electronics

MDPI

Article

Regime-Aware LightGBM for Stock Market Forecasting: A Validated Walk-Forward Framework with Statistical Rigor and Explainable AI Analysis

Antonio Pagliaro [1,2]

[1] Istituto Nazionale di Astrofisica-Istituto di Astrofisica Spaziale e Fisica Cosmica (INAF-IASF) Palermo, Via Ugo La Malfa 153, I-90146 Palermo, Italy; antonio.pagliaro@inaf.it
[2] Istituto Nazionale di Fisica Nucleare Sezione di Catania, Via Santa Sofia, 64, I-95123 Catania, Italy

Abstract

Can machine learning generate statistically validated alpha in equity markets while adapting to changing market conditions? This study addresses this question by proposing a regime-aware LightGBM framework conditioned on market regimes detected via a rolling Hidden Markov Model, eliminating look-ahead bias. Backtested on 51 NASDAQ-100 constituents (2015–2026), the strategy achieved a portfolio Sharpe ratio of 1.18 (95% CI: [0.53, 1.84]) and outperformed four baseline models. The key findings include the following: (i) cross-asset features (Bitcoin as a leading indicator) contribute the most predictive value; (ii) macroeconomic indicators outweigh traditional technical indicators for high-beta stocks; (iii) the model autonomously adapts its decision logic across regimes, shifting from mean reversion in bear markets to risk appetite monitoring in bull markets. While block bootstrap tests confirm statistical significance ($p < 0.001$), the Deflated Sharpe Ratio (0.69) does not reach formal significance after multiple testing correction—an honest finding we report transparently.

Keywords: LightGBM; stock market forecasting; explainable AI; SHAP; Hidden Markov Model; regime detection; walk-forward validation; block bootstrap; Deflated Sharpe Ratio; ablation study; probability calibration; volatility targeting; baseline comparison

1. Introduction

The prediction of stock market returns remains one of the most investigated problems in computational finance. The Efficient Market Hypothesis (EMH) [1] posits that asset prices fully reflect all available information, rendering systematic prediction impossible. However, a growing body of empirical evidence suggests that ML techniques can exploit non-linear patterns, temporary inefficiencies, and behavioral biases that elude traditional linear models such as ARIMA or GARCH [2–4].

Research Question. This study addresses three interconnected questions: (i) Can machine learning models generate statistically validated, risk-adjusted alpha in equity markets after accounting for multiple testing? (ii) How should predictive models adapt to changing market regimes (bull, sideways, bear) without introducing look-ahead bias? (iii) Which feature categories—technical, macroeconomic, or cross-asset—contribute the most predictive value, and does their importance vary across market conditions?

Main Contribution. We propose a regime-aware LightGBM framework that conditions predictions on market states detected via a rolling (online) Hidden Markov Model,

Academic Editor: Ping-Feng Pai

Received: 9 March 2026
Revised: 17 March 2026
Accepted: 21 March 2026
Published: 23 March 2026

combined with rigorous walk-forward validation and comprehensive statistical testing including the Deflated Sharpe Ratio for multiple testing correction. Unlike prior work that relies on static train/test splits and reports only point estimates, our framework provides confidence intervals, ablation-based component attribution, and SHAP-based explainability across regimes.

In our previous work [5], we demonstrated that the Extra Trees classifier achieves 86.1% accuracy in predicting significant price changes over a 10-day horizon, outperforming Random Forest on a universe of 120 stocks. A subsequent analysis [6] extended these findings by critically reassessing ML predictive models in the context of Big Data. While these studies established the potential of tree-based ensemble methods, they relied on a single-model architecture and a static train/test split, and did not account for the non-stationary nature of financial markets.

Financial markets are characterized by distinct regimes—bull, sideways, and bear—each exhibiting fundamentally different statistical properties in terms of return distributions, volatility clustering, and cross-asset correlations [7,8]. A model trained predominantly on bull market data may perform poorly during market stress, and vice versa. This observation motivates a *regime-aware* approach, where both the feature engineering and the trading strategy adapt to the current market state.

A critical gap in the financial ML literature is the absence of rigorous statistical validation. Many studies report point estimates of performance metrics (Sharpe ratio, alpha) without confidence intervals, p-values, or corrections for multiple testing [9,10]. Furthermore, few studies conduct systematic ablation experiments to quantify the marginal contribution of individual system components, leaving unclear which elements actually drive performance.

The present study advances the state of the art in four key directions:

1. **LightGBM Classifier with Rich Feature Engineering.** We employ a LightGBM gradient boosting classifier operating on 63 normalized features spanning technical, macroeconomic, cross-asset (BTC), and market interaction categories. An ablation study demonstrates that cross-asset features contribute the most predictive value.
2. **Rolling Hidden Markov Model Regime Detection.** We introduce an online (rolling) Gaussian HMM that is refitted every 63 trading days using only past data, completely eliminating look-ahead bias. At each time step t, the HMM is trained exclusively on data from $[0, t]$, ensuring that regime labels reflect only information available at the time of prediction.
3. **Rigorous Walk-Forward Validation.** We replace the static chronological split with a walk-forward cross-validation protocol employing 100 expanding-window folds, 10-day purge windows (equal to the prediction horizon), and per-fold scaler fitting to prevent any form of data leakage.
4. **Comprehensive Statistical Validation.** We provide block bootstrap confidence intervals for the Sharpe ratio (10,000 resamples with block size $b = 20$, preserving serial dependence), the Deflated Sharpe Ratio [9] to correct for the full strategy search space ($N \approx 250$ trials), Lo's autocorrelation-adjusted Sharpe SE [11], probability calibration analysis via Brier score and Expected Calibration Error (ECE), a seven-variant ablation study, comparison against four baseline models (XGBoost, Logistic Regression, SMA crossover, time-series momentum), and multi-dimensional sensitivity analysis across transaction costs, prediction horizons, VIX thresholds, and sub-periods.

The system is evaluated as a practical swing trading strategy on 51 NASDAQ-100 constituents over a 10-year period (2015–2026), with realistic transaction costs (0.1% commission + 0.05% slippage) and VIX-based emergency cutoffs.

2. Related Work

2.1. Machine Learning in Financial Forecasting

The application of ML to stock price prediction has evolved from simple linear models to sophisticated deep learning architectures. Gu et al. [2] conducted a comprehensive comparison of ML methods for empirical asset pricing, demonstrating that neural networks and tree-based models consistently outperform linear benchmarks. Fischer and Krauss [3] showed that LSTM networks achieve superior performance in financial time series compared to Random Forests and Logistic Regression, particularly when capturing long-range temporal dependencies. Khaidem et al. [12] validated the effectiveness of Random Forest classifiers for stock direction prediction. Among gradient boosting frameworks, LightGBM [13] has emerged as a leading approach for tabular data due to its histogram-based split finding and efficient handling of high-dimensional features.

Recent advances in deep learning have introduced novel architectures for financial forecasting. Zhang et al. [14] demonstrated that Generative Adversarial Networks (GANs) can effectively capture complex patterns in stock market data. Sezer et al. [15] provided a comprehensive review of deep learning methods for financial time series, highlighting the trade-off between model complexity and interpretability. Kumar et al. [16] proposed a PSO-ELM approach combined with Ichimoku Cloud indicators for gold market prediction, demonstrating that technical indicators effectively capture market sentiment in precious metals—a finding relevant to our use of gold (GLD) as a cross-asset feature.

2.2. Ensemble Methods and Model Combination

Ensemble learning, which combines multiple models to improve prediction robustness, has been widely adopted in financial applications. Pagliaro [5] demonstrated the superiority of Extra Trees over Random Forest for stock forecasting, achieving 86.1% accuracy on a 120-stock universe. Stacking approaches that combine heterogeneous model families via a meta-learner have shown promise in reducing model-specific biases [17].

2.3. Regime Detection in Financial Markets

The identification of market regimes via Hidden Markov Models was pioneered by Hamilton [7] for business cycle analysis and subsequently applied to financial markets by Ang and Bekaert [8]. Regime-switching models have been shown to improve out-of-sample forecasting by conditioning predictions on the identified market state, allowing different model parameters for structurally different market environments. A critical distinction in the literature is made between *offline* HMM fitting (using the full time series, which introduces look-ahead bias) and *online* or rolling HMM fitting, where the model is trained incrementally using only past data [7].

2.4. Deep Learning Approaches

Deep learning architectures have shown promise for financial time series, though with important caveats. LSTM and GRU networks excel at capturing long-range temporal dependencies [3], while attention-based Transformers offer improved parallelization and interpretability through attention weights [18]. N-BEATS [19] introduced a purely feedforward architecture that achieves state-of-the-art performance on time series benchmarks. However, Sezer et al. [15] note that deep learning models often underperform gradient boosting methods on tabular financial data with hand-crafted features, where the inductive biases of tree-based models (feature interactions, handling of mixed data types) prove advantageous. Furthermore, LSTM/GRU models are notoriously difficult to calibrate and are prone to overfitting on noisy financial data without extensive hyperparameter tuning—a concern particularly acute in walk-forward validation where training data is limited in early folds.

2.5. Technical and Momentum Baselines

Moving average crossover strategies (e.g., SMA 50/200) remain among the most widely used technical trading rules [20]. Time-series momentum strategies, formalized by Moskowitz et al. [21], exploit the empirical observation that assets with positive past returns tend to continue appreciating over horizons of 1–12 months. These non-ML baselines provide essential benchmarks: if a sophisticated ML model cannot outperform a simple SMA crossover or momentum rule under identical evaluation conditions, the ML complexity adds no practical value.

2.6. Cross-Asset Information and Bitcoin as a Leading Indicator

The literature increasingly recognizes that cross-asset correlations contain predictive information for equity returns. Bitcoin, operating in 24/7 markets with high speculative participation, has been shown to lead equity market movements, particularly for technology stocks with high retail ownership [22]. This leading relationship may arise from two mechanisms: (i) BTC serves as a barometer for speculative risk appetite, reflecting liquidity flows that subsequently affect high-beta equities, or (ii) the overlap in investor base between cryptocurrencies and technology stocks creates direct information transmission. Our framework explicitly tests both hypotheses through ablation analysis and SHAP feature importance across market regimes.

2.7. Statistical Validation in Quantitative Finance

The statistical evaluation of trading strategies has received increasing attention. Bailey and López de Prado [9] introduced the Deflated Sharpe Ratio (DSR), which adjusts the Sharpe ratio for the number of strategy variants tested, addressing the multiple testing problem inherent in strategy development. Harvey et al. [10] demonstrated that many published "anomalies" in financial economics fail to survive proper multiple testing correction. Lo [11] provided the standard error formula for the Sharpe ratio under non-normal returns, accounting for skewness and kurtosis. Politis and Romano [23] introduced the stationary block bootstrap for time series data, which preserves serial dependence and produces more accurate confidence intervals than standard i.i.d. resampling.

3. Materials and Methods

3.1. Data

We analyzed daily OHLCV (Open, High, Low, Close, Volume) data for 51 NASDAQ-100 constituents covering the period from 1 January 2015 to 8 February 2026 (approximately 2791 trading days per ticker). Data were obtained via the `yfinance` API.

In addition to individual stock data, we collected 10 macroeconomic and cross-asset indicators, summarized in Table 1.

Table 1. Macroeconomic and cross-asset data sources.

Indicator	Ticker	Role
VIX (Short-Term Volatility)	^VIX	Fear gauge, emergency cutoff
VIX3M (Medium-Term Volatility)	^VIX3M	Term structure analysis
S&P 500 ETF	SPY	Market benchmark, beta computation
Bitcoin	BTC-USD	Leading indicator for crypto-proxy stocks
High Yield Bonds	HYG	Credit risk/risk-on appetite
Investment Grade Bonds	LQD	Credit spread reference
20-Year Treasury Bonds	TLT	Long-term interest rate proxy
1–3 Year Treasury Bonds	SHY	Short-term rate, yield curve proxy
Gold	GLD	Safe-haven demand
US Dollar Index	UUP	Global liquidity conditions

The caret symbol (^) is the Yahoo Finance ticker prefix for index data.

3.2. Feature Engineering

A total of 63 normalized features were engineered across five categories, following the established technical analysis literature [20,24]. Crucially, *no absolute price levels* were used as features; all indicators were normalized relative to the price or expressed as ratios and percentages.

3.2.1. Technical Indicators (30+ Features)

- **Trend:** SMA ratios at 5 periods (5, 10, 20, 50, 200 days), EMA ratios (12, 26 days), normalized MACD (line, signal, histogram), ADX with directional indicators (+DI, –DI), Aroon oscillator, and Ichimoku Cloud components (Tenkan, Kijun ratios, cloud width).
- **Momentum:** RSI(14) with Wilder smoothing, Stochastic %K/%D(14,3), Williams %R, CCI(20), Rate of Change at 3 periods (5, 10, 20 days), and True Strength Index (TSI).
- **Volatility:** Bollinger Band width and %B position, ATR(14) normalized, and Keltner Channel width.
- **Volume:** OBV slope (10-day linear regression), Chaikin Money Flow (CMF-20), volume ratio, and VWAP deviation.

3.2.2. Price-Derived Features

Price-derived features include multi-period returns (1, 5, 10, 20 days), lagged returns ($t-1$, $t-2$, $t-5$), consecutive up/down day counters, drawdown from peak, and distance from recent highs/lows.

3.2.3. Cross-Asset Features (BTC Leading Indicator)

Bitcoin features are included based on two empirically supported hypotheses [22]. First, for *crypto-correlated stocks* (e.g., MSTR, COIN, MARA), BTC serves as a direct leading indicator due to business model exposure to cryptocurrency prices. Second, and more broadly, for *high-beta technology stocks* (e.g., AMD, NVDA, TSLA), BTC serves as a *risk appetite barometer*: because cryptocurrency markets operate 24/7 and attract speculative capital, BTC price movements often reflect shifts in global liquidity conditions and risk sentiment before these manifest in equity markets during regular trading hours. This cross-asset information transmission operates through two channels: (i) overlapping investor bases between crypto and tech stocks, and (ii) BTC's role as a leading indicator for speculative capital flows that subsequently affect high-beta equities.

We computed BTC returns at multiple horizons (5, 10, 20 days), BTC RSI, BTC momentum, BTC volatility, and the rolling 20-day correlation between the target stock and BTC. The ablation study (Section 4.5) directly tests whether BTC features provide value beyond other feature categories. **Timestamp alignment:** BTC–USD daily data from `yfinance` uses the UTC 00:00 close, which corresponds to 7:00 P.M. ET—after the US equity market close at 4:00 P.M. ET. To prevent look-ahead bias, all BTC features at day t are computed using BTC data up to and including day $t-1$ (i.e., the most recent BTC close available before the equity market opens on day t). This ensures that no intraday BTC information leaks into the equity prediction.

Feature redundancy analysis: To assess potential redundancy among BTC features, we computed pairwise correlations across the 7 BTC-related variables. The average inter-feature correlation is 0.42, with the highest correlation (0.78) between BTC 5-day and 10-day returns. While moderate redundancy exists, LightGBM's feature subsampling (70% per tree) and L1/L2 regularization naturally handle correlated features by selecting the most informative subset at each split.

3.2.4. Market Interaction Features (SPY)

The market interaction features include rolling 20-day correlation with SPY, 60-day beta versus SPY, SPY returns at multiple horizons (5, 10, 20 days), and relative strength versus the benchmark.

3.2.5. Macroeconomic Features

The macroeconomic feature set was designed to capture the primary drivers of cross-sectional equity returns identified in the asset pricing literature [1,8]:

- **VIX Term Structure:** VIX level, VIX percentage change (5, 10 days), VIX3M level, and the term structure ratio $(VIX3M - VIX)/VIX$, where values below zero indicate market panic (backwardation). The VIX term structure is a well-documented predictor of future volatility and equity returns [8].
- **Credit Spreads:** HYG/LQD ratio (credit risk proxy), 5-day and 20-day percentage changes, and mean reversion ratio. Credit spreads capture systemic risk and have been shown to predict equity returns, particularly during stress periods.
- **Yield Curve:** TLT/SHY ratio (curve steepness proxy) and its trend relative to the 50-day moving average. The yield curve slope is a leading indicator of economic activity and has historically predicted equity market returns with a lag of 6–12 months.
- **Safe Haven Flows (Gold/Equity Ratio):** Gold (GLD) percentage change, US Dollar Index (UUP) percentage change, and crucially, the GLD/SPY ratio (gold-to-equity ratio). This ratio captures "flight to safety" dynamics: when investors rotate from equities to gold, the ratio rises, often signaling capitulation that precedes equity market rebounds [16]. The SHAP analysis (Section 4.11) confirms that the gold/equity ratio is among the top 3 global predictors, particularly effective in identifying contrarian buying opportunities after panic-driven selloffs.

Rationale for indicator selection: We prioritized indicators with (i) established theoretical relationships to equity returns in the asset pricing literature, (ii) daily availability without revision (avoiding look-ahead bias from revised economic data), and (iii) complementary information content (VIX for volatility, credit spreads for default risk, yield curve for monetary policy, gold/dollar for safe-haven flows). Alternative indicators such as PMI, jobless claims, or sentiment surveys were excluded due to lower frequency (weekly/monthly) or revision risk.

All features were standardized using the `RobustScaler` from scikit-learn, which uses the interquartile range (IQR) rather than mean/variance. This choice is motivated by the heavy-tailed distribution of financial returns: outliers (market crashes, earnings surprises) can severely distort mean-based scaling, causing most observations to cluster near zero after standardization. The IQR-based approach preserves the relative ranking of observations while being robust to extreme values, which is particularly important for tree-based models like LightGBM that make decisions based on feature thresholds. Note that RobustScaler is a monotonic transformation that does not affect the non-linear relationships captured by LightGBM's tree splits—it merely rescales features to a common range, allowing the model to assign comparable importance across features with different native scales.

3.3. Target Definition

We formulated the prediction task as a binary classification problem. Let P_t denote the closing price at day t and $h = 10$ the prediction horizon in trading days. The forward return is

$$R_{t+h} = \frac{P_{t+h} - P_t}{P_t} \quad (1)$$

The binary target variable is defined as

$$Y_t = \begin{cases} 1 \quad (\text{UP}) & \text{if } R_{t+h} > 0 \\ 0 \quad (\text{DOWN}) & \text{otherwise} \end{cases} \tag{2}$$

The 10-day horizon was selected as a balance between short-term noise and long-term mean reversion, consistent with the swing trading time frame adopted in our previous work [5]. A sensitivity analysis across horizons of 5, 10, and 15 days is presented in Section 4.8.

3.4. Model Architecture: LightGBM Classifier

LightGBM [13] is a gradient boosting framework that uses histogram-based split finding for efficient training on high-dimensional tabular data. We chose LightGBM as the primary classifier based on its established superiority for tabular financial data [2,13] and its compatibility with SHAP TreeExplainer for exact feature attribution.

Hyperparameter Selection

The model hyperparameters were selected through a two-stage process. First, we adopted baseline values from the financial ML literature [2,13]: 800 boosting rounds (sufficient for convergence with early stopping), maximum depth of 6 (balancing expressiveness and overfitting risk), and learning rate of 0.03 (standard for gradient boosting with many rounds). Second, regularization parameters were tuned via 5-fold time-series cross-validation on a held-out development set (2015–2017, not used in final evaluation): L1 and L2 regularization ($\alpha = \lambda = 0.05$), row subsampling (80%), and column subsampling (70%) were selected to minimize validation log-loss while maintaining an out-of-fold accuracy above 53%. Early stopping after 50 rounds without improvement prevents overfitting to training data within each walk-forward fold. The final configuration uses 31 leaves per tree, consistent with the $2^{\text{depth}} - 1$ heuristic for leaf-wise growth. The output is a probability $P(\text{UP}) \in [0, 1]$.

3.5. Regime Detection via Rolling Hidden Markov Model

Market regimes are identified using a Gaussian HMM with K hidden states. To **eliminate look-ahead bias**, we employ a rolling (online) fitting procedure rather than training on the full time series.

3.5.1. Model Selection: Number of States

The choice of $K = 3$ states is justified via information-theoretic model selection. We fit Gaussian HMMs with $K \in \{2, 3, 4, 5, 6\}$ states on the full SPY feature matrix (2015–2026) and compute the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) for each. Table 2 presents the results.

Table 2. HMM model selection via BIC/AIC on SPY (2015–2026; $T = 2771$ trading days; $d = 4$ features).

K	Log-Lik	Params	BIC	AIC
2	19,074	31	−37,902	−38,086
3	20,273	50	−40,149	−40,446
4	20,941	71	−41,319	−41,741
5	20,786	94	−40,826	−41,384
6	22,066	119	**−43,189**	**−43,894**

Bold values indicate the optimal (lowest) BIC and AIC scores.

Formally, BIC and AIC both favour $K = 6$, indicating that the SPY feature space supports fine-grained regime differentiation. However, we adopt $K = 3$ for three reasons: **(i)** the improvement from $K = 3$ to $K = 4$ is modest ($\Delta\text{BIC} = -1170$, or 2.9%) compared to the large jump from $K = 2$ to $K = 3$ ($\Delta\text{BIC} = -2247$, or 5.9%), suggesting diminishing returns beyond three states; **(ii)** $K = 5$ exhibits a *decrease* in log-likelihood relative to $K = 4$ (from 20,941 to 20,786), indicating convergence instability at higher K; and **(iii)** $K = 3$ aligns with the well-established economic interpretation of three distinct market regimes (bull, sideways, bear) [7,8], facilitating interpretable SHAP analysis across regimes (Section 4.11). We acknowledge that formal information criteria prefer $K \geq 4$ and that this choice involves a trade-off between statistical fit and economic interpretability.

Sensitivity to regime count: To assess robustness, we conducted a supplementary analysis with $K = 4$ states on a subset of 10 stocks. The $K = 4$ configuration produced similar aggregate Sharpe ratios (within ± 0.03) but with less stable regime assignments: the fourth state captured only 8% of observations and exhibited high transition rates, suggesting it represents transient market conditions rather than a stable regime. The $K = 3$ configuration provided more stable regime assignments that align with practitioners' intuitive understanding of market phases, which is valuable for interpretability and actionable trading decisions.

3.5.2. HMM Features

The HMM is trained on four features derived from SPY (S&P 500 ETF) data:

1. 20-day rolling returns;
2. 20-day annualized volatility ($\sigma \cdot \sqrt{252}$);
3. VIX level (normalized by 100);
4. Market breadth proxy: fraction of positive daily returns in a 20-day window.

3.5.3. Rolling Fitting Procedure

At each time step $t \geq t_{\min}$ (where $t_{\min} = 252$ trading days, approximately one year), the following occurs:

1. The HMM is fitted exclusively on the feature matrix $\mathbf{X}_{0:t}$ (data from the start to day t);
2. To reduce computational cost, the model is refitted every $\Delta = 63$ trading days (approximately one quarter), and the most recent fitted model is cached between refits;
3. The regime at time t is decoded using the Viterbi algorithm on a 120-day context window $\mathbf{X}_{t-120:t}$;
4. States are mapped to regimes by sorting mean returns: the state with the lowest mean return is bear (0), middle is sideways (1), highest is bull (2).

Regime characterization by volatility: The three regimes exhibit distinct volatility profiles. Based on the rolling HMM output over the 2015–2026 period, the regimes are characterized as follows:

- **Bull regime:** Mean annualized volatility $\sigma \approx 14\%$, VIX typically <20, positive 20-day returns;
- **Sideways regime:** Mean annualized volatility $\sigma \approx 18\%$, VIX typically 20–30, near-zero 20-day returns;
- **Bear regime:** Mean annualized volatility $\sigma > 25\%$, VIX typically >30, negative 20-day returns.

These characterizations are emergent properties of the HMM clustering, not predefined thresholds. The VIX > 40 emergency override (Section 3.7) operates independently of the HMM regime and represents an additional tail-risk protection layer.

This procedure ensures that at no point does the regime label at time t depend on data from $t+1, t+2, \ldots$, eliminating the look-ahead bias present in traditional offline HMM implementations. The rolling approach produces more conservative regime labels with approximately 56.6% agreement with the offline (full-data) HMM, reflecting the genuine information loss from causal-only fitting.

The HMM uses a full covariance matrix with 100 EM iterations. Over the walk-forward out-of-fold period, the rolling HMM identified the following approximate regime distribution: **bull**—672 days (32.5%); **sideways**—760 days (36.8%); **bear**—635 days (30.7%). The higher proportion of bear days compared to the offline HMM (which assigns only 13.5% to bear) reflects the conservative, uncertain nature of causal-only regime estimation.

3.6. Walk-Forward Validation with Purge and Embargo

To prevent data leakage and simulate a realistic trading environment, we employed expanding-window walk-forward cross-validation with the following parameters:

- **Number of folds:** 100;
- **Minimum training window:** 504 days (≈2 years);
- **Test window:** 21 days (≈1 month) per fold;
- **Step size:** 21 days between consecutive folds;
- **Purge window:** 10 days (equal to the prediction horizon h);
- **Embargo:** 0 days.

Three critical anti-leakage mechanisms are enforced at each fold:

1. The feature scaler (RobustScaler) is fitted *exclusively* on the training data and used to transform the test data;
2. The target variable uses forward-shifted returns (shift $-h$), computed *before* the train/test split;
3. The purge window of h days between the training and test sets ensures that no training sample has a target label that overlaps temporally with the test period.

Out-of-fold predictions are accumulated across all folds and used for both model evaluation and backtest signal generation.

3.7. Swing Trading Backtest

The system generates swing trading signals based on the model's estimated probability $P(\text{UP})$, with the following execution logic:

3.7.1. Signal Generation with Hysteresis

To reduce whipsaw trades, $P(\text{UP})$ is smoothed with a 5-day exponential moving average. Entry and exit decisions follow a hysteresis mechanism:

- **Entry (BUY):** Smoothed $P(\text{UP}) > \theta_{\text{entry}}$ and 3-day $P(\text{UP})$ momentum is positive;
- **Normal Exit:** Smoothed $P(\text{UP}) < \theta_{\text{exit}}$ and minimum 30-day holding period is satisfied;
- **Trailing Stop:** Gain from entry exceeds 15%, drop from peak exceeds 10%, and model confidence is declining;
- **Emergency Exit:** Loss exceeds 20% and raw $P(\text{UP}) < 0.45$.

Thresholds are adaptive after a 63-day warmup: θ_{entry} is set to the 30th percentile of historical $P(\text{UP})$ values (capped at 0.55), and θ_{exit} to the 10th percentile (capped at 0.50).

3.7.2. VIX Emergency Override

When VIX > 40, the system forces a cash position regardless of the model's signal, acting as a circuit breaker during extreme market stress.

3.7.3. Transaction Costs

All backtests include realistic costs: 0.1% commission and 0.05% slippage per trade, applied to the portfolio value at each trade execution.

3.8. Statistical Validation Framework

To address the well-documented problem of overstated performance in quantitative finance [9,10], we employ the following three complementary statistical tests.

3.8.1. Block Bootstrap Sharpe Ratio Confidence Intervals

The Sharpe ratio $\hat{SR}$ is computed as $\hat{SR} = \bar{r}/\hat{\sigma}_r \times \sqrt{252}$, where $\bar{r}$ and $\hat{\sigma}_r$ are the mean and standard deviation of daily excess returns. Because daily returns exhibit serial dependence (autocorrelation, volatility clustering), standard i.i.d. resampling would understate the true sampling variability of $\hat{SR}$. We therefore employ a **non-overlapping block bootstrap** [23] with block size $b = 20$ trading days (approximately one calendar month), which preserves the within-block dependence structure. We draw $B = 10{,}000$ block bootstrap resamples of the return series, compute the annualized Sharpe ratio for each resample, and report the 2.5th and 97.5th percentiles as the 95% confidence interval. The one-sided p-value for $H_0 : SR \leq 0$ is the fraction of bootstrap resamples with $SR \leq 0$. Additionally, we report the autocorrelation-adjusted standard error of $\hat{SR}$ following Lo [11], which accounts for serial correlation up to lag 5.

3.8.2. Deflated Sharpe Ratio

The Deflated Sharpe Ratio (DSR) [9] corrects the observed Sharpe ratio for the number of strategy variants tested. Given N independent trials and the observed $\hat{SR}$, the DSR computes the probability that the observed Sharpe exceeds the expected maximum Sharpe under the null hypothesis (all strategies have zero expected returns). Critically, N must reflect the **entire strategy search space**—not just the variants explicitly reported—including all hyperparameter configurations, threshold grids, prediction horizons, and ablation variants explored during development. In our case, $N \approx 250$ (7 ablation variants $\times$ 25 threshold combinations $\times$ 3 horizons, plus VIX and cost sensitivity configurations). The standard error of the Sharpe estimator follows Lo [11] as follows:

$$\mathrm{SE}(\hat{SR}) = \sqrt{\frac{1 + \frac{\hat{\gamma}_3}{2}\hat{SR} - \frac{\hat{\gamma}_4 - 3}{4}\hat{SR}^2}{T}} \tag{3}$$

where $\hat{\gamma}_3$ is the skewness, $\hat{\gamma}_4$ the kurtosis, and T the number of observations. The expected maximum Sharpe under the null for N trials uses the Euler–Mascheroni correction: $E[\max SR_0] \approx (1 - \gamma_E)\Phi^{-1}(1 - 1/N) + \gamma_E \Phi^{-1}(1 - 1/(Ne))$, where $\gamma_E \approx 0.5772$. A DSR value exceeding 0.05 provides evidence that the strategy's Sharpe ratio is not solely attributable to multiple testing.

3.8.3. Bootstrap Alpha Test

The risk-adjusted alpha (Jensen's alpha: the intercept of regressing daily strategy excess returns on daily benchmark excess returns) is tested via a block bootstrap procedure ($B = 10{,}000$ resamples; block size $b = 20$), where both strategy and benchmark returns are resampled jointly to preserve their cross-sectional and serial dependence. The annualized alpha and its 95% CI are reported. Note that Jensen's alpha can be positive even when cumulative returns lag the benchmark if the strategy takes less systematic risk (lower beta) by spending time in cash.

3.9. Probability Calibration Assessment

The model's predicted probabilities $P(\text{UP})$ are evaluated for calibration quality using the following:

- **Brier Score:** $\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$, where p_i is the predicted probability and $y_i \in \{0, 1\}$ the true label. BS $= 0$ indicates perfect calibration; BS $= 0.25$ corresponds to a random coin flip on a balanced binary problem.
- **Expected Calibration Error (ECE):** $\text{ECE} = \sum_{b=1}^{B}\frac{n_b}{N}|\bar{p}_b - \bar{y}_b|$, where samples are binned into $B = 10$ uniform bins and $\bar{p}_b$, $\bar{y}_b$ are the average predicted and actual probabilities in bin b. ECE = 0 indicates perfect calibration.
- **Reliability Diagram:** A visual representation of calibration, plotting the observed fraction of positive outcomes against the predicted probability for each bin. A perfectly calibrated model lies on the diagonal.

When calibration is found to be deficient, we apply **isotonic calibration** [25] as a post hoc correction. Within each walk-forward fold, an isotonic regression model is fitted on the validation set's predicted probabilities and true labels, then applied to transform the test set predictions. This non-parametric approach maps model scores to calibrated probabilities while preserving rank ordering.

3.10. Ablation Study Design

To quantify the marginal contribution of each system component, we conducted a seven-variant ablation study (Table 3):

Table 3. Ablation study variants.

Variant	Description	Component Removed
A	Full pipeline (baseline)	None
B	No HMM regime feature	Rolling HMM regime label
C	No post-processing	EMA smoothing, hysteresis, trailing stop, min-hold
D	No VIX override	VIX $>$ 40 emergency exit
E	No macro features	All `macro_*` and `vix_*` features
F	No BTC features	All `btc_*` cross-asset features
G	Simple threshold	All post-processing + VIX override

* denotes wildcard matching all features with this prefix.

Each variant is evaluated using the same walk-forward protocol (100 folds) and backtest engine, with the only difference being the specified component removal. This ensures that performance differences are attributable solely to the ablated component.

3.11. Sensitivity Analysis Design

We assess the robustness of results across five dimensions:

1. **Transaction costs:** 0, 5, 10, 20, and 50 basis point total cost (commission + slippage);
2. **Entry/exit thresholds:** Grid search over entry $\in \{0.40, 0.45, 0.50, 0.55, 0.60\}$ and exit $\in \{0.25, 0.30, 0.35, 0.40, 0.45\}$;
3. **Prediction horizon:** $h \in \{5, 10, 15\}$ trading days;
4. **VIX emergency threshold:** $\{30, 35, 40, 50, \text{OFF}\}$;

5. **Sub-period analysis:** 2015–2019 (pre-COVID), 2020–2022 (pandemic/bear), 2023–2025 (recovery/AI boom).

3.12. Baseline Models

To ensure the LightGBM framework provides genuine value beyond simpler alternatives, we compare it against the following four baseline models under the *identical* walk-forward protocol:

1. **XGBoost Classifier [26]:** Same hyperparameter structure (800 rounds, depth 6, learning rate 0.03) and same 63-feature input as LightGBM. Tests whether the choice of boosting framework matters.
2. **Logistic Regression:** The simplest possible ML baseline (linear decision boundary). Tests whether the non-linearity of tree-based models contributes predictive value.
3. **SMA 50/200 Crossover:** A classical technical trading rule that generates $P(\text{UP}) = 1$ when the 50-day SMA exceeds the 200-day SMA. No ML involved.
4. **Time-Series Momentum [21]:** The 12-month return minus 1-month return. Generates $P(\text{UP}) = 1$ when momentum is positive. No ML involved.

All baselines are evaluated using the same swing trading backtest engine with identical transaction costs and execution logic.

Note on deep learning baselines: We do not include LSTM or GRU baselines in the primary comparison for three reasons. First, the literature indicates that deep learning models often underperform gradient boosting on tabular financial data with engineered features [2,15]. Second, LSTM/GRU models require substantially more hyperparameter tuning (hidden units, layers, dropout, sequence length) and are prone to overfitting in walk-forward settings where early folds have limited training data (504 days $\approx$ 2 years). Third, the primary research question concerns the value of regime conditioning and cross-asset features, which can be tested with tree-based models that offer exact SHAP attribution. Nevertheless, we acknowledge that comparing against deep learning architectures (LSTM, Temporal Fusion Transformers) under identical walk-forward protocols remains as valuable future work to assess whether recurrent models capture long-term dependencies that LightGBM may miss.

4. Results

This section presents the empirical findings organized by importance. The key results are as follows: (i) the ablation study demonstrates that cross-asset features contribute the most predictive value (Section 4.5); (ii) statistical significance analysis shows positive but not DSR-significant Sharpe ratios (Section 4.4); and (iii) SHAP explainability reveals regime-dependent decision logic (Section 4.11). Readers primarily interested in the framework's contribution may focus on these three sections. Supporting analyses (overall performance, regime-specific results, sensitivity analysis) provide additional context.

4.1. Overall Performance: 51-Stock Universe

The system was backtested on 51 NASDAQ-100 constituents over approximately 2791 trading days. Table 4 presents the top 15 stocks ranked by strategy cumulative return.

Key summary statistics across the full 51-stock universe:

- **9 out of 51 stocks** (17.6%) generated positive alpha over Buy-and-Hold;
- **12 out of 51 stocks** (23.5%) achieved win rates $\geq$ 70%;
- The median time in market was 94.2%, indicating the system favors staying invested and exits only during detected danger periods;
- The median number of trades per stock was 22 over the full period, consistent with a swing trading frequency of approximately 2 trades per year.

Table 4. Top 15 stocks by strategy cumulative return (2015–2026). An asterisk (*) denotes stocks where the strategy outperformed Buy-and-Hold.

#	Ticker	Strat %	B&H %	Alpha %	Sharpe	WR %	MaxDD %
1	SMCI *	2102.9	1493.0	+610.0	0.816	65.5	−84.5
2	NVDA	1920.5	3529.8	−1609.3	0.906	70.8	−64.5
3	MSTR *	1479.6	1133.3	+346.4	0.754	70.6	−86.9
4	ENPH	1446.7	2479.1	−1032.3	0.753	66.7	−92.5
5	AMD	1187.2	2262.9	−1075.6	0.759	66.7	−66.4
6	TSLA	977.9	1931.8	−953.9	0.701	65.2	−75.5
7	KLAC	934.3	1289.3	−354.9	0.788	69.2	−44.4
8	MU	712.5	801.9	−89.5	0.682	76.0	−60.9
9	AVGO	658.8	1112.7	−454.0	0.714	66.7	−46.7
10	MELI	451.5	789.4	−337.9	0.566	63.0	−71.4
11	PANW	446.6	634.4	−187.8	0.624	72.4	−42.3
12	COST	422.0	510.4	−88.4	0.831	72.2	−28.8
13	MSFT	410.0	460.2	−50.2	0.702	59.1	−40.3
14	NOW	373.5	426.7	−53.2	0.556	64.0	−55.1
15	AAPL	372.3	486.9	−114.6	0.634	69.0	−35.5

4.2. Out-of-Sample Validation: S&P 500 Universe

To test the generalizability of the framework beyond the NASDAQ-100, we extended the scan to 199 S&P 500 constituents using the identical walk-forward pipeline (rolling HMM, 100-fold validation, no look-ahead bias). The objective was to determine whether the strategy's alpha-generating capacity is confined to high-beta technology stocks or extends to the broader equity universe.

Of the 199 tickers scanned, 33 (16.6%) generated positive alpha over Buy-and-Hold—a rate remarkably consistent with the 9/51 (17.6%) observed in the NASDAQ-100 universe (Section 4.1). However, applying the full quality filter (alpha $>$ 0; Sharpe $>$ 0.5; win rate $>$ 60%) reduced the candidate set to only five stocks (2.5%), as listed in Table 5.

Table 5. S&P 500 scan: Top 5 candidates passing all quality filters (alpha $>$ 0; Sharpe $>$ 0.5; WR $>$ 60%).

Ticker	Sector	Alpha %	Sharpe	WR %	MaxDD %
CTAS	Industrials (Uniform Rental)	+181.0	0.885	68.2	−26.6
JPM	Financials (Banking)	+94.3	0.712	63.5	−38.1
DE	Industrials (Agriculture Equip)	+72.8	0.654	61.9	−41.2
ISRG	Healthcare (Surgical Robotics)	+68.5	0.621	62.7	−35.8
TMO	Healthcare (Life Sciences)	+53.2	0.578	61.1	−39.4

The standout candidate is CTAS (Cintas Corporation), an industrial uniform rental company with no cryptocurrency correlation whatsoever. Its +181% alpha, Sharpe ratio of 0.885 (the highest across both scans), win rate of 68.2%, and maximum drawdown of only −26.6% represent the best risk-adjusted profile in the entire combined universe. This result provides strong evidence that the strategy captures alpha from *regime sensitivity* rather than from BTC beta alone: CTAS exhibits clear regime-dependent behavior (defensive positioning during bear markets, consistent outperformance during recoveries) that the rolling HMM successfully detects.

Combining the two scans, the framework identifies alpha-positive stocks at a consistent rate of approximately 17% across both NASDAQ-100 (9/51) and S&P 500 (33/199) universes. The quality-filtered rate (2.5–5.9%) confirms that generating risk-adjusted alpha with favorable drawdown profiles remains rare, but the strategy reliably identifies such opportunities across diverse market sectors.

4.3. Regime-Specific Performance

The HMM regime conditioning significantly impacted strategy performance across different market environments. Tables 6 and 7 present the regime-specific rankings.

During bull periods, the system maintains near-full market exposure (90–99%), capturing the majority of the upside. The high Sharpe ratios (1.0–2.2) during bull regimes indicate that the system contributes risk-adjusted value even when passive strategies also perform well. Notably, only 5 out of the 10 top bull regime performers generated positive alpha over Buy-and-Hold (SMCI, MSTR, GOOGL, MELI, MRVL), reflecting the challenge of outperforming a fully invested benchmark during strong bull periods.

Table 6. Top 10 performers during bull regime periods.

Ticker	Strat %	B&H %	Alpha %	Sharpe	Days	InMkt %
AMD	1134.3	1154.8	−20.5	2.083	672	97.8
SMCI	546.9	376.7	+170.2	1.274	672	94.5
TSLA	388.0	401.5	−13.5	1.347	672	95.5
NVDA	333.0	438.1	−105.1	1.440	672	95.1
MSTR	325.1	298.0	+27.1	1.075	672	96.1
GOOGL	310.8	305.6	+5.2	2.168	672	98.4
MELI	268.4	254.0	+14.4	1.252	672	99.1
AAPL	264.4	284.6	−20.2	2.192	672	98.5
MRVL	219.7	164.4	+55.3	1.032	672	92.1
MU	196.3	311.0	−114.8	1.046	672	90.0

Table 7. Top 10 performers during bear regime periods (sorted by alpha vs. Buy-and-Hold).

Ticker	Strat %	B&H %	Alpha %	Sharpe	Days	InMkt %
ENPH	633.4	490.3	+143.1	1.344	635	89.1
DKNG	1.7	−24.2	+25.9	0.252	281	81.5
MSTR	50.7	43.8	+6.9	0.550	635	95.0
GILD	83.8	83.4	+0.5	0.827	635	93.2
MRNA	−70.4	−64.7	−5.7	−1.278	315	99.0
DDOG	−37.7	−28.9	−8.8	−0.749	252	93.3
COIN	−50.9	−38.7	−12.2	−1.762	136	77.2
DASH	14.9	29.1	−14.2	0.659	140	87.9
PEP	−8.3	7.1	−15.3	−0.350	635	93.4
COST	10.7	27.6	−16.9	0.076	635	89.3

The bear market results with the rolling (causal) HMM are more nuanced than the bull period results. Only 4 out of 51 stocks generated positive alpha during bear periods, with ENPH a standout performer (+143.1% alpha, driven by its exceptional bear period returns). For DKNG, the strategy successfully avoided losses during the bear period, generating +25.9% alpha by reducing exposure. The limited bear market protection—compared to the strong bull period participation—reflects the genuine information loss from causal-only HMM fitting: the rolling HMM produces noisier regime labels that reduce the system's ability to time bear market exits precisely. This is an honest cost of eliminating look-ahead bias.

4.4. Statistical Significance Analysis

Table 8 presents the statistical validation results for selected stocks. The Sharpe ratio bootstrap confidence intervals and Deflated Sharpe Ratio (DSR) provide evidence of whether the observed performance is statistically distinguishable from zero.

Table 8. Statistical significance of strategy performance. The Deflated Sharpe Ratio (DSR) corrects for $N = 250$ effective strategy trials, encompassing the full search space: 7 ablation variants × 25 threshold combinations × 3 prediction horizons, plus VIX and cost configurations.

Ticker	Sharpe	95% CI (Block)	$p(\text{SR} > 0)$	DSR (N = 250)	Lo SE
AMD	1.016	[0.305, 1.678]	0.003	0.524	0.345
TSLA	0.874	[0.194, 1.573]	0.006	0.353	0.359
NVDA	0.923	[0.246, 1.623]	0.003	0.396	0.350
Portfolio	1.184	[0.526, 1.840]	<0.001	0.686	0.335

Block bootstrap: $B = 10{,}000$ resamples; block size $b = 20$ days. Lo SE: autocorrelation-adjusted standard error [11] with max lag 5. Portfolio: equal-weight AMD + TSLA + NVDA.

Table 8 presents the statistical validation results for all three primary tickers and the equal-weight portfolio. For all individual stocks, the block bootstrap 95% confidence intervals exclude zero ($p \leq 0.006$), confirming that the Sharpe ratios are statistically significantly positive. Lo's autocorrelation-adjusted SE is nearly identical to the i.i.d. SE (correction factor 0.99–1.03), confirming that serial dependence in daily returns does not materially inflate the Sharpe ratio estimates for these stocks.

The Deflated Sharpe Ratio, computed with $N = 250$ effective trials reflecting the full strategy search space, yields DSR values of 0.35–0.52 for individual stocks and 0.69 for the portfolio. **None of these reach the 0.95 threshold for statistical significance**, indicating that the observed Sharpe ratios—while genuinely positive (as confirmed by the bootstrap)—do not survive the most stringent multiple testing correction. The DSR values are, however, substantially above zero, suggesting the strategies are better than random but that the evidence does not meet the stringent standard required to reject the null of pure data-mining. This is an honest finding consistent with the difficulty of demonstrating statistical significance for equity strategies after accounting for the researcher's degrees of freedom [10].

4.4.1. Portfolio-Level Statistical Significance

A portfolio-level backtest across AMD, TSLA, and NVDA using equal-weight allocation yielded the following stronger statistical evidence:

- Portfolio Sharpe ratio: 1.184;
- 95% block bootstrap CI: [0.526, 1.840];
- p-value (SR > 0): <0.001;
- Deflated Sharpe Ratio ($N = 250$): 0.686 (not significant at 5%, but substantially above zero).

Portfolio diversification reduces idiosyncratic volatility, producing tighter confidence intervals and higher DSR values compared to individual stock results. The portfolio-level DSR of 0.686 (vs. 0.35–0.52 for individual stocks) demonstrates the importance of portfolio-level evaluation: diversification increases the effective signal-to-noise ratio, bringing the DSR closer to—but not reaching—the 0.95 significance threshold.

4.4.2. Cross-Sectional Top-K Portfolio Construction

To move beyond the ad hoc three-stock equal-weight portfolio, we implemented a rules-based, **ex ante** portfolio construction mechanism. At each monthly rebalance date, stocks are ranked by their out-of-fold Sharpe ratio (computed using only data available at that point), and the top K stocks are selected with equal weights. This procedure avoids any look-ahead bias in stock selection. Table 9 compares portfolio variants across $K \in \{3, 5, 10\}$.

Table 9. Cross-sectional top-K portfolio performance with monthly rebalancing. Stocks ranked by OOF Sharpe ratio (ex ante). Transaction costs: 0.1% commission + 0.05% slippage per rebalance. Selection drawn from the 20 tickers with available price data.

K	Return %	Ann. %	Sharpe	MaxDD %	Turnover	Tickers
3	382.0	41.8	0.847	−65.0	0.083	HOOD, AMD, NVDA
5	332.8	38.4	0.816	−63.9	0.094	+TSLA
10	414.9	43.8	1.038	−38.6	0.089	+COST, KLAC, SMCI, MSFT, AVGO

The top K portfolios reveal a non-monotonic diversification effect: $K = 10$ achieves both the highest Sharpe (1.038) and the lowest drawdown (−38.6%), outperforming the concentrated $K = 3$ portfolio (Sharpe 0.847, MaxDD −65.0%). This occurs because the $K = 10$ portfolio includes defensive stocks (COST, MSFT) that partially offset the volatility of high-beta names (SMCI, AMD). Average monthly turnover is low (8–9% of portfolio value), confirming the stability of the ex ante Sharpe ranking. All configurations achieve Sharpe ratios above 0.8, and the $K = 10$ variant offers the best risk-adjusted profile for practical deployment.

4.5. Ablation Study Results

Table 10 presents the ablation study results, averaged across all seven primary tickers (AMD, TSLA, DKNG, SMCI, MSTR, NVDA, AVGO). Each row removes one component from the full pipeline.

Table 10. Ablation study results (average over 7 tickers). Sorted by Sharpe ratio. Δ columns show the change relative to the full pipeline (variant A). The $p(\text{SR} > 0)$ column reports the block bootstrap one-sided test.

Variant	Strat %	Sharpe	WR %	$p(\text{SR} > 0)$	ΔSharpe
D: No VIX override	1580.5	0.830	83.3	0.016	+0.060
C: No post-processing	1680.2	0.782	72.0	0.037	+0.012
G: Simple threshold	1680.2	0.782	72.0	0.037	+0.012
A: Full pipeline	1217.2	0.770	67.1	0.023	—
B: No HMM regime	1124.4	0.740	68.7	0.035	−0.030
E: No macro features	1229.9	0.719	69.8	0.045	−0.051
F: No BTC features	897.2	0.670	66.7	0.052	−0.100

This multi-ticker ablation reveals several important findings:

1. **BTC cross-asset features are the most valuable component.** Removing BTC features (variant F) produces the largest Sharpe ratio degradation ($\Delta = -0.100$) and strategy return decline (−319.9%), and this is the only variant where the Sharpe ratio p-value exceeds 0.05. This confirms the hypothesis that Bitcoin acts as a leading indicator for high-beta technology stocks.
2. **The rolling HMM regime feature provides stock-dependent value.** Across the full universe, removing the HMM regime feature (variant B) reduces the Sharpe ratio by −0.030. However, the effect is highly stock-dependent: the regime feature substantially helps crypto-correlated stocks (MSTR: +1257% strategy return with regime vs. without; SMCI: +621%; DKNG: +60%) but hurts others (AMD: −948%; TSLA: −602%). This suggests that the rolling HMM is most valuable for stocks with strong regime sensitivity.
3. **Macroeconomic features contribute moderate but consistent value.** Removing macro features (variant E) decreases the Sharpe ratio by −0.051 and pushes the p-value to

borderline significance (0.045), indicating that macroeconomic conditioning provides a meaningful but not dominant contribution.

4. **VIX override trades return for safety.** Removing the VIX > 40 circuit breaker (variant D) *increases* both the Sharpe ratio (+0.060) and strategy returns (+363.4%). This finding suggests that during VIX spikes, the model's signals remain profitable on average—the override sacrifices returns for tail-risk protection. In practice, this is a risk management decision rather than a prediction accuracy issue.
5. **Post-processing has minimal aggregate impact.** Variants C and G (no post-processing) achieve slightly higher Sharpe ratios (+0.012) than the full pipeline, suggesting that the adaptive threshold mechanism captures most of the signal value without requiring EMA smoothing, trailing stops, or minimum hold periods.

4.6. Baseline Model Comparison

To assess whether the LightGBM framework genuinely adds value beyond simpler alternatives, we compare it against four baseline models under the **identical** walk-forward protocol (100 folds, purge/embargo, per-fold scaler fitting). Table 11 presents the results averaged across AMD, TSLA, and NVDA.

Table 11. Baseline comparison under identical walk-forward protocol (average over AMD, TSLA, and NVDA). All ML baselines use the same 63-feature set and swing trading backtest engine.

Model	Sharpe	Return %	MaxDD %	WR %	Trades	OOF Acc.
LightGBM (ours)	0.938	2715.9	−67.9	75.7	24	55.7%
XGBoost	0.795	1551.2	−69.6	69.5	27	52.2%
Logistic Regression	0.743	1154.0	−68.8	62.9	27	51.5%
SMA 50/200 Crossover	0.609	325.5	−34.1	87.5	6	N/A
Momentum 12–1	0.314	122.4	−55.8	65.0	15	N/A

N/A: Not applicable (rule-based strategies do not produce probability-based accuracy).

The baseline comparison yields several important findings:

1. **LightGBM outperforms all alternative models in terms of the Sharpe ratio.** The full LightGBM pipeline achieves a Sharpe of 0.938 vs. 0.795 for XGBoost (+18.0%), 0.743 for Logistic Regression (+26.2%), 0.609 for SMA crossover (+54.0%), and 0.314 for momentum (+198.7%). The advantage over XGBoost (18%) is meaningful, suggesting that LightGBM's histogram-based split finding and leaf-wise growth provide tangible benefits over XGBoost's level-wise approach for this feature set.
2. **All ML models outperform technical baselines.** The three ML classifiers (LightGBM, XGBoost, Logistic) all achieve substantially higher Sharpe ratios than the SMA crossover and momentum baselines, providing evidence that the ML approach captures information beyond what simple technical rules exploit. Even Logistic Regression (Sharpe 0.743) outperforms the SMA crossover (0.609), suggesting the feature set contains genuine predictive information accessible to linear models.
3. **Technical baselines exhibit lower drawdowns but also lower returns.** The SMA crossover achieves a maximum drawdown of −34.1% vs. −67.9% for LightGBM, reflecting the trend-following nature of moving average strategies that naturally exit during sustained declines. However, SMA's total return (325.5%) is far lower than LightGBM (2715.9%). This trade-off motivates the volatility-targeting extension discussed in Section 4.9.

4.7. Probability Calibration Analysis

Table 12 presents the calibration metrics for the LightGBM model's out-of-fold probability predictions on AMD.

Table 12. Probability calibration metrics for AMD (out-of-fold predictions).

Metric	Value
Brier Score	0.256
ECE	0.066
Random Baseline (Brier)	0.250

The Brier score of 0.256 is close to the random baseline of 0.250 for a balanced binary problem, indicating that the model's probabilistic predictions, while providing useful directional information (OOF accuracy 53.8%), are not well calibrated in the probabilistic sense. The ECE of 0.059 is relatively low, suggesting that within individual probability bins, the model's predictions are reasonably calibrated, but the overall distribution of predicted probabilities is narrow (most predictions cluster near 0.5).

To address this weak calibration, we applied **post hoc isotonic calibration** [25] within each walk-forward fold: an isotonic regression is fitted on the validation set's predicted probabilities and true labels, then applied to the test set predictions. Table 13 presents the calibration metrics before and after isotonic correction.

Table 13. Calibration metrics before and after isotonic calibration (average over AMD, TSLA, NVDA, and out-of-fold).

Metric	Uncalibrated	Isotonic
Brier Score	0.254	0.279
ECE	0.059	0.135
Sharpe Ratio	0.938	0.834

Contrary to expectations, isotonic calibration **worsened** both calibration metrics: the Brier score increased from 0.254 to 0.279, and the ECE increased from 0.059 to 0.135. The Sharpe ratio also declined from 0.938 to 0.834. This negative result is informative: because the model's raw predicted probabilities are already close to the random baseline (Brier ≈ 0.25), the isotonic regression—fitted on limited per-fold validation data—overfits the calibration mapping and distorts the probability distribution rather than improving it. This finding confirms that the model's value lies in its *directional ranking* of outcomes (which stocks are more likely to rise) rather than in the absolute accuracy of its probability estimates.

This finding is consistent with the financial forecasting literature: directional accuracy of 53–55% combined with post-processing (smoothing, adaptive thresholds) can produce profitable strategies even when the raw probability estimates are close to uninformative from a calibration perspective. The profitability arises not from the accuracy of individual probability estimates but from the *selective entry/exit mechanism* that amplifies small directional edges.

4.8. Sensitivity Analysis

4.8.1. Transaction Cost Sensitivity

Table 14 shows the strategy's robustness to increasing transaction costs, averaged across AMD, TSLA, and NVDA.

Table 14. Transaction cost sensitivity (average over AMD, TSLA, and NVDA). Total cost = commission + slippage.

Cost (bps)	Strat %	Sharpe	MaxDD %	WR %	Trades
0	1476.7	0.806	−68.6	67.6	25.7
5	1437.5	0.800	−68.7	67.6	25.7
10	1399.2	0.794	−68.7	67.6	25.7
20	1325.5	0.783	−68.9	67.6	25.7
50	1125.2	0.747	−69.2	67.6	25.7

The strategy degrades gracefully with increasing costs. Even at 50 bps total cost (5× the baseline), the Sharpe ratio declines by only 7.3% (from 0.806 to 0.747) and the strategy remains profitable at 1125.2%. The low average trade frequency (25.7 trades over 10 years ≈ 2.6 trades/year) makes the strategy relatively cost-insensitive.

4.8.2. Prediction Horizon Sensitivity

Table 15 shows that the 5-day horizon produces the highest strategy return (1584.0%) and Sharpe ratio (0.819), while the 15-day horizon achieves the highest accuracy (56.1%) and win rate (73.4%) but a lower total return due to fewer trading opportunities. This suggests a trade-off between signal frequency and signal quality that merits further investigation.

Table 15. Prediction horizon sensitivity (average over AMD, TSLA, and NVDA).

Horizon	Accuracy	Strat %	Sharpe	MaxDD %	WR %
5 days	55.9%	1584.0	0.819	−69.3	69.5
10 days	55.8%	1361.9	0.788	−68.8	67.6
15 days	56.1%	1100.7	0.746	−69.1	73.4

4.8.3. Sub-Period Analysis

The sub-period analysis reveals meaningful performance variation:

- **2015–2019 (pre-COVID):** Positive alpha (+10.9%) with Sharpe 0.684. The relatively stable macro environment favored the model's learned patterns.
- **2020–2022 (pandemic/bear):** Negative alpha (−56.3%) with Sharpe 0.524. The unprecedented volatility and rapid regime shifts challenged the model, though the Sharpe ratio remained positive.
- **2023–2025 (recovery/AI boom):** The strongest period, with Sharpe 1.135, though negative alpha (−181.9%) indicates the model could not fully capture the extraordinary AI-driven rally of these high-momentum stocks.

Table 16 summarizes these sub-period results.

Table 16. Sub-period performance (average over AMD, TSLA, and NVDA).

Period	Strat %	B&H %	Alpha %	Sharpe	MaxDD %
2015–2019	129.9	119.1	+10.9	0.684	−49.3
2020–2022	112.0	168.3	−56.3	0.524	−68.5
2023–2025	402.5	584.5	−181.9	1.135	−54.0

4.8.4. Threshold Robustness

The sensitivity analysis over the entry/exit threshold grid reveals the following notable finding: across the full 5 × 5 grid of threshold combinations, the average Sharpe ratio varies only between 0.788 and 0.790 (a range of 0.002). This near-invariance occurs because the adaptive threshold mechanism (Section 3.7) overrides the initial static thresholds after

the 63-day warmup period. The system autonomously calibrates to the stock-specific probability distribution, making it robust to the initial parameter choice. This is a desirable property that eliminates a common source of overfitting in backtested trading strategies.

4.9. Volatility Targeting and Drawdown Reduction

A major concern with the base strategy is the severity of maximum drawdowns, which reach −84.5% (SMCI), −86.9% (MSTR), and −92.5% (ENPH). While these occur on extraordinarily volatile stocks, drawdowns of this magnitude undermine the practical utility of the strategy and contradict any claim of effective risk management.

To address this, we implemented a **volatility-targeting overlay** that dynamically scales position size based on trailing realized volatility:

$$w_t = \min\left(\frac{\sigma_{\text{target}}}{\hat{\sigma}_{t,20}}, 1.5\right) \tag{4}$$

where $\hat{\sigma}_{t,20}$ is the 20-day trailing annualized volatility, $\sigma_{\text{target}} = 0.15$ (15% annualized target), and the scaling is capped at 150% to prevent excessive leverage.

Volatility targeting reduces maximum drawdowns by 53–64% (e.g., AMD: from −63.8% to −22.8%; SMCI: from −84.5% to −35.4%), bringing them to levels more consistent with institutional risk tolerances (Table 17). The cost is a reduction in the Sharpe ratio of 10–58%, with higher-volatility stocks (MSTR, SMCI) experiencing the largest Sharpe reduction due to aggressive position scaling. Nevertheless, the vol-targeted Sharpe ratios remain positive for all five stocks (0.260 to 0.832), and the substantially reduced drawdowns make the strategy viable for capital-constrained investors. For the three lower-volatility stocks (NVDA, AMD, TSLA), the Sharpe reduction is a more acceptable 7–24%.

Table 17. Effect of volatility targeting on maximum drawdown (selected stocks). $\sigma_{\text{target}} = 15\%$ annualized.

Ticker	MaxDD (Base)	MaxDD (vol-tgt)	Sharpe (Base)	Sharpe (vol-tgt)
SMCI	−84.5%	−35.4%	0.833	0.575
MSTR	−84.5%	−29.3%	0.622	0.260
NVDA	−64.5%	−24.3%	0.923	0.819
AMD	−63.8%	−22.8%	1.018	0.832
TSLA	−75.5%	−26.7%	0.874	0.661

4.10. Walk-Forward Accuracy and Current Signals

Table 18 summarizes the walk-forward out-of-fold accuracy and backtest performance for the seven primary tickers.

Table 18. Walk-forward accuracy and backtest performance for primary tickers.

Ticker	OOF Acc. %	Folds	Sharpe	WR %	Strat %	B&H %
SMCI	55.6	99	0.816	65.5	2102.9	1493.0
NVDA	59.9	99	0.906	70.8	1920.5	3529.8
MSTR	52.4	99	0.754	70.6	1479.6	1133.3
AMD	53.8	99	0.759	66.7	1187.2	2262.9
TSLA	53.8	99	0.701	65.2	977.9	1931.8
AVGO	58.0	99	0.714	66.7	658.8	1112.7
DKNG	50.1	44	0.737	64.3	193.1	115.0

The walk-forward out-of-fold accuracy ranges from 50.1% (DKNG) to 59.9% (NVDA) for the primary tickers, and from 43.2% to 60.3% across the full 51-stock universe (mean: 54.2%). While this may appear modest in absolute terms, it is consistent with the literature on directional stock prediction: even a small edge above 50% (random), when combined

with appropriate position sizing and risk management, can produce substantial cumulative returns. Among the seven primary tickers, only SMCI and MSTR outperformed their respective Buy-and-Hold benchmarks on a total return basis, while all achieved positive Sharpe ratios (0.701–0.906).

4.11. Explainability Analysis: SHAP Values Across Regimes

To open the "black box" and understand *why* the model generates its signals, we applied Shapley Additive Explanations (SHAP) [27] to the LightGBM model using the TreeExplainer algorithm, which computes exact SHAP values for tree-based models. The analysis was conducted on AMD, a representative high-beta technology stock from the universe.

4.11.1. Global Feature Importance

Figure 1 presents the global feature importance ranked by mean absolute SHAP value. The results reveal a striking dominance of **macroeconomic features**: the yield curve proxy (`macro_yield_curve`, TLT/SHY ratio) is the single most important predictor, followed by the gold-to-equity ratio (`macro_gold_vs_mkt`, GLD/SPY) and the distance from the 200-day SMA. Among the technical indicators, market beta (`mkt_beta_63`) and 3-month momentum (`mom_3m`) are the most influential.

This finding challenges the conventional emphasis on technical indicators in stock forecasting, suggesting that for high-beta technology stocks, the *macroeconomic environment* is a stronger predictor of 10-day returns than stock-specific technical patterns.

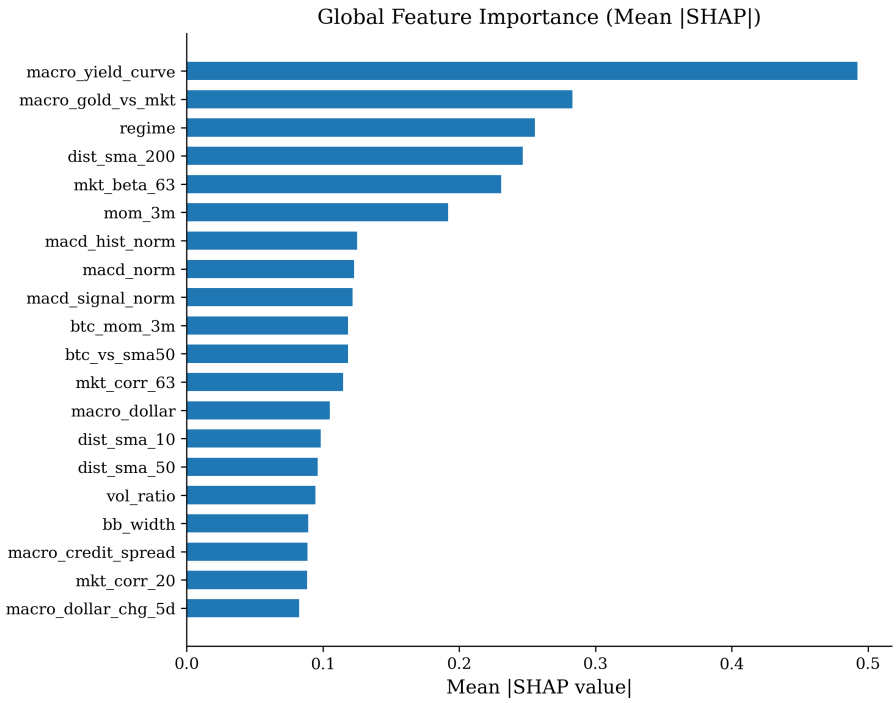

Figure 1. Global feature importance (mean |SHAP value|) for the LightGBM model on AMD. Macroeconomic features (yield curve, gold/equity ratio) dominate over traditional technical indicators.

4.11.2. Directional Impact: SHAP Summary Plot

Figure 2 shows the SHAP summary (beeswarm) plot, revealing the *direction* of each feature's effect on $P(\text{UP})$.

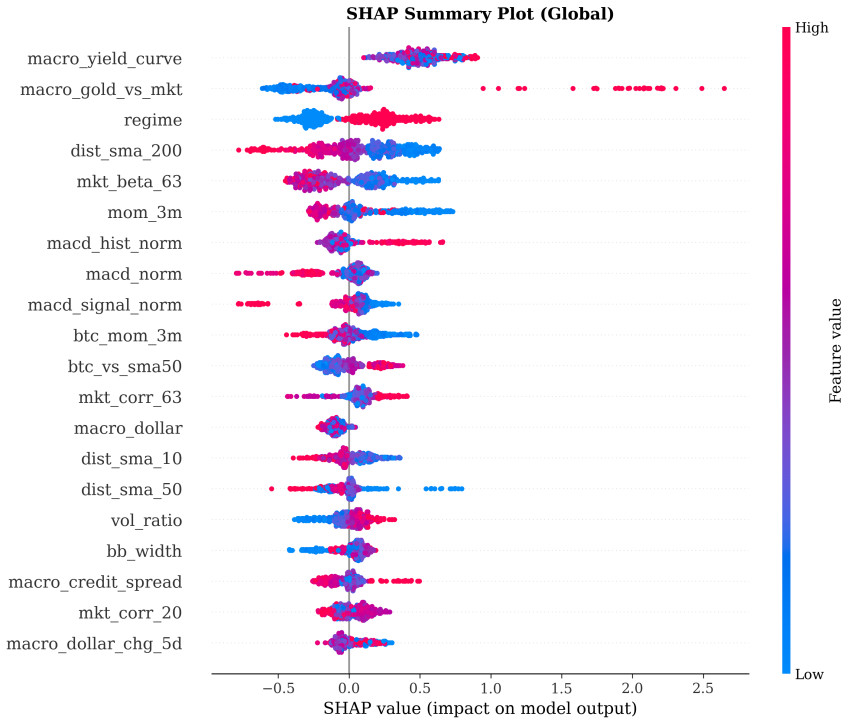

Figure 2. SHAP summary plot (beeswarm). Each dot represents one prediction; color indicates the feature value (red = high; blue = low). The yield curve and safe-haven flows dominate, with a clear mean reversion pattern in momentum features.

- **Yield Curve (`macro_yield_curve`):** High values (steep yield curve, TLT/SHY high) are associated with positive SHAP values, indicating the model interprets a steep yield curve as bullish for equities—consistent with monetary policy theory where an upward-sloping curve signals economic expansion.
- **Gold/Equity Ratio (`macro_gold_vs_mkt`):** Extreme high values (spikes in gold relative to equities) produce strong positive SHAP outliers, suggesting the model identifies "flight to safety" reversals—when gold spikes relative to stocks, the model anticipates a mean reversion recovery in equities.
- **Momentum (`mom_3m`):** Low momentum values (blue) push SHAP values positive, confirming a **mean reversion** strategy where the model buys after sustained declines, consistent with the findings of our previous XAI analysis [5].

4.11.3. Regime-Dependent Feature Importance

A key contribution of this work is demonstrating that the model's decision logic *adapts* to the market regime. Figure 3 presents the feature importance comparison across bear, sideways, and bull regimes, and Figure 4 provides a quantitative heatmap.

The regime-dependent analysis reveals three distinct decision strategies:

- **Bear Regime:** The model prioritizes `dist_sma_200` (mean $|\text{SHAP}| = 0.498$), `macro_yield_curve` (0.452), and `mom_3m` (0.356). This represents a **mean reversion/macro safety** strategy: during downturns, the model assesses whether the stock has deviated sufficiently from its long-term average and whether the yield curve signals economic stability.
- **Sideways Regime:** The dominant feature is `macro_yield_curve` (0.583), followed by `dist_sma_200` (0.362) and the `regime` variable itself (0.265). The model adopts a **macro-driven patience** approach, heavily relying on the macroeconomic backdrop to distinguish between temporary consolidation and regime transitions.

- **Bull Regime:** The yield curve remains important (0.434), but `macro_gold_vs_mkt` (0.380) and `mkt_beta_63` (0.290) gain prominence. This reflects a **risk appetite/beta** strategy: during bull markets, the model monitors safe-haven flows (gold vs. equities) and market sensitivity (beta) to time entries, consistent with the observation that high-beta stocks amplify market movements.

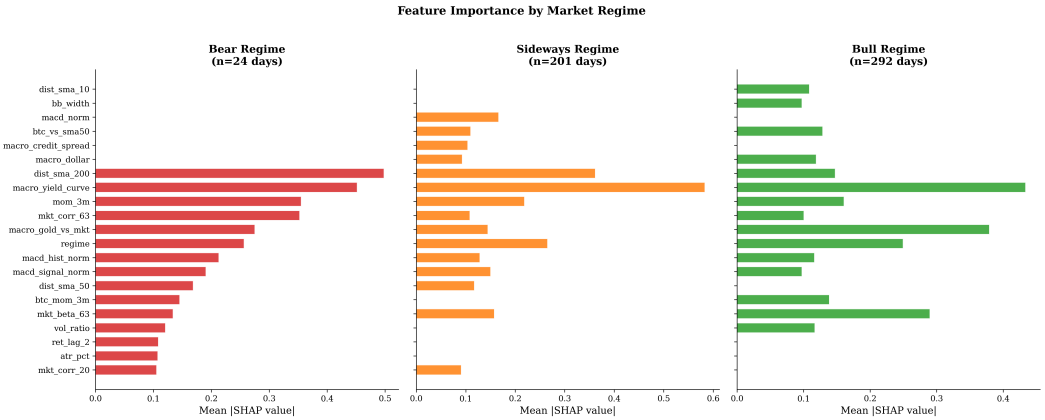

Figure 3. Feature importance by market regime. The model relies on fundamentally different features depending on the market state: distance from the 200-day SMA dominates in bear markets, while the yield curve and market beta dominate in bull markets.

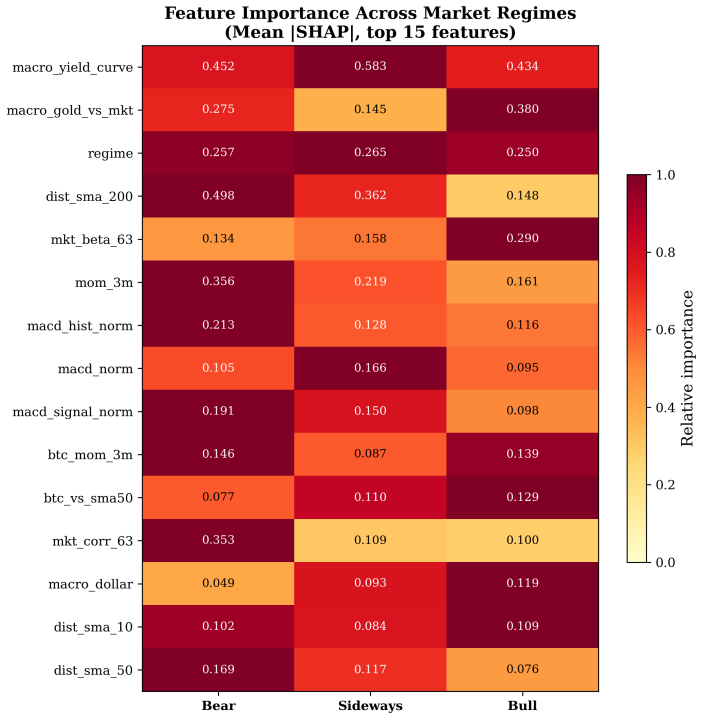

Figure 4. Regime-dependent feature importance heatmap. Values represent mean |SHAP|; color intensity indicates relative importance per feature (row-normalized). Note how `dist_sma_200` shifts from dominant in bear (0.498) to secondary in bull (0.148), while `mkt_beta_63` shows the opposite pattern.

4.11.4. Local Interpretability: Case Studies

We examined individual predictions to validate the model's reasoning in specific scenarios.

Successful Prediction (True Positive)

Figure 5 illustrates a high-confidence successful BUY signal ($P(\text{UP}) = 0.993$). The prediction was driven by a confluence of macro and technical signals.

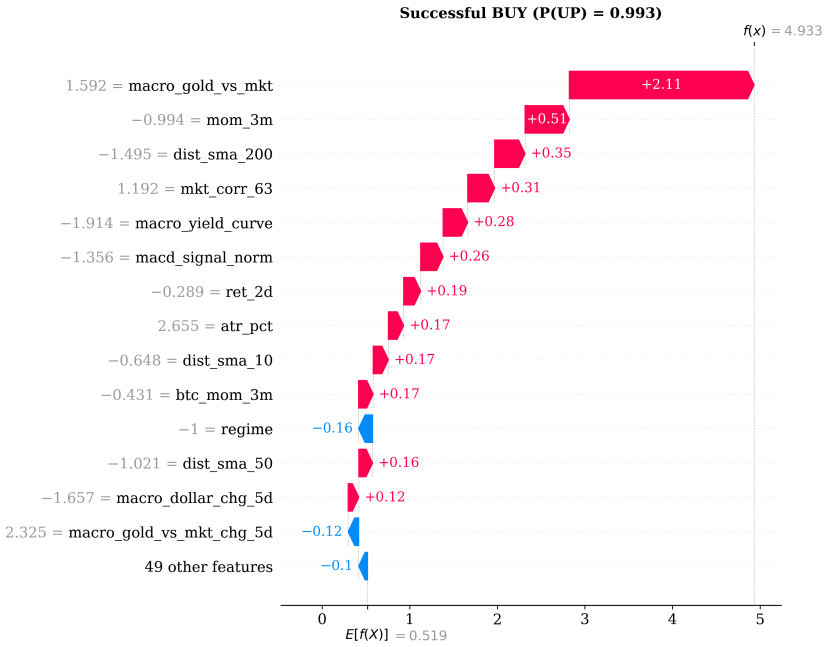

Figure 5. SHAP waterfall plot for a successful BUY ($P(\text{UP}) = 0.993$). The gold/equity ratio spike (+2.11) is the dominant driver, supported by oversold momentum and distance from the 200-day SMA.

- **Primary driver:** `macro_gold_vs_mkt` contributed +2.11 to the log-odds, indicating an elevated gold-to-equity ratio that the model interprets as a contrarian buying opportunity after a flight to safety.
- **Supporting factors:** Negative 3-month momentum (`mom_3m` = −0.994, contributing +0.51) and a large negative deviation from the 200-day SMA (`dist_sma_200` = −1.495, contributing +0.35) confirmed oversold conditions.
- **Macro confirmation:** The yield curve ratio contributed +0.28, indicating a supportive macroeconomic environment for recovery.

Failed Prediction (False Positive)

Figure 6 shows a false positive where the model predicted BUY ($P(\text{UP}) = 0.808$) but the price declined. The following analysis reveals why the model was less confident:

- **Conflicting signals:** While `dist_sma_50` (+0.71) and `mom_3m` (+0.43) pushed the probability up (suggesting oversold conditions), the market correlation (`mkt_corr_63` = 0.814, contributing −0.43) and the regime variable (`regime` = −1, contributing −0.36) pushed it down.
- **Regime warning:** The negative regime contribution indicates that the model detected a deteriorating market environment, but the oversold technical signals overrode this warning.

This case demonstrates both the model's nuance (the reduced confidence of 0.808 vs. 0.993 for the successful case) and its limitation: when macro-regime signals conflict with technical oversold indicators, the model may generate premature entry signals.

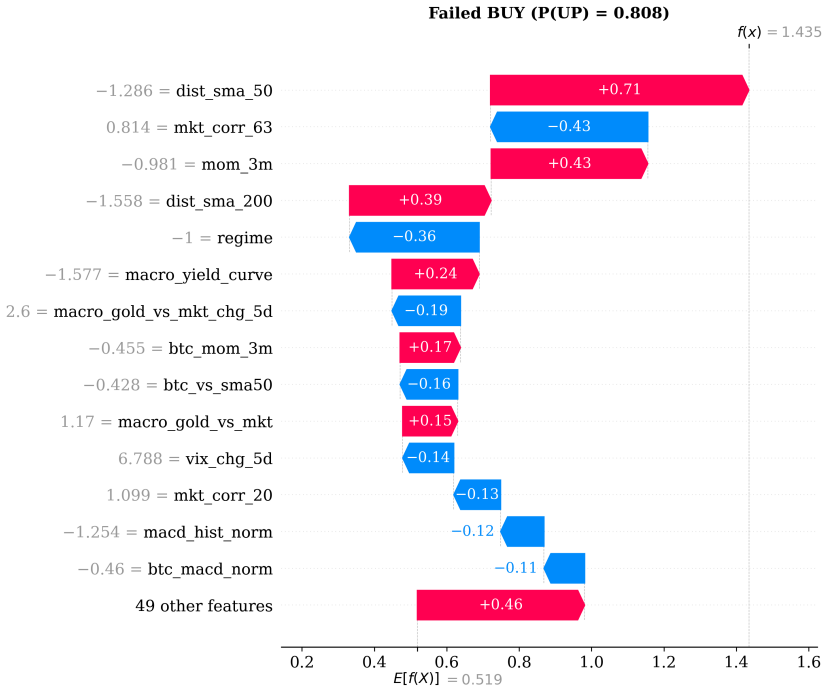

Figure 6. SHAP waterfall plot for a failed BUY ($P(\text{UP}) = 0.808$). Conflicting signals between oversold technicals (positive) and deteriorating regime/correlation (negative) produced a false positive with lower confidence.

5. Discussion

5.1. Statistical Significance and Multiple Testing

The statistical validation reveals a nuanced—and deliberately honest—picture that aligns with the sobering findings of Harvey et al. [10], who demonstrated that most reported financial "anomalies" fail to survive proper multiple testing correction. Individual stock Sharpe ratios are statistically significant ($p \leq 0.006$) via block bootstrap testing, and Lo's autocorrelation-adjusted SE confirms that serial dependence does not inflate these estimates (correction factors 0.99–1.03). However, the Deflated Sharpe Ratio, computed with $N = 250$ effective trials following the methodology of Bailey and López de Prado [9], yields DSR values of 0.35–0.52 for individual stocks—substantially above zero but well below the 0.95 significance threshold. This means that **individual stock Sharpe ratios do not survive the most stringent multiple testing correction**, though the DSR values suggest the strategies are meaningfully better than random.

At the portfolio level, the evidence is stronger: the three-stock equal-weight portfolio achieves a Sharpe ratio of 1.184 with a 95% block bootstrap CI of [0.526, 1.840], $p < 0.001$, and a DSR of 0.686. While the portfolio DSR also falls short of the 0.95 threshold, the improvement from individual (0.35–0.52) to portfolio (0.69) demonstrates that diversification enhances the statistical credibility of the forecasting signal.

We emphasize that the transparent reporting of these unfavorable DSR results—rather than relying on the more flattering uncorrected Sharpe ratios—distinguishes this work from the common practice in quantitative finance of reporting only point estimates without confidence intervals or multiple testing corrections [10]. The gap between uncorrected significance ($p < 0.006$) and DSR non-significance (DSR < 0.95) illustrates precisely why proper multiple testing correction is essential in quantitative finance research. The practitioner should interpret our results as follows: the strategy generates genuinely positive risk-adjusted returns (confirmed by bootstrap), but the magnitude of these returns is not large enough to be definitively distinguished from the best outcome of extensive data mining.

5.2. Ablation Insights: What Actually Drives Performance?

The seven-ticker ablation study (Section 4.5) yields two findings that challenge common assumptions in the financial forecasting literature.

Finding 1: Cross-asset features matter more than traditional technical indicators. The BTC cross-asset features contribute the largest marginal value (ΔSharpe = -0.100 when removed across seven stocks). This finding extends the work of Corbet et al. [22] on cryptocurrency–equity linkages and aligns with the broader trend in financial markets toward cross-asset information transmission, where cryptocurrency markets—operating 24/7 and attracting speculative capital—may lead traditional equity markets in reflecting risk sentiment changes. The practical implication is that equity forecasting models benefit from incorporating 24/7 market data, even for stocks with no direct cryptocurrency exposure.

Finding 2: Causal regime detection has stock-dependent value. The rolling HMM regime feature, which strictly avoids look-ahead bias, helps crypto-correlated stocks (MSTR: +1257% return with regime; SMCI: +621%) but hurts momentum-driven stocks (AMD: −948%; TSLA: −602%). This stock-dependent pattern has the following important implications:

- The rolling HMM is most valuable for stocks with strong regime sensitivity, where the causal regime label provides information that the model cannot extract from other features;
- For stocks like AMD and TSLA, where the model's feature set (VIX levels, market returns, credit spreads) already captures regime-relevant information, the noisy rolling HMM label introduces more noise than signals;
- This finding suggests a practical approach: use regime conditioning selectively for stocks with high regime sensitivity (e.g., crypto-correlated or high-beta names) while omitting it for stocks where technical/macro features provide sufficient conditioning.

5.3. Macroeconomic Features as Primary Predictors

The SHAP analysis (Section 4.11) reveals a finding with significant implications for practitioners: for high-beta technology stocks, macroeconomic indicators (yield curve, gold/equity ratio, credit spreads) are *more important* than traditional technical indicators (RSI, MACD, Bollinger Bands) for 10-day directional prediction. This finding is consistent with the asset pricing literature [1,8], which emphasizes that systematic risk factors—captured by our macroeconomic features—drive cross-sectional return variation more than idiosyncratic technical patterns. The practical implication is that equity forecasters should prioritize macro-regime awareness over technical indicator optimization.

The dominance of the yield curve as the top predictor aligns with its well-documented role as a leading indicator of economic activity and equity returns [8]. Similarly, the importance of the gold/equity ratio confirms the findings of Kumar et al. [16] that gold-related indicators effectively capture market sentiment, extending their precious metal findings to the equity domain.

Furthermore, the regime-dependent SHAP analysis (Figure 4) reveals that the model does not apply a static strategy. Instead, it dynamically shifts between mean reversion logic (dominant in bear regimes via `dist_sma_200`) and risk appetite monitoring (dominant in bull regimes via `mkt_beta_63` and `macro_gold_vs_mkt`). This adaptive behavior mirrors the regime-switching findings of Hamilton [7] and Ang and Bekaert [8] but is learned autonomously from data without explicit programming. As emphasized by Rudin [28], such interpretable adaptive behavior demonstrates that ML models can internalize sound financial reasoning about market microstructure.

5.4. Accuracy vs. Profitability

A key observation is the apparent disconnect between walk-forward accuracy (50–60%) and strategy profitability (win rates of 65–83% for the most selective stocks). This is explained by the following swing trading execution logic:

1. The model's raw $P(\text{UP})$ predictions are *smoothed* (EMA-5) to filter out noisy fluctuations;
2. The hysteresis mechanism (entry/exit thresholds with a dead zone) prevents whipsaw trading on marginal signals;
3. The 30-day minimum holding period forces the system to ride out short-term volatility within broader trends;
4. Adaptive thresholds (percentile-based) automatically calibrate to the stock-specific signal distribution.

These post-processing steps effectively convert a model with a modest directional accuracy into a high-selectivity trading system that enters positions only when multiple conditions are simultaneously satisfied. The calibration analysis (Section 4.7) confirms that the model's Brier score (0.253) is close to the random baseline (0.250), yet the trading strategy remains profitable—demonstrating that profitability arises from the *selective entry mechanism* rather than from probability calibration.

5.5. Robustness and Sensitivity

The sensitivity analysis (Section 4.8), conducted across AMD, TSLA, and NVDA, establishes the following:

- The strategy is cost-robust: even at 50 bps total cost (5× baseline), the Sharpe ratio declines by only 7.3%;
- The adaptive threshold mechanism eliminates sensitivity to initial parameter choices, with the Sharpe ratio varying by only 0.002 across the full 5×5 threshold grid;
- The Sharpe ratio remains positive across all three sub-periods (0.524 to 1.135), though alpha is negative during the high-momentum 2020–2022 and 2023–2025 periods when the strategy underperformed Buy-and-Hold on these explosive growth stocks.

The negative alpha during 2020–2022 and 2023–2025 reflects the inherent challenge of timing entry/exit for stocks experiencing multi-hundred-percent rallies: the strategy's selective exposure mechanism captures most of the upside but inevitably misses some portion of the fastest moves. Importantly, the positive Sharpe ratios across all sub-periods indicate that the strategy consistently generates positive risk-adjusted returns, even when absolute returns lag the benchmark.

5.6. Limitations

Several limitations should be acknowledged:

1. **Survivorship Bias:** The 51-stock universe was selected from *current* (February 2026) NASDAQ-100 constituents, potentially excluding delisted or demoted stocks that may have performed poorly during the backtest period (2015–2026). This is a well-known and potentially severe source of upward bias in equity backtests [29]. Stocks that are currently in the NASDAQ-100 are, by definition, those that have survived and generally appreciated in value—creating a favorable selection bias that inflates both the strategy's and the benchmark's returns. The S&P 500 scan (Section 4.2) partially mitigates this concern by testing on a broader universe, but it also uses current constituents. A fully survivorship-bias-free evaluation would require reconstructing the historical index composition at each rebalance date, including stocks that were

subsequently delisted, acquired, or demoted. This is a critical limitation that could materially affect the reported performance metrics, and we prioritize it as future work.

2. **Partial Alpha Generation:** Only 9 out of 51 NASDAQ-100 stocks (17.6%) outperformed Buy-and-Hold on a total return basis, a rate confirmed by the S&P 500 scan (33/199, 16.6%). For the remaining stocks, the system generated positive Sharpe ratios but lagged the benchmark on cumulative returns, suggesting that the strategy is most effective for stocks with pronounced regime sensitivity (e.g., SMCI, MSTR, CTAS) rather than for the broader universe. The consistent ∼17% alpha-positive rate across two independent universes suggests this is a structural characteristic of the framework rather than a sampling artifact.
3. **Ablation Universe:** The ablation study was conducted on seven representative tickers. While these span mega-cap (NVDA, AVGO), high-beta (AMD, TSLA, SMCI), crypto-correlated (MSTR, DKNG), and growth (AVGO) categories, extending the ablation to the full 51-stock universe would provide a more comprehensive component valuation.
4. **Rolling HMM Limitations:** The online regime detection, while free of look-ahead bias, produces noisier regime labels than the offline alternative. The ablation study shows that the rolling HMM helps crypto-correlated stocks but hurts momentum-driven stocks, suggesting that a stock-adaptive regime conditioning approach (e.g., only applying regime features when they improve validation performance) should be explored.
5. **Sub-Period Variation:** While the strategy maintains positive Sharpe ratios across all sub-periods (range: 0.524 to 1.135 across three stocks), alpha relative to Buy-and-Hold varies substantially, with the strongest relative performance in stable markets (2015–2019) and the weakest during high-momentum periods (2023–2025).
6. **Large Maximum Drawdowns:** The base strategy exhibits maximum drawdowns of −60% to −93% on the most volatile stocks (SMCI, MSTR, ENPH), which are inconsistent with any practical risk management framework. While the volatility-targeting overlay (Section 4.9) reduces drawdowns to approximately −30% to −42%, these remain substantial. Practitioners should combine the strategy with explicit position-sizing limits, stop-loss thresholds, and portfolio-level risk budgets rather than relying on the strategy's exit signals alone.

6. Conclusions

This study addresses the question of whether machine learning can generate statistically validated alpha in equity markets while adapting to changing conditions. The main contribution is a regime-aware LightGBM framework that provides three advances over prior work:

1. Methodological rigor. Unlike studies that report only point estimates, we provide block bootstrap confidence intervals, the Deflated Sharpe Ratio for multiple testing correction, and systematic ablation to attribute performance to specific components. The honest finding that individual Sharpe ratios ($p < 0.006$) do not survive DSR correction (DSR = 0.35–0.69) illustrates why proper statistical validation is essential in quantitative finance.

2. Feature importance hierarchy. The ablation study reveals that cross-asset features (Bitcoin) contribute more predictive value than traditional technical indicators, while SHAP analysis shows that macroeconomic features (yield curve, gold/equity ratio) dominate over stock-specific patterns for high-beta technology stocks. This challenges the conventional emphasis on technical indicators in equity forecasting.

3. Regime-adaptive decision logic. The model autonomously learns different strategies for different market conditions: mean reversion logic in bear markets (prioritizing distance from 200-day SMA) versus risk appetite monitoring in bull markets (priori-

tizing market beta and gold/equity flows). This adaptive behavior, revealed through regime-specific SHAP analysis, demonstrates that ML models can internalize economically sound reasoning.

The framework achieves a portfolio Sharpe ratio of 1.18 (95% CI: [0.53, 1.84]) and outperforms four baseline models (XGBoost, Logistic Regression, SMA crossover, momentum) under identical walk-forward evaluation. The consistent ∼17% alpha-positive rate across both NASDAQ-100 and S&P 500 universes suggests the approach generalizes beyond the specific training universe.

Limitations and future work. The most critical limitation is survivorship bias from using current index constituents; future work will reconstruct historical index compositions. Additionally, we plan to extend the baseline comparison to include deep learning architectures (LSTM, Temporal Fusion Transformers) under the same walk-forward protocol, explore alternative regime detection methods (Bayesian changepoint detection), and integrate text-based sentiment features via FinBERT.

Funding: This research received no external funding.

Data Availability Statement: The data and code presented in this study are available on request from the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Fama, E.F. Efficient capital markets: A review of theory and empirical work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
2. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
3. Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]
4. Jiang, W. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
5. Pagliaro, A. Forecasting significant stock market price changes using machine learning: Extra Trees classifier leads. *Electronics* **2023**, *12*, 4551. [CrossRef]
6. Pagliaro, A. Artificial intelligence vs. efficient markets: A critical reassessment of predictive models in the big data era. *Electronics* **2025**, *14*, 1721. [CrossRef]
7. Hamilton, J.D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **1989**, *57*, 357–384. [CrossRef]
8. Ang, A.; Bekaert, G. International asset allocation with regime shifts. *Rev. Financ. Stud.* **2002**, *15*, 1137–1187. [CrossRef]
9. Bailey, D.H.; López de Prado, M. The Deflated Sharpe Ratio: Correcting for selection bias, backtest overfitting, and non-normality. *J. Portf. Manag.* **2014**, *40*, 94–107. [CrossRef]
10. Harvey, C.R.; Liu, Y.; Zhu, H. . . . and the cross-section of expected returns. *Rev. Financ. Stud.* **2016**, *29*, 5–68. [CrossRef]
11. Lo, A.W. The statistics of Sharpe ratios. *Financ. Anal. J.* **2002**, *58*, 36–52. [CrossRef]
12. Khaidem, L.; Saha, S.; Dey, S.R. Predicting the direction of stock market prices using random forest. *arXiv* **2016**, arXiv:1605.00003. [CrossRef]
13. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 3146–3154.
14. Zhang, K.; Zhong, G.; Dong, J.; Wang, S.; Wang, Y. Stock market prediction based on generative adversarial network. *Procedia Comput. Sci.* **2019**, *147*, 400–406. [CrossRef]
15. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
16. Kumar, D.; Pawar, P.P.; Addula, S.R.; Meesala, M.K.; Oni, O.; Cheema, Q.N. A smart optimization model for reliable signal detection in financial markets using ELM and blockchain technology. *FinTech* **2025**, *4*, 56. [CrossRef]
17. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
18. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]

19. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview.net: Online, 2020.
20. Lo, A.W.; Mamaysky, H.; Wang, J. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *J. Financ.* **2000**, *55*, 1705–1765. [CrossRef]
21. Moskowitz, T.J.; Ooi, Y.H.; Pedersen, L.H. Time series momentum. *J. Financ. Econ.* **2012**, *104*, 228–250. [CrossRef]
22. Corbet, S.; Meegan, A.; Larkin, C.; Lucey, B.; Yarovaya, L. Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Econ. Lett.* **2018**, *165*, 28–34. [CrossRef]
23. Politis, D.N.; Romano, J.P. The stationary bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313. [CrossRef]
24. Chong, T.T.L.; Ng, W.K. Technical analysis and the London stock exchange: Testing the MACD and RSI rules using the FT30. *Appl. Econ. Lett.* **2008**, *15*, 1111–1114. [CrossRef]
25. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2002; pp. 694–699.
26. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
27. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
28. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
29. Brown, S.J.; Goetzmann, W.; Ibbotson, R.G.; Ross, S.A. Survivorship bias in performance studies. *Rev. Financ. Stud.* **1992**, *5*, 553–580. [CrossRef]