\text{Equity}_t = \text{Equity}_{t-1} + w_t^{\text{exec}}\,\text{Equity}_{t-1}R_t - \text{Cost}_t,$$

where transaction costs are proportional to dollar turnover,

$$\text{Cost}_t = c\left|w_t^{\text{exec}} - w_{t-1}^{\text{exec}}\right|\text{Equity}_{t-1},$$

with c denoting the per-dollar transaction cost rate.

This framework embeds machine-learning-based directional forecasts within a disciplined risk management structure. By allowing the LSTM to focus on predicting return direction and delegating exposure control to a GARCH-based volatility model, the strategy aims to translate weak predictive signals into economically meaningful, risk-adjusted portfolio performance.