report

December 15, 2025

0.1 XGBoost Direction Prediction & Trading Strategy Report

0.2 1) Executive Summary

This report evaluates an **XGBoost binary classifier** (predicting the future return direction) and its translation into a trading strategy, benchmarked against several baselines (Linear Regression, Random Walk, Smoothed RW, Calibrated RW). Conclusions are grounded in **observable evidence from the backtest figures**.

The key takeaway is: the model is not "highly accurate" in an absolute sense, but it helps the strategy act **more aggressively when confidence is higher** and **more conservatively when confidence is lower**, which can improve overall outcomes. Key figure-supported points:

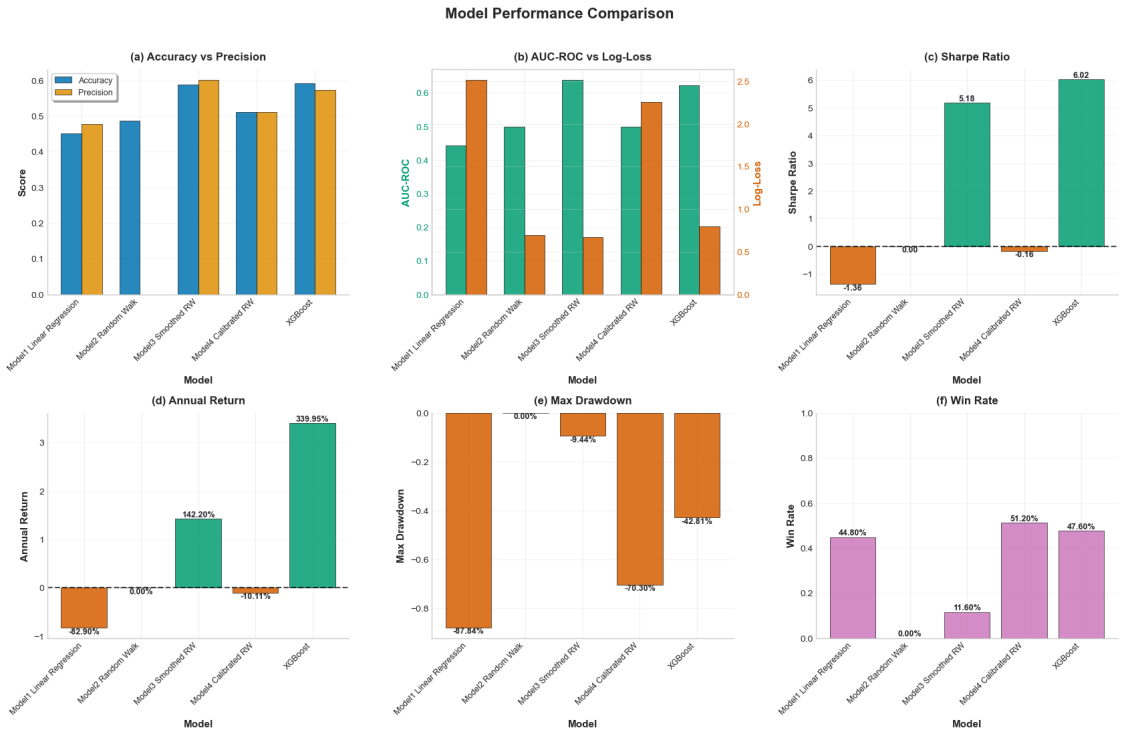
- **Overall performance**: Sharpe (roughly "return per unit risk") reaches **6.02**, with annual return **339.95%**; however, max drawdown (peak-to-trough loss) is also large at **-42.81%**. In short: **strong upside, but meaningful downside risk**.

- **Prediction quality**: ROC AUC (ability to rank up-moves ahead of down-moves) is about **0.623**—useful, but far from "perfect". The predicted probability distributions shift between classes but still overlap materially.

- **Robustness**: Under Bull/Bear regimes, XGBoost shows positive Sharpe and positive annual return in both regimes, suggesting performance is not driven by a single market segment. At the same time, some baselines show unusually high values in specific regimes—this is a reminder to be cautious about **definitions, sample size, and calculation choices** that can inflate metrics.

- **Advanced improvements**: Optuna tuning reports a best-trial Sharpe of **15.7196**; dynamic thresholds (vs static thresholds) show a tendency toward higher Sharpe/annual return in the figure. Importantly: **faster growth does not necessarily mean more stable** (drawdown improvement is not consistent), so interpretation should be cautious.

Key figure:



0.3 2) Problem & Data

0.3.1 2.1 Problem setup

- **Goal**: use features to predict the future return direction (up/down), then map predicted probabilities into trading actions (long/short/flat).

- **Evaluation**:
  - **ML metrics**: Accuracy, Precision, AUC-ROC, LogLoss

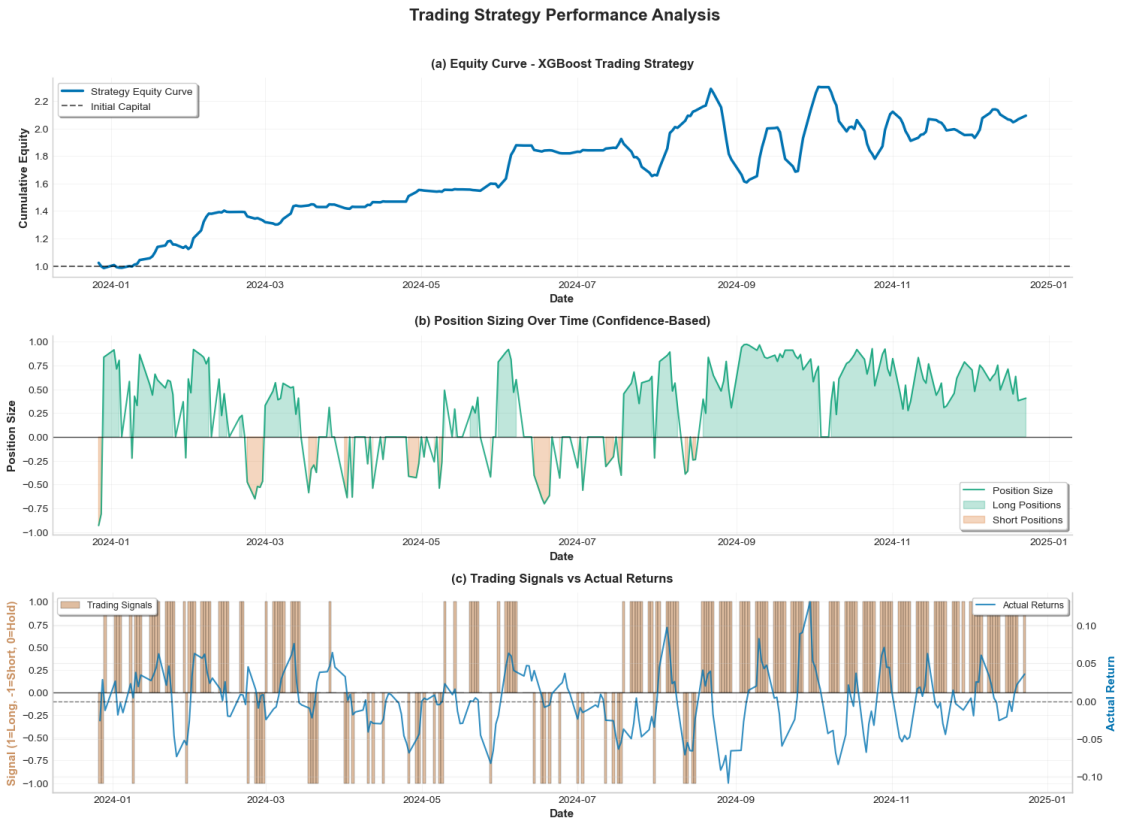
  - **Trading metrics**: Sharpe, Annual Return, Max Drawdown, Win Rate

We do not aim to "be correct every day". Instead, we care more about **earning when we are right** and **limiting losses when we are wrong**. Therefore, the same ML score can translate into very different trading outcomes depending on the execution rules.

0.3.2 2.2 Data and sample window (as reflected in plots)

From the date axis in the strategy equity curve, the shown backtest window spans roughly **2024-01 to 2025-01** (trading days). Conclusions should be interpreted as evidence **within this sample window**.

Key figure:



0.4 3) Method & Backtest Hygiene

The goal here is not to "maximize Sharpe", but to ensure the evaluation is **credible**.

The biggest risk in backtesting is "answer leakage"—even a small inadvertent use of future information can make the curve look unrealistically strong, while being impossible to reproduce in live trading.

0.4.1 3.1 Walk-forward and time causality

- **Principle**: training windows come first, test windows come later; all features/labels must respect time causality.

- **Common pitfall**: rolling/smoothing statistics (moving averages, rolling volatility, quantile thresholds). If computed once over the full sample, test-period information can leak into training-period statistics.

0.4.2 3.2 Preprocessing (missing values and scaling)

- **Principle**: any scaler/encoder must be `fit` on training data only, and `transform` applied to test data. Missing value fills (ffill/bfill) should avoid crossing the train/test boundary in a way that effectively "backfills with the future."

- **Interpretation rule in this report**: we do not treat a high-performing equity curve as proof of clean backtesting. Hygiene is a prerequisite; the performance plots matter only after hygiene is satisfied.

0.4.3 3.3 Cross-checking with outcome plots (supporting evidence only)

We use the XGBoost summary figure as an "observational" cross-check (**not a substitute for code-level leakage auditing**), to spot patterns that look inconsistent with normal trading behavior.

If an equity curve "only goes up" with near-zero drawdowns, that would be a red flag—in real markets, such perfection often indicates future information leakage or definition bias.

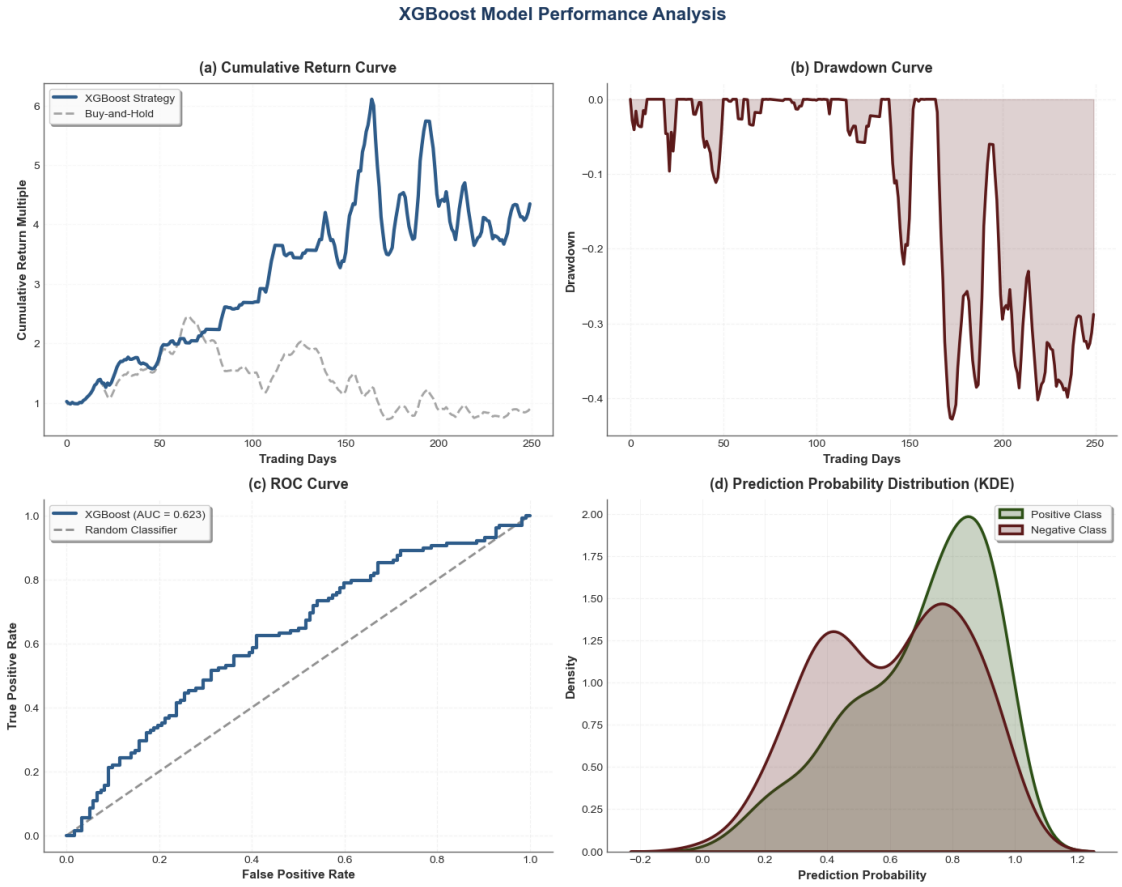
- **Cumulative return**: strategy equity rises from 1 to about **4.2**, while buy-and-hold drifts down to about **0.9**, indicating meaningful timing/directional gains in-sample.

- **Drawdown**: sizable drawdowns appear later; the low point is around **-0.42**, implying gains are concentrated in certain phases and the curve is volatile.

- **ROC**: AUC is labeled **0.623**, suggesting non-trivial information but not a strong predictor; therefore, the strategy's high Sharpe depends heavily on thresholding and position mapping.

- **Probability (KDE)**: class distributions shift, but overlap remains large—typical for financial direction prediction. The practical value often comes from converting a small edge into controlled, tradable exposure.

Key figure:



0.5 4) Baseline Results

This section separates "model quality" into two layers: **predictive metrics (ML)** and **tradability (strategy metrics)**.

Importantly, higher predictive scores (e.g., Accuracy/AUC) do not guarantee profitability. Profitability depends on how scores are mapped into orders and how risk is controlled.

0.5.1 4.1 Predictive metrics (Accuracy / Precision / AUC / LogLoss)

From the comparison plot: - **XGBoost**: Accuracy **0.59**, Precision **0.57**, AUC-ROC **0.62**—slightly better than random and potentially usable as a signal.
- **Smoothed RW**: AUC is also around **0.64**, showing simple time-series baselines can sometimes achieve comparable classification performance under certain constructions.

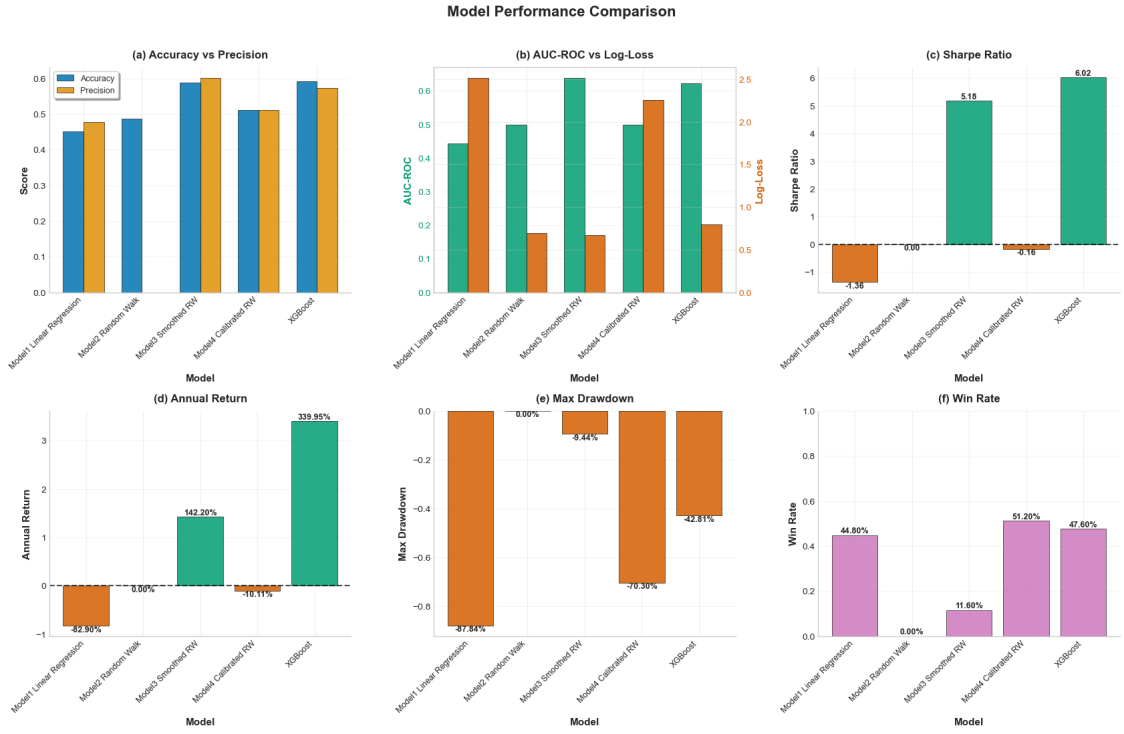
0.5.2 4.2 Strategy metrics (Sharpe / Annual Return / Max Drawdown / Win Rate)

The plot provides direct strategy-level conclusions (values annotated in the figure): - **XGBoost**: Sharpe **6.02**; Annual Return **339.95%**; Max DD **-42.81%**; Win Rate **47.60%**.
- **Smoothed RW**: Sharpe **5.18**; Annual Return **142.20%**; Max DD **-9.44%**; Win Rate **11.60%** (very low).

How to interpret "high Sharpe but low win rate / low drawdown"? ()
- A very low win rate with high Sharpe/annual return often implies a highly skewed PnL distribution: a **small number of large wins** offsets many small losses or flat periods.
- This introduces two key risks:
- **Sample dependence**: if the few large wins do not occur in a different period, performance can collapse.
- **Cost sensitivity**: if trading is frequent, many small-loss trades can be heavily impacted by transaction costs.
- In contrast, XGBoost's win rate near 50% looks more "typical" for a direction-based strategy, but its drawdown is larger, suggesting leverage/thresholds may amplify risk in certain volatility regimes.

More intuitively: - **Low win rate but still profitable** usually means "many small losses/flat periods, plus a few big opportunities."
- **Win rate around 50%** often means "frequent small wins/losses with a slight edge." Both can work, but the former relies more on rare large opportunities.

Key figure:



0.6 5) Strategy Construction & Execution (Signal → Position → PnL)

0.6.1 5.1 From probabilities to actions: thresholds and the action space

The project maps the model's "up-move probability" into trading actions: long/short/flat (or reduced exposure). The goal is not higher Accuracy per se, but fewer trades under high uncertainty and larger exposure under higher confidence, thereby improving risk-adjusted performance.

You can interpret the model output as a "confidence scale": higher confidence → more aggressive;

moderate confidence → trade less or stay out.

0.6.2 5.2 Position sizing: confidence-based continuous exposure

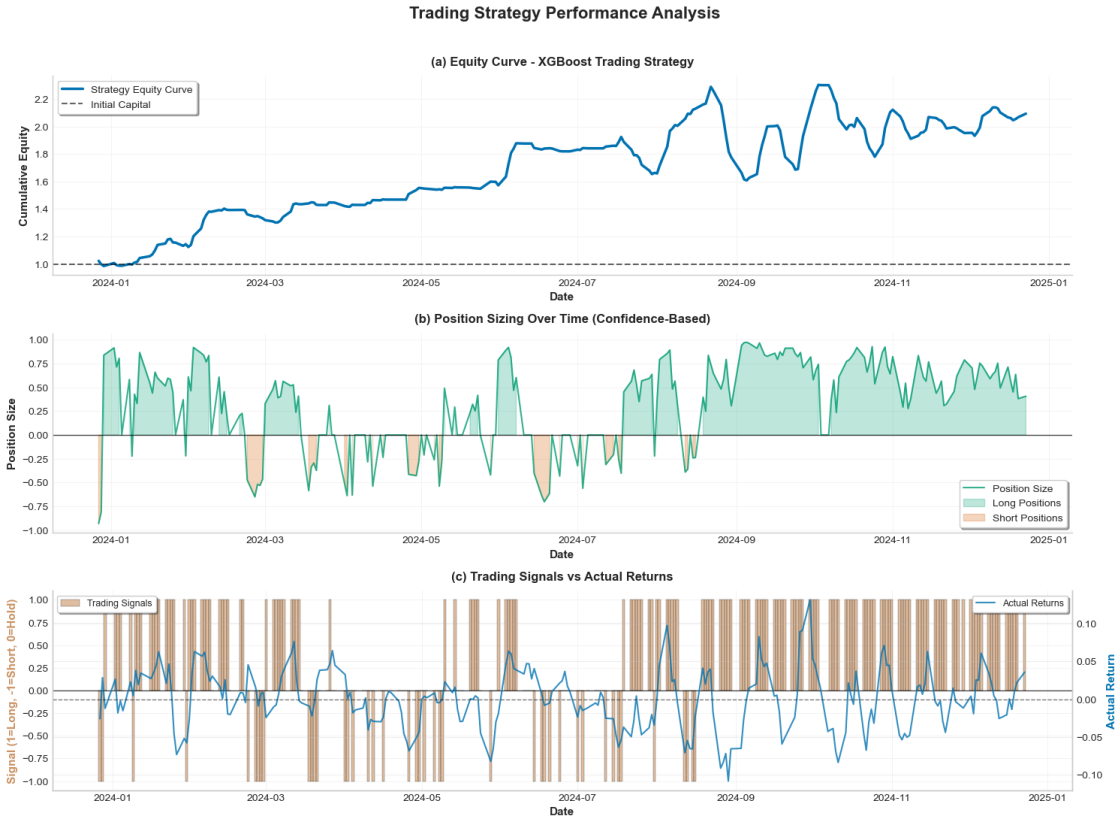
From "Position Sizing Over Time", positions vary continuously within ([-1, 1]) and tilt more long in the second half. Benefits of this design: - **Reduces binary switch noise**: avoids full flips when probabilities barely cross a threshold.
- **Lets strong signals contribute more**: extreme probabilities lead to larger exposure, concentrating returns in higher-quality predictions.

0.6.3 5.3 Trading frequency sanity check

From the signal bars in "Trading Signals vs Actual Returns", signals are fairly dense in several periods (especially after August), indicating the strategy is not ultra-low-frequency "betting."
This matters: too few trades can be luck-driven; too many trades can be dominated by costs. The plot suggests a higher switching frequency, so conclusions must explicitly acknowledge cost assumptions (see Conclusion).

The denser the bars, the more frequent the trading. If transaction costs are not modeled, the backtest can be materially overstated.

Key figure:



0.7 6) Robustness & Market Regime Consistency

The key question is whether performance is concentrated in a single phase, or whether it holds across different market regimes.

We prefer results that are not simply "one lucky big trend", but are relatively consistent across different regimes.

0.7.1 6.1 Bull vs Bear consistency

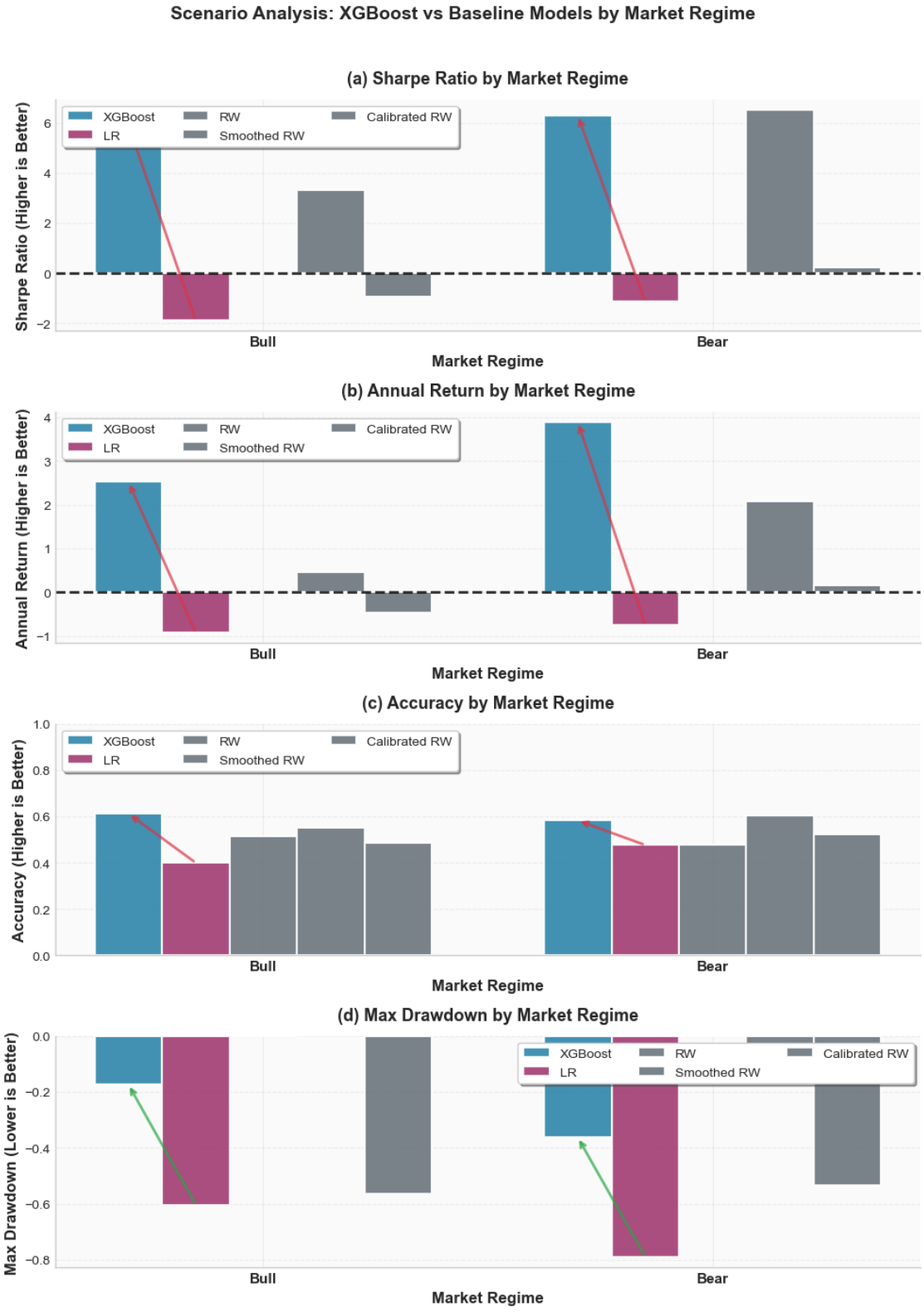
From the regime comparison plot: - **XGBoost**: Sharpe is clearly positive in both Bull and Bear (roughly **5.8** and **6.3**), and annual return is also positive in both (roughly **2.5** and **3.9**).
- **Risk**: max drawdown is around **-0.17** in Bull and **-0.35** in Bear, indicating higher risk during bear/down regimes.

0.7.2 6.2 Why to be cautious about "apparent regime robustness"

The same plot also shows some baselines with seemingly extreme values (e.g., unusually high Sharpe in certain regimes). Common reasons: - regime statistics not computed strictly with time-causal data (leakage risk);
- small subsample sizes causing Sharpe to explode;
- near-zero trading in certain segments distorting metrics.

Therefore, the correct use of this figure is to **surface robustness questions**, rather than to declare "the best model in Bear." XGBoost's value is its positive performance in both regimes, but it still must be judged jointly with trading frequency, costs, and drawdowns.

Key figure:


Scenario Analysis: XGBoost vs Baseline Models by Market Regime
(a) Sharpe Ratio by Market Regime
Sharpe Ratio (Higher is Better)
XGBoost
LR
RW
Smoothed RW
Calibrated RW
Bull
Bear
Market Regime
(b) Annual Return by Market Regime
Annual Return (Higher is Better)
(c) Accuracy by Market Regime
Accuracy (Higher is Better)
(d) Max Drawdown by Market Regime
Max Drawdown (Lower is Better)

0.8 7) Advanced Improvement: Optuna Hyperparameter Tuning

0.8.1 7.1 Observation: a "sharp" objective landscape

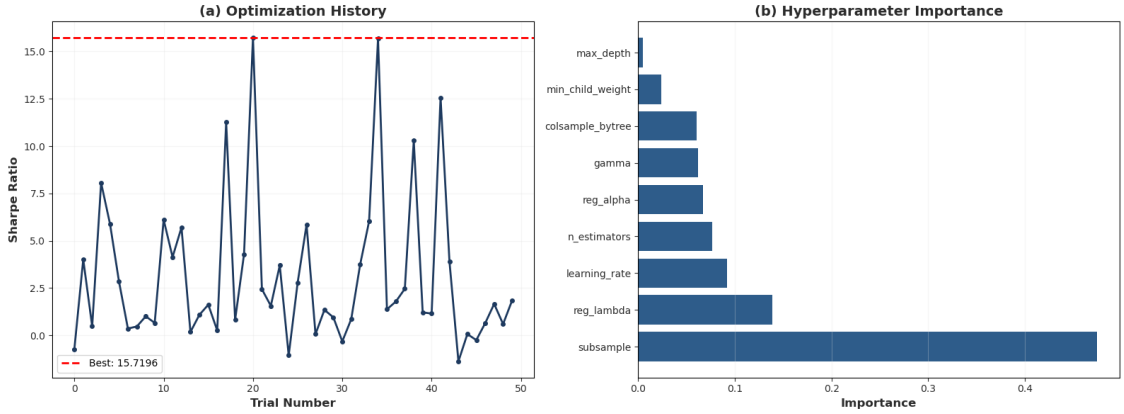
In the optimization history, Sharpe varies dramatically across trials, with a best trial Sharpe of **15.7196**. This suggests: - **High sensitivity to hyperparameters**: small changes can swing performance from negative to very high.
- **Potential window dependence**: if improvements come mainly from certain periods, tuning can amplify this dependence.

This implies we can indeed find better parameters, but the solution can also degrade easily if tuned "slightly off", making it more sensitive to definitions and robustness checks.

0.8.2 7.2 Which hyperparameters matter most

From the hyperparameter importance plot: - **`subsample`** dominates (**0.48**), indicating row sampling strongly affects generalization and noise suppression;
- followed by **`reg_lambda`**, **`learning_rate`**, **`n_estimators`**, etc.—the classic regularization / learning rate / tree-count trade-off.

Key figure:



0.9 8) Advanced Improvement: Dynamic Thresholding

The purpose is clear: **be more conservative under higher volatility** (higher long threshold / lower short threshold / wider no-trade zone), and **more active under lower volatility**.

Intuitively: when risk (volatility) is higher, reduce aggressive trading; when risk is lower, allow moderately higher activity.

0.9.1 8.1 How thresholds change with volatility (from the plot)

- The top-left panel shows piecewise thresholds over time: high threshold roughly **0.5–0.8**, low threshold **0.3–0.5**, compared to static thresholds (high 0.6 / low 0.4).

- The top-right panel shows thresholds vs rolling volatility: as volatility moves roughly **0.02–0.06**, thresholds shift upward with volatility (consistent with the design).

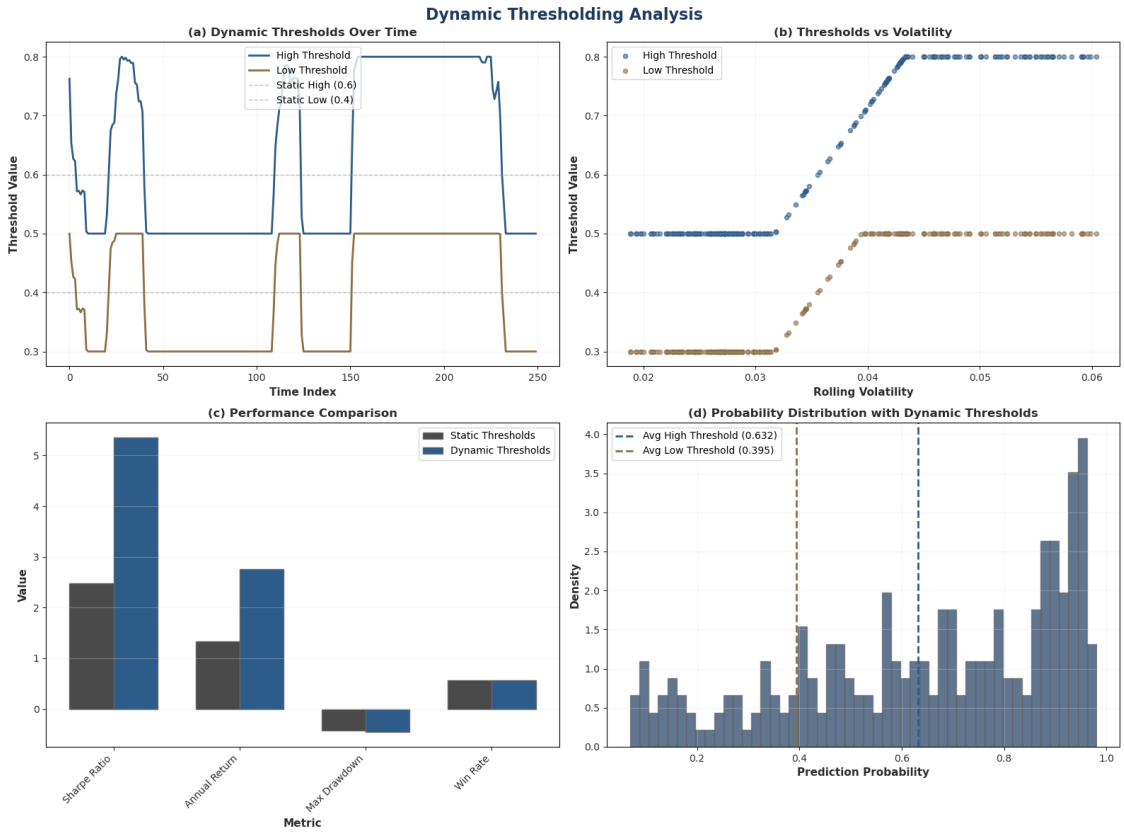
- The bottom-right panel reports average thresholds: Avg High **0.632**, Avg Low **0.395**.

0.9.2 8.2 Performance impact (reading the plot into conclusions)

The performance comparison indicates that, relative to static thresholds, dynamic thresholds show:
- **higher Sharpe** and **higher annual return**;
- **max drawdown not necessarily lower** (unstable changes, potentially slightly worse);
- **similar win rate** (improvement comes more from exposure control than hit-rate).

This matches a common financial intuition: dynamic thresholding is a form of time-varying risk control; it often improves return-to-volatility ratios, but does not guarantee lower drawdowns.

Key figure:



0.10 9) Optional Diagnostics: Regression/Correlation Metrics

Although the main task is direction classification, regression-style metrics help evaluate whether the model captures any structure in return magnitudes.

From the regression comparison plot (annotated values): - **XGBoost**: RMSE **0.0379** (best), MAE **0.0300** (best), correlation between predicted and actual returns **0.2753** (positive).
- **Smoothed RW**: higher correlation **0.3565**, but this does not necessarily imply more robust classification/trading performance.

This suggests XGBoost is more stable in error magnitude, while not necessarily maximizing linear

correlation—consistent with financial data where noise dominates and predictability is weak.

More intuitively, it is better at judging "directional tilt" than at precisely forecasting return size.

Key figure:

0.11 10) Conclusion

Based on the figures, the conclusion can be summarized in three statements:

- **Usability**: XGBoost has limited but non-trivial information content (AUC **0.623**). The strategy can amplify this into strong risk-adjusted returns (Sharpe **6.02**, annual return **339.95%**), implying that the signal-to-position mapping materially contributes to results.

- **Risk profile**: this is not a "low-risk steady arbitrage" strategy. Max drawdown reaches **-42.81%**, with more pronounced volatility/drawdowns later in the sample. Any "high annual return" interpretation must be accompanied by drawdown and trading frequency.

- **Effect of improvements**: Optuna and dynamic thresholding show potential further gains (best Sharpe **15.7196**; dynamic thresholds outperform static ones in Sharpe/annual return). However, these enhancements can be more sensitive to sample definitions and measurement choices, so stricter robustness validation is required before trusting the uplift.

The same conclusion in a more compact form: - The model contains signal, but it is not strong (AUC 0.623).
- Execution rules amplify the edge (Sharpe 6.02, annual return 339.95%), while drawdowns remain substantial (-42.81%).
- More complex tuning/dynamic thresholds may improve results further, but can be more sample/definition-sensitive and thus require stricter robustness checks.

Limitations that must be stated (to avoid over-interpretation): - High annual return / Sharpe can be overstated if transaction costs and slippage are not explicitly modeled.
- The displayed sample window (roughly 2024-01 to 2025-01) may not cover multiple full market cycles; conclusions should be framed as in-sample evidence for this period.
- Any regime split or rolling-statistic computation can introduce leakage if not handled with strict time causality; final interpretation should follow code-level hygiene auditing.