

Predicting Session Range Expansion in SPY: A Walk-Forward Study

Martial Systems LLC | Working paper | Not investment advice

Revisions. 2026-08-01: initial walk-forward diagnostics. 2026-08-19: research architecture freeze of Optuna trial 22. 2026-08-20: promotion-gate design (PSI, golden master, shadow, dual-control); GraphForge pin of fail-closed promote laws. 2026-08-20: paper hold-out ledger on post-2026-07-31 closes; path to live. Not live.

Diagnostics timestamp: 2026-08-19T15:50:17.075496+00:00. Out-of-sample window in tables below: 2019-01-02 to 2026-07-31 ($n = 1905$). Research architecture freeze dated 2026-08-19 (spec 01_optuna_trial22_lgbm_21d). Promotion-gate design dated 2026-08-20. Paper hold-out ledger on closes after 2026-07-31 (live=false). Related work: short-horizon realized-vol forecasts (production architecture frozen separately); open-to-close direction (rejected under pre-specified baselines).

1. Introduction

Let R_t denote the open-normalized session range of SPY on trading day t . We study a binary classification problem: whether R_t exceeds a trailing average of recent ranges. The target is deliberately not the sign of the open-to-close return. Sign prediction at the open was examined in a companion study and failed pre-specified accuracy and probability-score criteria relative to a constant long baseline; that line is not promoted.

The economic motivation is standard. Realized volatility and range measures exhibit serial dependence; directional returns at daily frequency are closer to a martingale increment after simple baselines. If next-session range is predictable after those baselines, the result is informative for risk forecasting even if the conditional mean return remains unpredictable.

We report walk-forward classification and ranking diagnostics only. We do not evaluate option strategies, stop-based breakout rules, or transaction costs of any trading rule. A separate production system forecasts multi-day realized volatility under a frozen specification; that system is not modified here.

2. Data and sample design

Prices are daily open, high, low, close, and volume for SPY (adjusted series from a public vendor API). Features observed at the close of day $t-1$ are used to form predictions for day t . An initial estimation sample ends 2018-12-31. Forecasts from 2019-01-02 onward are out of sample. Models are refit every 21 trading days on an expanding window of complete history strictly prior to the forecast origin. K-fold cross-validation with shuffled days is not used.

In the reported diagnostics, the out-of-sample size is $n = 1905$ with empirical expansion frequency 0.432 under the default label.

3. Label and features

Define $R_t = (H_t - L_t) / O_t$. The default label is $y_t = 1\{R_t > \mathbf{R}_{t-1}^{(10)}\}$, where $\mathbf{R}_{t-1}^{(10)}$ is the mean of R over the ten sessions ending on $t-1$. Equality is coded as non-expansion. Alternative labels used only for robustness replace the ten-day mean by a twenty-day mean or by the trailing sixty-day 70th percentile of R .

The baseline feature vector includes lagged range moments (one-day range, multi-horizon means, standardized range), ratios of recent range to prior average range, close-to-close returns and annualized realized volatility at five and twenty days, volume z-scores, Bollinger bandwidth and its trailing rank, a one-day lag of the overnight gap, and day-of-week indicators for the forecast day. Same-day high, low, and close of t are excluded from the feature set.

4. Models and baselines

We estimate L2-regularized logistic regression (standardized inputs) and gradient-boosted trees (LightGBM, with a histogram-gradient fallback if LightGBM is unavailable). Hard-rule baselines include always-expand, always-contract, persistence of the expansion event, a high trailing realized-vol rule, and a Bollinger-squeeze rule. Probability scores are

compared to a climate forecast that assigns the out-of-sample empirical expansion frequency to every day (for Brier score construction we use the realized OOS frequency as in the diagnostic export).

An ablation retains only returns, realized volatility, and a small range core (RV-only). The ablation is intended to test whether the full feature set improves on information already present in trailing volatility and recent range.

5. Pre-specified evaluation criteria

Criteria are module-specific. For open-to-close direction, the binding benchmark was a constant long classification. For range expansion, the binding benchmarks are (i) climate Brier score, (ii) persistence of the expansion event, (iii) the RV-only model, and (iv) rank correlation between predicted expansion probability and realized R. Zero-one accuracy against the majority class is reported but is not treated as the primary decision rule. Subsample rules that condition on $|P - 1/2|$ exceeding a threshold are reported as sensitivity checks only.

#	Criterion (primary)	Outcome
1	$\text{Brier} \leq \text{climate} - 0.002$	satisfied ($\Delta = -0.0316$)
2	Improve on persistence (accuracy or Brier)	satisfied
3	Improve on RV-only (Brier by ≥ 0.001 or acc by ≥ 0.5 pp)	satisfied
4	$\text{Spearman}(P, R) \geq 0.05$	satisfied (0.1642)

Under these four primary criteria jointly, the tree specification is recorded as **a research architecture freeze** (spec 01_optuna_trial22_lgbm_21d). That freeze authorizes documentation and a promotion-gate control plane. It does not authorize live promotion of a range model or any change to the production volatility architecture.

6. Out-of-sample results

Specification	Acc.	Brier	Climate Brier	AUC	$\rho_s(P,R)$
Frozen tree (trial 22)	0.6514	0.2138	0.2454	0.7076	0.1642
Logistic, full features	0.6577	0.2129	0.2454	0.7150	0.1345
Tree, RV-only	0.6514	0.2191	0.2454	0.6900	0.1674
Best hard rule (persistence)	0.5974	n/a	n/a	n/a	n/a
Persistence	0.5974	n/a	n/a	n/a	n/a
Always contract	0.5680	n/a	n/a	n/a	n/a

Both full-feature models improve on the climate Brier score and on persistence. The full-feature tree improves modestly on the RV-only tree in Brier and accuracy, so non-RV features contribute, but the RV-only model already improves on climate. In this sample the logistic model attains slightly lower Brier and higher accuracy than the frozen tree. The freeze still selects the tree: it was the pre-declared stage-1 architecture, and Optuna refinement was conducted on that family. Logistic remains a reported comparator, not the frozen recipe.

6.5 Architecture freeze (2026-08-19)

After a 40-trial Optuna search on the locked 21-day walk-forward, trial 22 (full features; n_estimators 300, learning_rate approximately 0.0209, num_leaves 26, min_child_samples 117, subsample 0.880, colsample_bytree 0.716, reg_lambda 0.156, reg_alpha 9.01) was re-evaluated with no further HPO. Out-of-sample Brier is 0.21378 versus 0.21675 for the named 2026-08-01 baseline (lift 0.00297, above the pre-declared 0.002 bar). Spearman(P, R) is 0.164. All four primary bars still pass. The 2026-08-01 tree is retained as a named baseline. The freeze file records live=false, auto_update=false, unfreeze_vol=false.

6.6 Promotion gate (2026-08-20)

A promote path, if it ever exists, must fail closed. Order: (1) PSI of freeze scores and of stationary features (range_z_20, atr10_ratio) on the last 63 trading days versus the prior 252; (2) golden-master invariance on 1000 static OOS rows (correlation at least 0.95, median absolute probability gap at most 0.05); (3) 48-hour shadow dual-score; (4) dual-control approvals from two distinct actors; (5) live write, which is refused while the freeze is not live. Raw range and RV levels are reported and are not hard blocks: they track the volatility regime. A clean pass is READY_FOR_HUMAN, not auto-update. Those laws are also encoded in GraphForge (rangeforge/) so an illegal transition cannot be silent.

6.7 Paper hold-out ledger (2026-08-20)

The 2019–2026 walk-forward selected the freeze. Closes after 2026-07-31 are therefore a true hold-out. A paper job scores the locked recipe at each new session close and appends one row to logs/paper_ledger.csv with date, predicted expansion probability P, model_id, recipe_hash, live=false, and scored_at. When the next session is observed, realized y and R are filled on the prior origin. The job prefers the SPY OHLCV parquet already written by the live vol pipeline; if that file is absent it re-ingests SPY causally. It is a separate GitHub Actions workflow (weekday 22:30 UTC, fifteen minutes after the vol cron) and is not a step inside the vol workflow. A range failure cannot take vol down. The paper job does not write frozen_best_config.json, the model registry, promotion artifacts, or a public live badge.

6.8 Path to a live range line (not this revision)

Live range would be a second production forecast, not a patch on vol. The freeze file stays live=false until all of the following hold. First, the paper ledger must contain enough post-2026-07-31 closes for Brier versus climate, persistence, RV-only, Spearman of P against realized R, and calibration to be interpretable on unseen days. If that hold-out is weak, the research freeze remains and the hold-out is documented; there is no live promote. If the four primary bars still clear, a shadow period on new closes (not a self-check on the original OOS) and dual-control approval would precede any flip of live=true. Daily scoring would remain an independent job after the vol close. A live board, if built, would report probabilities and Brier, not vol MAE. Joining frozen vol σ as a feature, and any option or breakout evaluation, remain out of scope until those decisions exist.

7. Additional diagnostics

Conditional mean range is 0.0136 on days classified as expansion and 0.0114 on days classified as contraction (unconditional mean 0.0122). Precision and recall for the expansion class are 0.610 and 0.535. Among the top quintile of realized R, the fraction that also fall in the top quintile of predicted probability is 0.281 (uniform random assignment would yield approximately 0.200).

Reliability diagrams (decile bins of predicted probability) are consistent with roughly monotone calibration. Conditioning on $|P - 1/2|$ larger than a positive threshold tends to raise accuracy and lower Brier as the threshold increases; those subsamples are not used as promotion criteria.

8. Crisis months

Month	n	Acc.	Persist.	Brier	Climate Brier
2020-03	22	0.5909	0.6818	0.2129	0.2479
2020-04	21	0.8571	0.7143	0.1379	0.1224
2022-06	21	0.5238	0.1905	0.2671	0.2449
2022-10	21	0.6190	0.5714	0.2066	0.2358

Month-level samples are small. Performance is uneven across 2020-03, 2020-04, 2022-06, and 2022-10. Satisfaction of full-sample criteria does not imply uniform performance in every stress window.

9. Label construction

We re-estimate logistic walk-forward models after replacing the default ten-day mean threshold with a twenty-day mean and with a sixty-day seventieth percentile. In each case the climate-Brier criterion remains satisfied in the diagnostic export. That is evidence against pure dependence on a single arbitrary threshold, not evidence that all range labels are equivalent for trading applications.

Label	Acc.	Brier	Climate Brier	Climate criterion
y_expand	0.6577	0.2129	0.2454	yes
y_expand_20	0.7013	0.1949	0.2408	yes
y_expand_p70	0.7654	0.1617	0.2080	yes

10. Limitations

- 1 Daily bars only; no residual range after the open on minute data.
- 2 No evaluation of options prices, implied volatility, or path-dependent breakout rules.
- 3 Vendor-adjusted OHLCV; microstructure and corporate-action conventions are not audited here.
- 4 Crisis-month inference is under-powered.
- 5 A multi-day realized-vol production model is maintained separately and is not jointly estimated.
- 6 The paper ledger after 2026-07-31 is still short; it is not yet a substitute for the 2019–2026 walk-forward.

11. Conclusion

Under the pre-specified primary criteria, next-session range expansion for SPY is statistically distinguishable from climate, persistence, and a narrow RV-only feature set in 2019–2026 walk-forward evaluation. The result is recorded as a research architecture freeze of LightGBM trial 22, not as a live product. Promotion is gated and cannot write live. New closes are scored in paper mode into a ledger so that a later hold-out evaluation is possible; that is not a live product. This is not a recommendation to trade, and it does not alter the frozen multi-day realized-vol architecture. Open-to-close direction remains a rejected specification under its own criteria.

12. References (internal)

RANGE_DESIGN.md; FREEZE_DECISION.md; PROMOTION_GATE.md; OPTUNA_RESULT.md; logs/range_diagnostics.json; logs/experiments/frozen_best_config.json; logs/paper_ledger.csv; spy_vol_forecast methodology note and FREEZE_DECISION.md; spy_direction_intraday FREEZE_DECISION.md and direction null note; MODULE_EVAL_PATTERN.md.

Disclaimer

This document is for research documentation. It is not investment advice, not a solicitation to buy or sell any security or derivative, and not a performance record of live trading. Copyright (c) 2026 Martial Systems LLC. All rights reserved.