White Paper - Technical

Simplified version available here: White Paper - Simplified

1. Introduction

This document outlines a classification based machine learning pipeline for forecasting fund directionality. The framework unifies historical data ingestion, feature engineering, model training, regime identification, and prediction generation within a continuous, end-to-end workflow. Its primary objective is to produce robust, multi-horizon forecasts with measurable confidence. This design supports the implementation of a swing trading strategy, alongside an associated risk management framework described in the following sections.

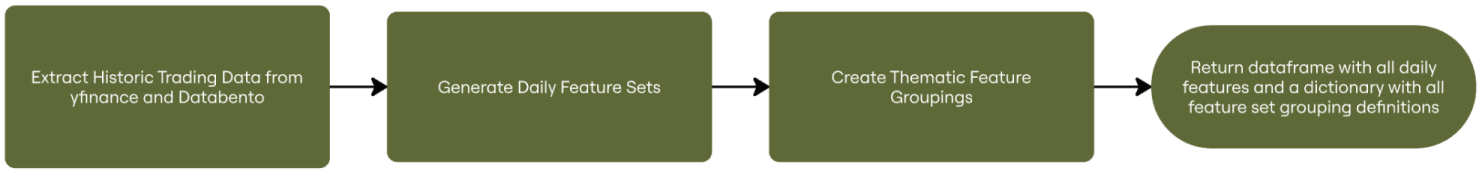
Certain elements of the methodology are intentionally abstracted or omitted to preserve the proprietary nature of the approach and safeguard its competitive advantage, limiting the potential for direct replication of its core components.

2. Data Extraction

The pipeline begins with the extraction of historical trading data from external providers, including Yahoo Finance (via yfinance) and Databento. From this data, a comprehensive set of features is derived using five core market variables: Open, Close, High, Low, and Volume.

Data is collected across multiple temporal resolutions, including daily, hourly, and minute intervals, and is stored in a structured format to support downstream processing. To ensure reproducibility across runs, particularly for features that rely on incremental calculations, a consistent start date is applied during feature generation.

3. Feature Engineering



Daily feature sets are derived from the underlying price and volume data through a series of systematic transformations. These features span multiple categories, including:

- Technical indicators (e.g., moving averages, relative strength, trend signals)
- Volatility and dispersion metrics
- Momentum and rate-of-change indicators
- Volume based and flow derived signals

Feature construction applies rolling window operations, relative positioning measures, and normalization techniques to ensure comparability across time. Where appropriate, features are scaled or standardized using rolling statistics to stabilize distributions and improve model interpretability.

3.1 Feature Set Groupings

Constructed features are organized into thematic groupings based on shared methodology and informational content, including trend, momentum, volatility, volume, and exogenous signals. A structured mapping is maintained in a dictionary that defines:

- Feature group identifiers
- Corresponding feature column sets within each group

Feature Group (Dictionary Key)	**Feature Name (Reference Columns)**
Moving Average	200 Day Simple Moving Average (SMA)
Moving Average	50 Day SMA / 200 Day SMA
Volume	On Balance Variance
Volume	Cumulative Positive Day Volume
Duration	Time since breached 200 Day SMA
Duration	Time since 180 day minimum

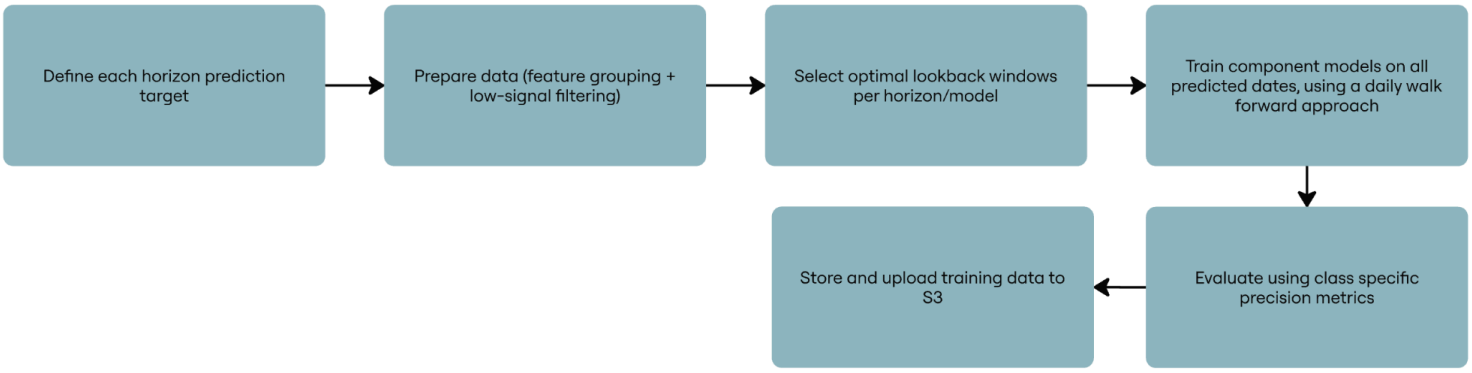
This modular design enables flexible model training, controlled experimentation, and systematic evaluation of feature subsets without altering the underlying data pipeline. A primary motivation for grouping similar feature types is the premise that combining features with aligned informational characteristics enhances signal stability, while reducing noise. By limiting interactions between too many highly heterogeneous features, which may produce conflicting indications, the framework promotes more consistent and interpretable model behavior across different conditions.

Component models are then compiled and trained independently using restricted feature subsets spanning two to three thematic categories. This design promotes specialization within models while reducing overfitting risk associated with overly broad feature inclusion that was seen to degrade out of time generalization and reduce performance.

Outputs:

- Consolidated feature dataframe (currently comprised of more than 500 features)
- Feature grouping dictionary mapping (currently comprised of more than 13 groupings)
- Component models (currently comprise of more than 60 combinations)

4. Model Training



The system employs a walk-forward training framework with dynamically selected lookback windows tailored to each feature grouping and forecast horizon. Lookback periods are optimized for each horizon and feature set and training is optimized using balanced accuracy to ensure performance is not biased toward the majority class, enabling more reliable signal generation across both directional outcomes. To reduce noise from low signal environments, the smallest 5% of positive and negative return observations are excluded from the training data, removing periods where price movement is minimal and less informative for learning directional structure. The entire modeling pipeline is retrained on a daily cadence, allowing it to continuously adapt to evolving market conditions.

4.1 Multi-Horizon Approach

The system is built around a multi horizon forecasting structure, where independent models are trained to predict outcomes at different future time intervals. Each horizon defines a distinct distance from the prediction date based on how far ahead the outcome is measured. For example, the 5-day horizon is trained to predict market direction five days into the future, while the 30-day horizon is trained to predict direction thirty days ahead.

This separation ensures that each model learns patterns specific to its forecast horizon, rather than attempting to generalize across all time scales. Shorter horizons emphasize higher-frequency dynamics and are more sensitive to rapidly evolving features, whereas longer horizons prioritize slower-moving structural relationships and regime level signals.

Each horizon operates as an independent modeling layer within the broader system, with its own training data alignment, feature interactions, and performance characteristics. This design allows the system to maintain specialized predictive behavior across multiple time scales while producing a unified ensemble output during inference.

Note: The shortest horizon meeting stability criteria for deployment was the 5-day model. Shorter horizons were evaluated, but no consistent patterns were identified from the underlying feature sets that produced sustained outperformance, defined by both individual class precision metrics exceeding 0.6 across a sufficient number of component models to justify inclusion.

4.2 Training Process

For all forecast horizons, a set of component models is constructed and tested. Prior to deployment eligibility, each model is trained on a rolling historical window of at least 484 trading days (approximately two years). This training period provides a sufficient basis for evaluating model performance and determining both its suitability for inclusion and the conditions under which it should be applied. This window length is selected to balance sufficient exposure to prevailing market structure with responsiveness to more recent dynamics.

The system incorporates multiple component models derived from distinct combinations of feature groupings. This introduces heterogeneous but controlled perspectives on market behavior, capturing distinct dimensions such as trend, momentum, volatility, and flow. The design avoids excessive reliance on any single feature representation while also constraining unnecessary cross-interaction between highly disparate feature sets within individual models.

Empirical results from internal validation studies indicate that overly broad feature aggregation can reduce out-of-sample performance and decrease responsiveness to regime changes. In contrast, smaller and more targeted component models that are stacked demonstrate improved stability and predictive consistency. This supports a modular architecture in which specialized models operate over constrained feature subsets rather than a single monolithic representation.

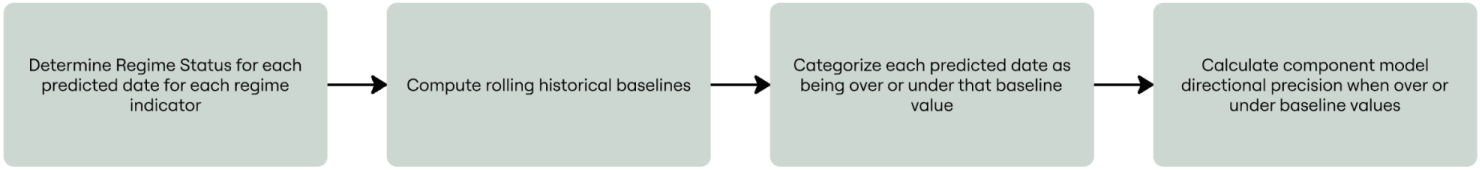
4.3 Evaluation

Model performance is evaluated over the full training window using class-specific metrics, with particular emphasis on precision for both positive and negative outcomes. Precision is prioritized over recall and composite measures such as the F1 score, reflecting a preference for lower prediction frequency with higher conditional accuracy.

This evaluation framework ensures that performance is assessed in a way that aligns with the decision objective, where false positives are more costly than reduced coverage. Class-level evaluation also enforces balanced discriminative performance across both outcome classes, reducing bias toward a single directional signal. This is particularly important given the asymmetric nature of market movements, where declines tend to occur much more rapidly, albeit less frequent, than advances. As a result, bias toward the positive majority class can materially increase downside risk.

This structure also supports a consistent assessment across varying market regimes, which are further analyzed in the subsequent section.

5. Regime Detection



Market conditions vary over time, making regime recognition a critical component for adaptive system performance. The regime detection framework characterizes prevailing market states using the historical behavior of selected indicators and enables context aware model selection through a structured routing mechanism.

5.1 Methodology

To detect regimes, the system uses a subset of features identified as having strong segmentation power on model performance. For these features, rolling historical baselines are computed to serve as reference distributions.

At each prediction date, current regime feature values are evaluated relative to these baselines to determine their position within a longer term historical context. This comparison provides a normalized measure of deviation from typical behavior, which is used to infer the prevailing market regime, as well as determine which component models perform best under those conditions.

5.2 Regime Classification

Based on this comparison, each regime feature is assigned to one of two regime states:

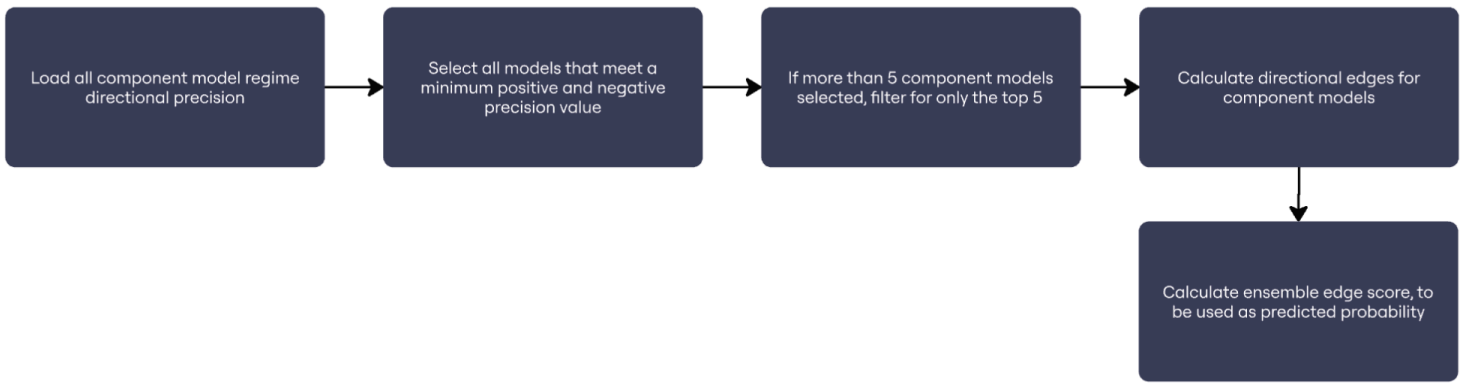
- **“O” (Over):** Current value exceeds a statistically derived threshold associated with differentiated performance behavior
- **“U” (Under):** Current value is below the corresponding threshold

Regime Feature	Baseline Value	Predicted Date	PD Value	Regime Status
200_SMA	1.03	2026-01-01	1.12	Over
200_SMA	1.03	2026-04-01	0.98	Under
VIX	18.2	2026-01-01	14.3	Under
VIX	18.2	2026-04-01	27.2	Over

These thresholds are selected based on their demonstrated ability to separate model performance across historical observations while maintaining a relatively balanced number of samples on either side of the cutoff. This ensures that regime definitions are both empirically grounded and statistically stable.

This binary classification framework provides a consistent and interpretable representation of market context. It enables a regime-aware routing process in which only models that have historically performed well under similar conditions are activated, while those with degraded performance in the current regime are excluded from consideration.

6. Model Selection and Ensemble Construction



Not all trained component models are retained for production inference. A model selection layer is applied each predicted date to ensure that only sufficiently robust and consistently performing models contribute to downstream predictions for that day. This selection process is conditioned on the regime detection framework, allowing model eligibility to vary dynamically with market context.

Selection is based on multiple stability and performance criteria, including:

- Historical predictive thresholds
- Consistency of performance across both positive and negative classes
- Stability across specific forecast horizon

This filtering process produces a reduced, high-confidence subset of models eligible for inclusion in the production ensemble. For each forecast horizon, retained models are ranked using performance based metrics and a dynamic subset is selected per evaluation date. Model outputs are filtered by minimum confidence thresholds to reduce noise from marginal signals.

7. Prediction and Inference Pipeline

The inference pipeline executes the selected ensemble system over time to generate and store production ready forecasts.

7.1 Input Preparation

The pipeline initializes by loading the outputs of the regime detection layer along with the filtered set of models produced by the model selection stage. These inputs define the active model universe for each predicted date and ensure that prediction generation is conditioned on the prevailing market regime for each predicted date.

7.2 Prediction Scope Definition

Forecasting is structured by horizon, with predictions generated across a rolling set of predicted dates whose outcomes have not yet been realized. Each horizon defines the forward looking window over which predictions are made. For example, the 5-day horizon predicts outcomes five days ahead, while the 30-day horizon predicts outcomes thirty days ahead.

At any point in time, each horizon maintains a set of active, unresolved predictions. For the 5-day horizon, this includes predictions made at different lead times over the previous five days that have not yet reached their outcome date. The 5-day horizon models have a prediction made five days ago that expires today, four days ago that expires in one day, three days ago that expires in two days and so on. At market close, a new prediction will be made that expires in five days time, while all other predictions get one day closer to actualizing.

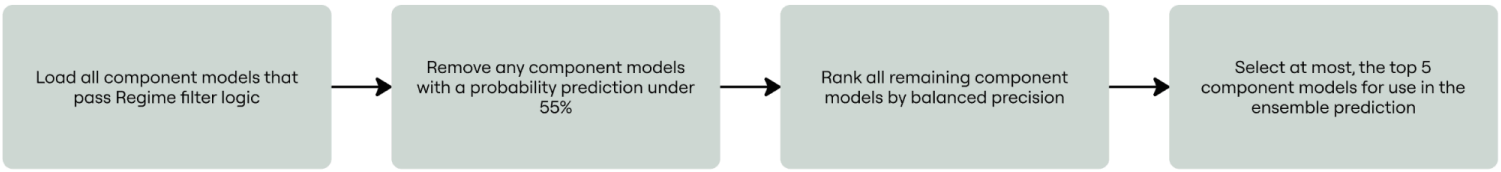
The same structure applies to the 30-day horizon, which maintains thirty active predictions ranging from those made thirty days ago (expiring today) through to those made yesterday day (expiring in twenty nine days).

Horizon	Current Date	Predicted Date	Days to Expiry
5	2026-04-01	2026-03-31	4
5	2026-04-01	2026-03-30	3
5	2026-04-01	2026-03-28	1
30	2026-04-01	2026-03-31	29
30	2026-04-01	2026-03-15	15
30	2026-04-01	2026-03-02	1

7.3 Model-Level Prediction Generation

For each combination of predicted date and horizon, all eligible models are executed to produce individual prediction outputs. These outputs include directional probabilities along with associated performance metrics and feature-level metadata. Predictions are retained at the model level to preserve traceability prior to aggregation.

7.4 Ensemble Construction and Filtering



Model outputs are filtered using a confidence threshold to remove low-conviction signals. For each date, models are ranked by a predefined lookback period precision score and only the top performing subset of models are retained. Further, if more than 5 component models pass through the prevailing filtering logic, the number of models that are used for the ensemble prediction is reduced to the top 5, as measured by balanced precision.

Predictions are transformed into directional edge values, defined as the directional precision values minus 0.5, and aggregated across models to form a single ensemble signal per date and horizon. This produces a unified prediction score representing consensus strength across the active model pool.

7.5 Output Storage and Persistence

Final ensemble outputs, along with underlying model level predictions, are appended to a historical prediction store indexed by date and horizon. These records are persisted for downstream analysis, monitoring, and performance evaluation, enabling continuous tracking of both model and ensemble behavior over time.

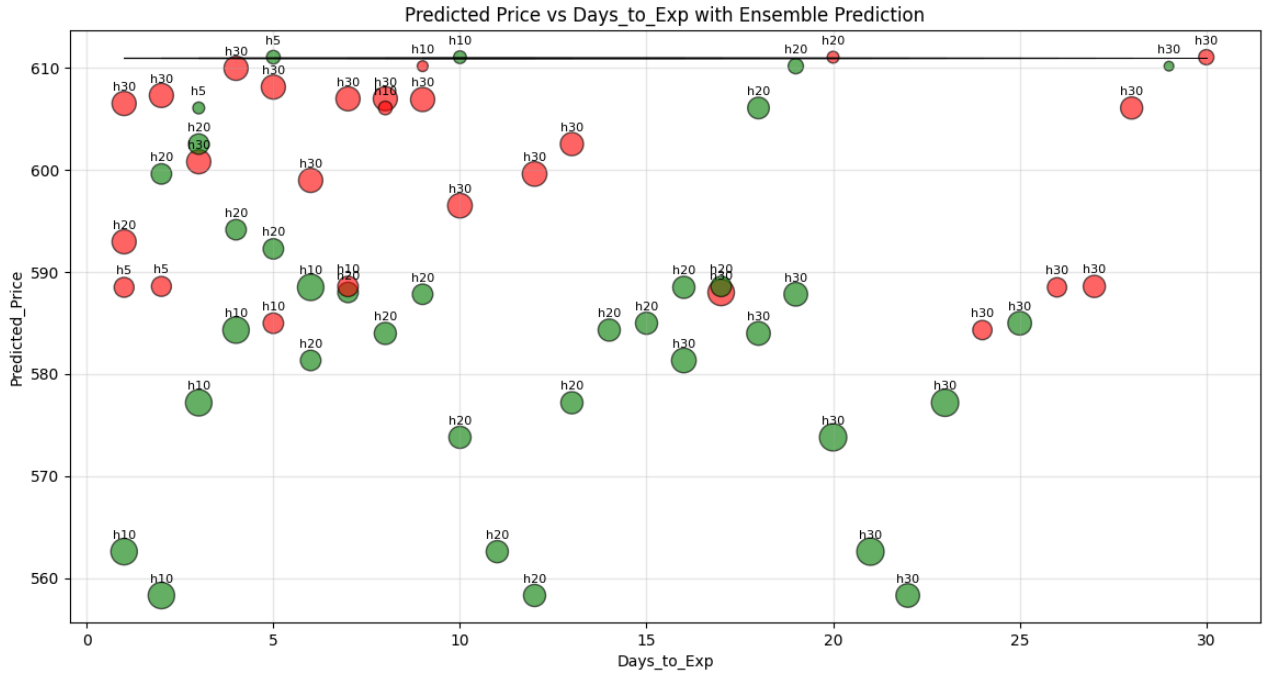
8. Visualization and Metrics Tracking

This stage provides both interpretability of ensemble forecasts and ongoing performance monitoring across horizons and time.

8.1 Global Ensemble Visualization

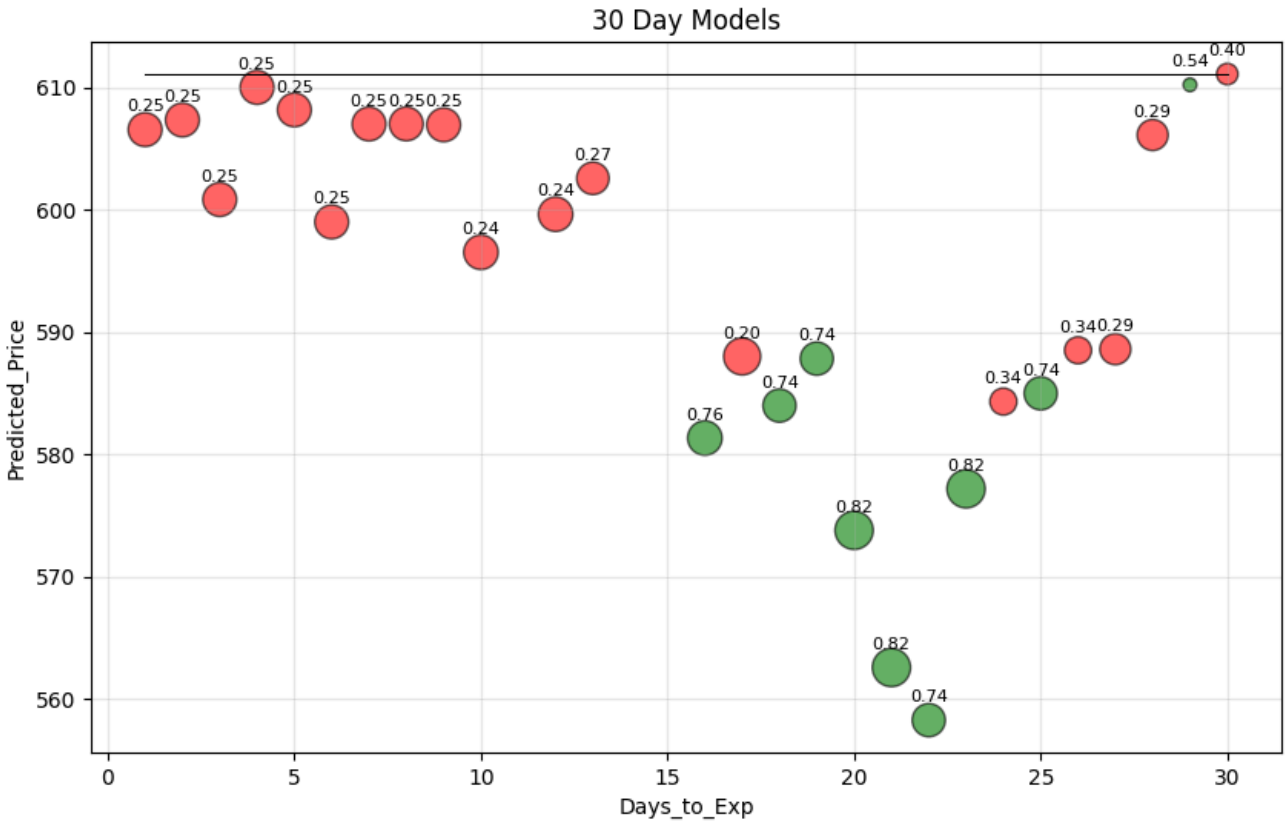
A consolidated visualization is generated to present ensemble predictions across all forecast horizons within a unified framework. Predicted values are plotted against days-to-expiry, with bubble size encoding signal strength and color encoding predicted direction. The y-axis shows the price the model expects the asset to be above or below when that prediction expires.

The last close is overlaid as a reference baseline using a black line. Each point is annotated with its corresponding forecast horizon, enabling direct comparison of short- and long-horizon predictions within a unified visualization framework.



8.2 Horizon-Level Visualization

For each forecast horizon, a dedicated plot is generated showing predicted price paths against time-to-expiry. Each point is annotated with the ensemble prediction score, providing a direct view of model confidence evolution across the horizon, based on the aggregated directional edge of the component models used that day. The same market baseline is included to anchor predictions relative to observed price levels.



8.3 Rolling Performance Evaluation

Model performance is evaluated over multiple rolling windows using fixed cutoffs, producing a time-sensitive view of model reliability under different lookback periods. For each horizon and cutoff window, predictions are merged with realized returns and filtered into directional classes based on ensemble predictions.

Performance is measured separately for positive and negative predictions:

- Positive precision (pprec): accuracy of upward signals
- Negative precision (nprec): accuracy of downward signals
- Support counts (pcnt and ncnt): number of evaluated predictions in each class

Below we see, over the last 10 days, the 30 day horizon (h30) models only had 7 predictions that were deemed strong enough to deploy, of which all were negative predictions. Nprec was 0.857, telling us 6/7 of those predictions were correct.

Over the last 255 days, 239 predictions were deemed strong enough to deploy, with 172 being for the positive class and 65 for the negative class. The positive class predictions were correct 89.5% of the time and the negative class predictions were correct 81.5% of the time.

```
h30 Performance Last 10 Days | pprec: nan | pcnt: 0 | nprec: 0.857 | ncnt: 7
h30 Performance Last 32 Days | pprec: 0.0 | pcnt: 4 | nprec: 0.917 | ncnt: 24
h30 Performance Last 53 Days | pprec: 0.455 | pcnt: 11 | nprec: 0.912 | ncnt: 34
h30 Performance Last 85 Days | pprec: 0.562 | pcnt: 32 | nprec: 0.854 | ncnt: 41
h30 Performance Last 150 Days | pprec: 0.812 | pcnt: 96 | nprec: 0.854 | ncnt: 41
h30 Performance Last 255 Days | pprec: 0.895 | pcnt: 172 | nprec: 0.815 | ncnt: 65
```

8.4 Performance Output Logging

For each horizon, performance summaries are written to text files containing precision metrics and sample counts for each cutoff window. These outputs are stored as persistent artifacts and updated incrementally as new predictions are added. A final aggregation step uploads both chart and performance artifacts for external monitoring and review.

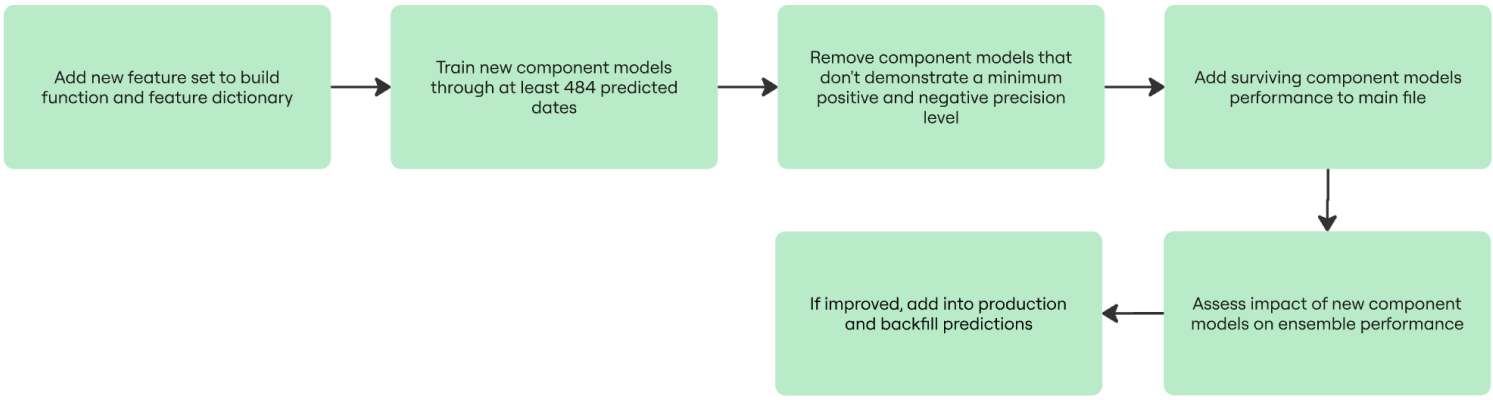
As of April 10th, 2026, over the last 255 days performance for each ensemble horizon model was as follows:

Horizon	Positive Precision	Negative Precision	Coverage
5	73.8%	69.0%	231/255
10	82.6%	79.5%	251/255
20	89.2%	85.3%	254/255
30	89.5%	81.5%	237/255

8.5 Incremental Prediction Storage

As new ensemble outputs are generated, they are merged into horizon-specific historical datasets. Only previously unseen predicted dates are appended, ensuring the dataset remains deduplicated while continuously growing over time. These stored outputs serve as the foundation for downstream evaluation and backtesting.

9. Scalability and Performance



The system is designed to support incremental expansion of both features and models without requiring changes to the core inference or evaluation pipeline. New components can be introduced in a controlled, modular manner while preserving backward compatibility with existing ensemble logic.

9.1 Feature and Model Extension

Scalability begins with the creation of new feature sets, which are added to a centralized feature dictionary. These feature sets define the inputs for newly introduced component models and allow them to operate alongside existing model families without structural changes to the pipeline.

9.2 Component Model Training

For each new feature set, a corresponding set of component models is trained using the walk-forward approach. Each model is required to be trained on a minimum history to ensure sufficient exposure across regimes and market conditions. The training history also overlaps with the existing component models to ensure a direct 1:1 comparison in performance. These steps provide consistency with the training standards applied to existing models.

9.3 Model Selection and Integration

Newly trained models are evaluated using the same performance and stability criteria as the existing model pool. Only models that meet minimum performance thresholds and demonstrate consistent behavior across evaluation windows are selected for inclusion in downstream inference.

9.4 Backfilling and Historical Alignment

Once validated, new component models are backfilled into historical performance datasets. This ensures that all evaluation files remain complete and comparable across model generations, enabling consistent benchmarking against legacy models.

9.5 Ensemble Re-Evaluation Under Regime Logic

After integration, the full ensemble pipeline is rerun with regime-aware selection logic enabled. This step determines where and when newly introduced models are eligible for inclusion based on market context. The impact of new models is assessed at both the ensemble level for each horizon to ensure compatibility with existing structure.

9.6 Performance Impact Assessment

The updated ensemble outputs are evaluated to determine whether the addition of new models improves predictive performance in production, as was indicated in preliminary testing. Metrics are compared against baseline ensemble behavior across multiple rolling windows and cutoffs. Based on this evaluation, models are either retained in the production ensemble or discarded if they unexpectedly degrade stability and predictive accuracy.

This process ensures that system scalability is governed by measurable performance impact rather than model accumulation, maintaining both extensibility and robustness over time.

10. Risk, Capital Control, and Model Limitations

This section defines the system’s capital constraints, signal validation rules, and loss management framework, along with key structural limitations that affect performance under non-stationary market conditions. The approach leverages a swing trading strategy, which aims to minimize the number of trades made, while maximizing profit. On average, a position is built up over 2-3 days and held for 3-7 days before exiting.

10.1 Capital at Risk (CAR) Management

Capital at Risk (CAR) defines the maximum proportion of capital exposed to active directional positions at any time. The system targets a CAR range of 70–85%, with a cap of 90% that is rarely exceeded.

CAR is dynamically adjusted based on system conditions. Exposure is reduced when:

- Cross-horizon model disagreement increases
- Model confidence dispersion widens across the ensemble
- The expected outcomes have been realized prior to expiration

This mechanism ensures that capital exposure contracts during periods of elevated uncertainty or regime instability.

10.2 Position Construction and Exposure Scaling

Position sizing is based on an Expected Value Weighted Exposure (EVWE) framework rather than fixed drawdown rules. Exposure increases only when expected value improves, incorporating:

- Model confidence
- Conditional historical accuracy at forecast distance
- Stability of current regime
- Trend persistence strength

Because the shortest deployable horizon is 5 days, model predictions may indicate a reversal before short term price action has fully turned. In practice, this creates a timing mismatch where momentum can persist in the opposite direction of the forecast in the early days following the initial prediction. To manage this, initial position sizing is constrained to a maximum of 33%. This allows exposure to the signal while limiting risk during the early phase of the prediction, with additional capital deployed only as subsequent predictions provide confirmation.

Condition	Distance from Original Cost Basis	Exposure Tier
Initial signal, confidence alignment across multiple horizons	0%	33%
Newer predictions continue to support position + strong recovery stats	-1.0%	50%
High historical reversal accuracy and continued agreement between horizon models	-2.5%	67%
Deep adverse move driven by expected temporary conditions with newest predictions still expecting a reversal	-4.0%	85%
LONG ONLY	-7.5%	90-100%

Higher allocation is only justified when supported by strong model agreement and historically validated reversal or continuation behavior.

10.3 Signal Gating and Model Trust

This section governs whether a trade is initiated and how it is sized. A signal is admissible only if it meets minimum ensemble quality thresholds:

- Minimum ensemble confidence (≥ 55%)
    - Ensemble probabilities are structurally dampened by model disagreement and typical balanced precision being in the 60-65% range for any single model. As a result, extreme high confidence consensus across all deployed models for a given predicted date is rare.

    - Backtesting shows that increasing the threshold to 65% improves ensemble balanced precision by ~2–3 points but reduces signal frequency by ~15%, while 75% yields an additional ~3–4 point gain at a ~30% reduction in coverage.
    - A 55% threshold is therefore used to balance signal quality and opportunity set breadth.
- Cross model and cross horizon alignment
    - Signals must exhibit directional consistency across models and overlapping forecast horizons. Conflicts indicate weak or unstable predictive structure and are filtered out at entry.
- Directional stability at the horizon level
    - Signals are excluded if horizon specific models already exhibit instability or recent directional flipping.

Position sizing (conditional on admissibility):

- Weighted by historical directional stability
- Weighted by probability consistency (low variance in predicted probabilities)
- Increased with cross model and cross horizon agreement
- Downweighted in high disagreement or shock sensitive regimes

10.4 Loss Management and Exit Logic

This section governs how positions are managed after entry. Risk is monitored across three dimensions: price movement, signal integrity, and time decay.

Partial reductions are triggered by deterioration in signal quality:

- Ensemble confidence drops below the entry threshold
- Predictive distributions show rising entropy (loss of conviction) or reversals
- Horizon level models weaken or begin to diverge

Full exits are triggered by structural invalidation:

- Broad model consensus shifts direction (true reversal, not noise)
- Cross horizon agreement breaks down in a sustained manner
- Time decay invalidates the original trade thesis without renewed confirmation

Time-based constraint:

- Each position has a maximum holding period
- Positions exceeding this duration require full revalidation under current ensemble conditions

10.5 System Limitations and Structural Risks

Despite layered controls, the system remains exposed to inherent risks:

- Model overfitting: Even with walk-forward validation, models may adapt too closely to historical regimes.
- Regime Classification Volatility: Rapidly changing regime assignment can activate inconsistent models.
- Market structural shifts: Non-stationary dynamics may invalidate learned relationships over time or influences outside of the model's scope can materially change the outlook.
    - I.e. Rapid inflation during covid, tariffs, war, economic condition deterioration, rapidly rising borrowing costs, etc

These risks are partially mitigated through ensemble redundancy, regime conditioning, continuous retraining, and monitoring, but cannot be fully eliminated in dynamic market environments.