Box A.5: Walk-Forward Cross-Validation

The walk-forward evaluation framework provides a time-robust assessment of model performance that a single random end split cannot. While a conventional 80/20 (or 75/25) split is leakage-free, it concentrates evaluation in *one recent window* it does not establish whether the model generalizes across time. To address this, a walk-forward cross-validation regime was adopted. The walk-forward protocol executes *a sequence of chronologically ordered train–test experiments*, evaluating the model across several market states rather than a single recent period. While this approach may lower headline hit rate, it reduces false confidence and error risk, yielding *more reliable performance estimates*. It also reveals regime sensitivity – folds that span macro or event shocks (e.g., tariffs, wars, Federal Reserve actions) often show dips in hit rate – information that a single split would average away.

In this project, folds use approximately one trading year for training, one quarter for testing, and advance by roughly one trading month between folds. An embargo equal to the prediction horizon removes rows whose forward returns would overlap the test labels, eliminating look-ahead via overlapping targets. The selection–pruning blueprint governs each fold, including leakage-safe, model-based pruning in training (see Box A.4)

Evaluation is reported per fold and in aggregate

(A) Directional accuracy (hit rate) measures the share of test-set days where the sign of the prediction matches the realized sign at the chosen horizon.

(B) Spearman's ρ captures the monotonic alignment of scores with outcomes.

(C) The conditional long-only return (`strat_ret`) reports the average realized return on the subset of test days for which the model's prediction is positive.

For each metric the cross-fold mean, and its variability are provided, while the most recent fold is highlighted as the operational proxy because it is trained on the largest history and assessed in the latest market regime. *In equity forecasting* at ten-day horizons, mean hit rates in the low-to-mid fifties are typically commercially relevant when stable, Spearman values above roughly 0.10 are notable, and conditional long-only returns that materially exceed unconditional averages indicate useful selection before costs.

Overall, the walk-forward regime – combined with path-aware feature scoping, deterministic prefilters, leakage-safe model-based pruning, and a compact hybrid XGB model—produces an evaluation that is both realistic and defensible. It measures *generalization across changing regimes*, enforces *timing integrity* through embargo and warm-ups, and offers *two complementary perspectives on performance*: while the last fold uses the most history and tests on the most recent regime – providing a benchmark for live short-term deployment – the cross-fold mean hit, averaged across many past regimes, supplies a long-run baseline when the future state is unknown.