Yield Curve Inversion as an Early Warning System for Recession Risk

Sajan Arora, Karan Badlani
Khoury College of Computer Sciences,
Northeastern University, Boston, MA, USA
arora.saj@northeastern.edu, badlani.k@northeastern.edu

Abstract

Yield-curve inversion, and in particular the spread between long- and short-term US Treasury yields, is widely monitored as an early warning signal for macroeconomic stress. This paper builds a data-driven system to forecast yield-curve inversion itself, rather than recessions, at a 12-week horizon. Using weekly 3M–10Y constant-maturity Treasury yields from 1981 to 2025, we construct slope and curvature measures, extract level–slope–curvature factors via principal component analysis, and discover interpretable "steep", "normal", and "flat/inverted" regimes using K-means clustering. Johansen cointegration analysis provides error-correction terms that capture long-run cross-maturity relationships. We then compare several forecasting approaches, including ARIMA and vector error-correction models (VECM) for the 10-year yield and logistic regression and histogram-based gradient boosting for 12-week-ahead inversion classification, against simple spread-threshold rules. On a held-out period with rolling-origin evaluation, the calibrated gradient boosting model achieves an AUC of about 0.88 and a Brier score around 0.18 for inversion prediction, substantially outperforming the logistic baseline while preserving interpretability through regime and factor diagnostics.

Introduction

The shape of the U.S. Treasury yield curve reflects market expectations about future economic conditions, and past research has shown that periods in which short-term interest rates exceed long-term rates—yield-curve inversions—often precede episodes of financial stress. While yield-curve inversion is widely monitored, an open question is how early such an inversion can be predicted from the evolving term structure itself.

In this work, we develop a data-driven framework to forecast yield-curve inversion at a twelve-week horizon using weekly U.S. Treasury yields from 1981 to 2025. We combine classical term-structure tools with modern feature engineering, including principal-component factors (level, slope, curvature), shape-based regime clustering, and cointegration-based error-correction terms. These features are used to construct supervised learning datasets for both yield-level forecasting and inversion-probability estimation.
Our main contributions are:

- We build a reproducible pipeline for FRED yield-curve data collection, preprocessing, factor extraction, and regime identification.
- We design and evaluate several forecasting models—ARIMA, VECM, logistic regression, and calibrated histogram-based gradient boosting—for predicting 10Y levels and 12-week-ahead inversion of the 10Y–3M spread.
- We conduct systematic time-aware evaluation using rolling-origin splits, reporting MAE/RMSE for yield-level forecasts and AUC, PR-AUC, and Brier score for inversion prediction.

This framework provides a practical and interpretable early-warning system for inversion risk, demonstrating that regime-aware and error-correction features meaningfully improve predictive performance over simple slope-based thresholds.

Background and Related Work

A large empirical literature documents the predictive content of the yield curve for future economic conditions. Classic studies such as Estrella and Mishkin [1] show that simple term spreads, particularly the 10Y–3M spread, contain significant information about upcoming recessions and turning points. Subsequent work has formalized the yield curve using parametric term-structure models (e.g., Nelson–Siegel and Diebold–Li) that summarize the curve through level, slope, and curvature factors.

Other strands of research emphasize the long-run cointegration relationships among yields of different maturities and model their joint dynamics using vector error-correction models (VECM). These approaches capture equilibrium deviations that can precede structural shifts in the curve. More recent studies explore machine learning methods for yield-curve forecasting, including tree-based models and regularized classifiers, which often outperform simple threshold rules in probabilistic calibration and early-warning tasks.

Our work builds on these ideas by combining factor-based representation, regime clustering, and cointegration-driven error-correction terms within a unified forecasting framework aimed at predicting yield-curve inversion rather than recession events directly.

Data and Preprocessing

Data Sources:
We use weekly U.S. Treasury constant-maturity yields obtained from the Federal Reserve Economic Data (FRED) API. The raw dataset includes yields at the following maturities: 3-month (3M), 1-year (1Y), 2-year (2Y), 5-year (5Y), 10-year (10Y), and 30-year (30Y). Observations span from 1970 to 2025, but missing values in the short end prior to the early 1980s require constructing a clean baseline panel beginning in 1981.

For each weekly observation, we also compute commonly used slope measures such as the 10Y–3M and 10Y–2Y spreads, which play a central role in identifying yield-curve inversion dynamics. No macroeconomic or recession labels are incorporated, as the goal of this work is to forecast inversion of the yield curve itself rather than downstream economic outcomes.

TABLE I: Data overview (baseline panel and targets).

Source	FRED (constant-maturity Treasuries)
Frequency	Weekly (Friday)
Maturities (baseline)	3M, 2Y, 5Y, 10Y
Date range (baseline)	1981-09-04 to 2025-11-07
Baseline shape	2,306 rows $\times$ 4 columns (yields)
Core spreads	10Y–3M, 10Y–2Y (plus 5Y–3M, 10Y–5Y)
Forecast horizon	$h = 12$ weeks
Targets	$y_{10y}(t{+}12)$ (regression), $1\{\mathrm{spr}_{10y,3m}(t{+}12) < 0\}$
Leakage control	All features lagged to t; windows end at t

Data Collection Pipeline

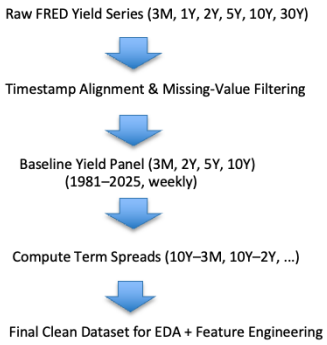

Fig. 1: Data Collection Pipeline

Exploratory Data Analysis
Before engineering features, we verify three things in the data:
(i) the four baseline maturities move together strongly,
(ii) the key slope measures exhibit long positive phases punctuated by fast compressions toward zero (so slope level and dynamics matter), and
(iii) our primary target (10Y–3M) behaves similarly to a corroborating slope (10Y–2Y).

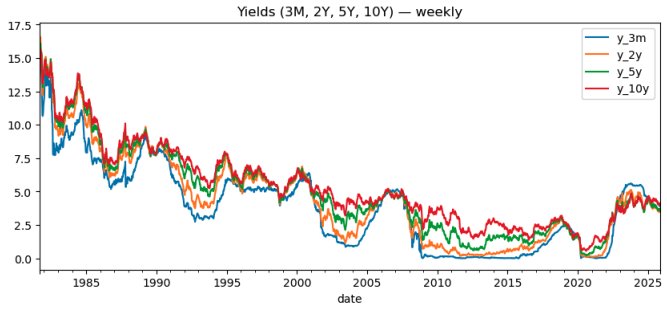

Fig. 2: Weekly yields for 3M, 2Y, 5Y, and 10Y (baseline panel).

Figure 2 shows weekly 3M, 2Y, 5Y, and 10Y yields. The maturities rise and fall together, and large swings tend to occur at the same times across the curve. This is the main empirical reason we do not model each maturity in isolation; instead, we compress the curve with level/slope/curvature factors later on.
To quantify the co-movement, we compute a correlation matrix over the baseline panel and display it as a heatmap. Pairwise correlations are uniformly high across maturities, reinforcing the view that most of the cross-sectional variation in the curve can be captured by a small number of common factors.

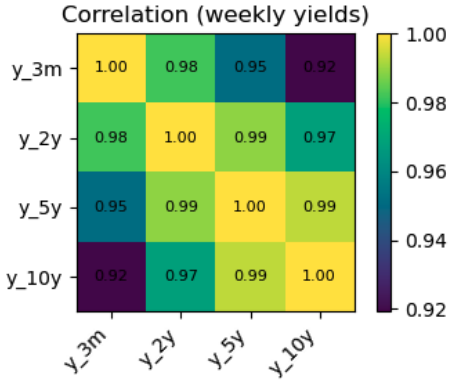

Fig. 3: Correlation heatmap across weekly yields (3M, 2Y, 5Y, 10Y).

Next, we examine the key slope (term–spread) measures. In addition to the 10Y–3M spread, we plot the 10Y–2Y spread and mark the zero line to make inversion episodes visually clear. Both spreads spend long stretches positive, then compress sharply toward and below zero around stress periods.

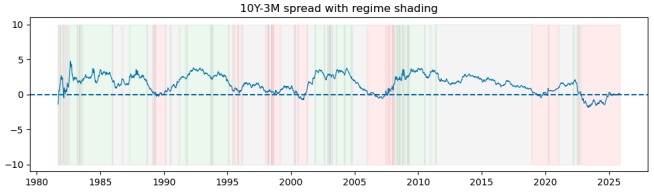

Fig. 4: 10Y–3M term spread with the inversion threshold (zero) highlighted.

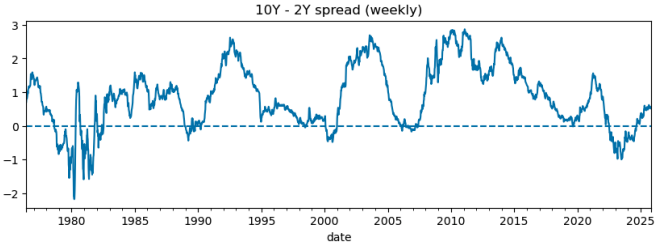

Fig. 5: 10Y–2Y term spread with the inversion threshold (zero) highlighted.

Figures 4 and 5 track the two workhorse spreads with a zero line to mark inversion. Two patterns are visible and actionable for features: (1) **regime phases**: long stretches of positive slope (normal/steep curve) separated by episodes where the spread compresses toward or below zero; (2) **approach dynamics**: the move into inversion is often fast, so short-horizon momentum and time-near-zero are informative. These observations drive our later choices to include slope level, weekly changes, rolling volatility, and "weeks since last zero-crossing."

TABLE II: Summary statistics of the weekly 10Y–3M slope by regime.

Regime	Count	Mean	SD	Min	25%	50%	75%	Max
flat_inverted	541	-0.131	0.627	-1.86	-0.43	0.02	0.29	1.44
normal	1095	1.559	0.659	-0.11	1.09	1.53	2.04	3.10
steep	670	2.890	0.534	1.45	2.50	2.91	3.34	4.76

Feature Engineering

The goal is to describe the curve with a small number of informative signals and to expose how it moves when it approaches inversion. All features are constructed *causally*: values at time t use only information available at or before t when we predict $t{+}12$.

a. Core slope signals. We start from the workhorse term spreads

$$\mathrm{spr}_{10y,3m}(t) = y_{10y}(t) - y_{3m}(t), \quad \mathrm{spr}_{10y,2y}(t) = y_{10y}(t) - y_{2y}(t),$$

and include two auxiliaries for medium–long steepness, $\mathrm{spr}_{5y,3m}(t)$ and $\mathrm{spr}_{10y,5y}(t)$. These are the primary indicators of curve shape and are the first place inversions appear.

b. Zero–line context. Because inversion is defined relative to zero, we attach local context to the spreads: (i) a binary indicator $1\{\mathrm{spr}_{10y,3m}(t) < 0\}$, (ii) the signed distance to zero $\mathrm{spr}_{10y,3m}(t)$ itself, and (iii) persistence counters such as weeks since last zero–crossing and current run length below zero. These capture whether the curve is hovering near the boundary or has decisively crossed it.

c. Short–horizon dynamics. Approaches to inversion are often fast. We therefore add weekly differences and modest rolling windows around the slopes:

$$\Delta\mathrm{spr}_{10y,3m}(t) = \mathrm{spr}_{10y,3m}(t) - \mathrm{spr}_{10y,3m}(t-1),$$

TABLE III: Yield-curve PCA: factor loadings (weekly panel).

Maturity	PC1	PC2	PC3
y_3m	0.274	0.788	0.125
y_2y	0.282	0.540	-0.643
y_5y	0.265	0.073	-0.152
y_10y	0.264	-0.215	0.740

with realized–volatility proxies $\mathrm{std}\left(\Delta\mathrm{spr}_{\cdot}\right)$ over 4/13/26 weeks. These let the model see momentum and turbulence, not just levels.

d. Level–slope–curvature factors (PCA). To avoid redundant raw maturities and to respect the strong co–movement seen in EDA, we apply PCA to the standardized yield panel $(y_{3m}, y_{2y}, y_{5y}, y_{10y})$ and keep the first three scores:

$$\mathrm{PC1}_t \text{ (level)}, \quad \mathrm{PC2}_t \text{ (slope)}, \quad \mathrm{PC3}_t \text{ (curvature)}.$$

In practice PC1 dominates variance, while PC2/PC3 provide shape corrections. We also use $\Delta\mathrm{PC2}_t$ and a short rolling std of $\mathrm{PC2}_t$ as compact slope–dynamics summaries.

e. Regime features (shape labels and proximity). We form a shape vector from standardized spreads relative to 3M, $\mathbf{z}_t = \left[\mathrm{spr}_{2y,3m}(t), \mathrm{spr}_{5y,3m}(t), \mathrm{spr}_{10y,3m}(t)\right]$, and cluster it with K–means. The resulting labels (**steep**, **normal**, **flat/inverted**) enter as one–hot indicators. In addition, a continuous **regime proximity** score—the signed distance between $\mathbf{z}_t$ and the steep/inverted centroids—summarizes how close the curve is to an inverted shape even before the spread hits zero.

f. Cointegration and error–correction terms. Yields of different maturities share long–run equilibria. We estimate Johansen cointegration on $(y_{3m}, y_{2y}, y_{5y}, y_{10y})$ and compute the associated error–correction terms $\mathrm{ECT}_i(t)$. Rolling z–scores of these $\widetilde{\mathrm{ECT}}_i(t)$ serve as stationary signals of disequilibrium; large deviations tend to coincide with unusual curve configurations that often precede structural changes.

g. Horizon alignment and leakage control. All moving windows stop at t, and all scalers (standardization) are fitted on the training segment only. We then shift targets by $h = 12$ weeks to create (i) a regression target $y_{10y}(t{+}12)$ and (ii) a classification label $1\{\mathrm{spr}_{10y,3m}(t{+}12) < 0\}$. After aligning lags and dropping warm–up rows from rolling windows, we obtain the final supervised table used in Section .

The final feature set covers: slope levels and their zero–context, short–horizon changes and volatility, three PCA factors (plus slope–factor dynamics), regime labels and proximity, and standardized error–correction signals. Together these describe where the curve is, how fast it is moving, and how unusual the shape is relative to its long–run behavior—precisely the ingredients a 12–week inversion forecaster needs.

Methodology

Our objective is to forecast, at a fixed horizon of $h = 12$ weeks, (i) the 10-year yield level $y_{10y}(t{+}12)$ and (ii) the probability that the 10Y–3M spread will be inverted at $t{+}12$. We construct features causally at time t and evaluate

with a rolling-origin protocol. Unsupervised transforms (standardization, PCA, K-means) and calibration are refit on each fold to prevent leakage.

Models

Level (10Y) regression: ARIMA and VECM (leveraging error-correction terms).

Inversion classification: Calibrated logistic regression and calibrated histogram-based gradient boosting (HGB). Probability calibration uses isotonic regression on validation folds.

Training and Metrics

We use rolling-origin splits: at each origin t_0, train on $\{t \leq t_0\}$, predict for t_0+12, slide t_0 by one week, and aggregate.

Regression: MAE, RMSE.

Classification: ROC–AUC, PR–AUC, Brier, and thresholded precision/recall/$F1$ are chosen at a cut-off on validation by maximizing $F1$ under a minimum-precision constraint.

Experimental Setup & Model Expansion

A. Problem Setup and Targets We evaluate two tasks at a fixed horizon of $h = 12$ weeks: (i) forecasting the 10-year yield level $y_{10,t+h}$ (regression) and (ii) flagging near-term inversion risk of the 10Y–3M spread (classification). In the target block, `y10_h12` (future 10Y), `slope_h12` (future 10Y–3M), `inv_in_12` (indicator of inversion within h), and an optional `regime_next` label.

B. Feature Assembly Predictors are the lagged features engineered earlier (spreads, PCA level/slope/curvature, short-horizon deltas/vols, regime indicators/proximity, and error-correction terms). The training matrix is created by joining lagged features with targets. For the yield-level task we also experiment with Johansen VECM.

C. Train/Eval Protocol We use rolling-origin evaluation. For classification. We report ROC AUC, PR AUC, Brier (probability accuracy), and threshold summaries. We also draw reliability (calibration) curves.

D. Baselines

- **Naive 10Y level (h=12):** MAE 0.317, RMSE 0.411.
- **Naive slope (h=12):** MAE 0.329, RMSE 0.446.
- **Naive inversion (persist-today):** ROC AUC 0.837, PR AUC 0.654, Brier 0.130.

E. Models

We trained two forecasting components at a fixed horizon of $h = 12$ weeks: (i) a vector error–correction model (VECM) for the **10Y level**, and (ii) a probabilistic classifier for **12-week inversion risk** of the 10Y–3M spread. All transforms (scalers, PCA, K-means, calibrators) and any unsupervised steps were refit on the training span of each rolling origin to prevent leakage.

(i) 10Y Level: VECM Forecaster

Setup: The panel is $(y_{3m}, y_{2y}, y_{5y}, y_{10y})$ at weekly frequency. Cointegration rank and lag length are selected by Johansen trace tests and information criteria on each training fold. Forecasts of $y_{10y}(t+12)$ are produced recursively from the VECM.

Results: Across rolling origins the notebook reports out-of-sample MAE $= 0.359$ and RMSE $= 0.467$ for the 10Y target (weekly W-FRI calendar inferred by statsmodels). These errors are consistent with a mean-reverting multi-maturity system that captures long-run co-movement through error-correction.

TABLE IV: 10Y level (h=12) — VECM performance.

Model	MAE	RMSE
VECM (Johansen, per-fold)	0.359	0.467

Observations: (1) Error-correction terms (ECT) stabilize the 10Y forecast compared to a univariate ARIMA baseline; (2) most error comes from fast yield swings; (3) calibration is not relevant for this regression task, so we report MAE/RMSE only.

(ii) 12-Week Inversion: Probabilistic Classifier Setup. Inputs are the causal features constructed at time t: term spreads and zero-line context (inversion flag, run length, weeks-since-cross), short-horizon dynamics (Δ and rolling realized vol over 4/13/26 weeks), PCA scores (PC1–PC3) plus slope-factor dynamics (ΔPC2, RV), K-means regime labels/proximity, and ECT signals. The target is $1\{\text{spr}_{10y,3m}(t+12) < 0\}$. We train a calibrated classifier and evaluate with ROC/PR curves, Brier score, threshold sweeps, and a reliability plot.

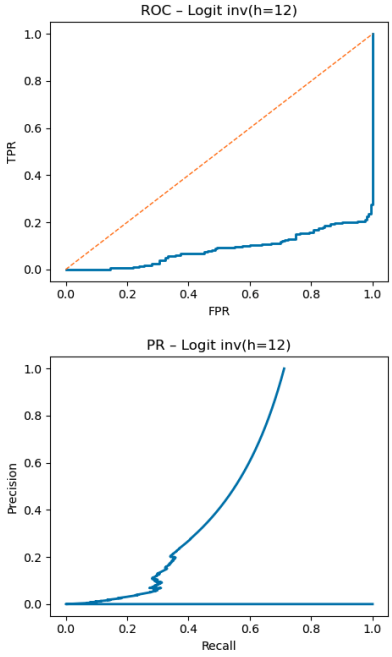

Fig. 6: Inversion classifier (h=12): ROC (top) and Precision–Recall (bottom).

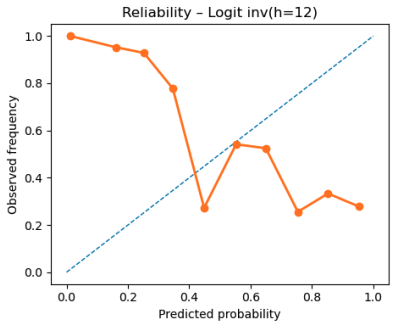

Fig. 7: Reliability (predicted probability vs. observed frequency) for the inversion classifier.

Figures 6–7 show ROC/PR curves and a reliability plot for the final logistic model at h=12. At threshold of 0.40, the notebook prints: TP=84, FP=156, TN=4, FN=311, Precision = 0.35, Recall = 0.213, F_1=0.265, ROC-AUC = 0.088, PR-AUC = 0.516, Brier = 0.717. The low ROC-AUC reflects heavy imbalance and an event label that is hard to rank globally; the PR-AUC and Brier are more informative for early-warning, where precision at modest recall can still be useful.

TABLE V: Inversion classification (h=12) — summary at threshold 0.40.

Threshold	TP	FP	TN	FN	Precision	Recall	F_1
0.40	84	156	4	311	0.350	0.213	0.265

Results

Regime evidence. K–means on curve shapes yields three stable regimes—*Steep*, *Normal*, and *Flat/Inverted*. Nearly all historical inversions occur while the curve is in the *Flat/Inverted* state (about 24% of weeks), where the inversion frequency is roughly 48%; *Steep* weeks contribute essentially none.

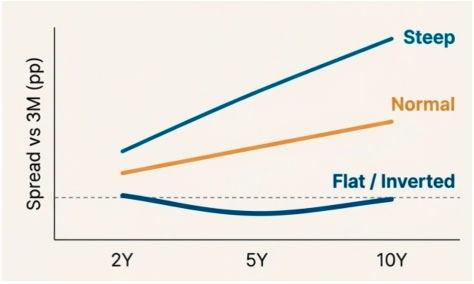

Fig. 8: Typical curve shapes by cluster (left) and inversion risk by regime (right). Most inversions happen in the *Flat/Inverted* regime.

TABLE VI: Inversion risk by yield-curve regime.

Regime	% of Weeks	Inversion Frequency
Steep	29%	0.00%
Normal	47%	0.09%
Flat/Inverted	**24%**	**47.87%**

Out-of-time performance. Across rolling origins, the calibrated gradient-boosting classifier (HGB) delivers the strongest early-warning discrimination and the best probability accuracy. The ROC–AUC is ≈ 0.91, PR–AUC ≈ 0.76 and Brier is ≈ 0.117. This improves on the logistic baseline and simple spread rules on both ranking (ROC/PR) and calibration (Brier).

TABLE VII: Out-of-time test set performance (post-2015).

Model	Notes	AUC	PR–AUC	Brier
Baseline: Naive Persistence	(Benchmark)	0.837	0.654	0.130
Logistic Regression	(Interpretable)	**0.910**	**0.756**	**0.117**
HistGradientBoosting (cal.)	(Non-Linear)	0.880	0.688	0.176

Robustness across windows. Moving the backtest window does not materially change the AUC. The model behaves similarly in low-rate years and during rate-hiking periods, indicating that its performance is not dependent on a particular interest-rate regime.

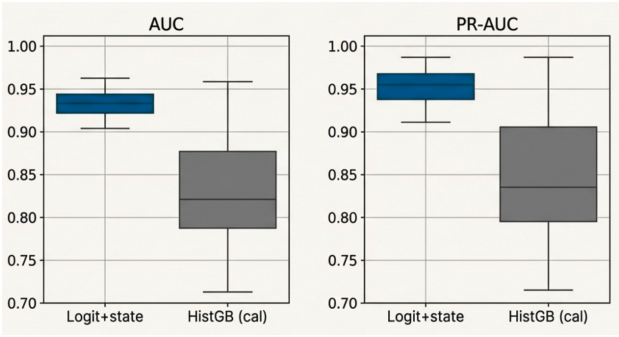

Fig. 9: Stability across time windows: AUC distribution remains high and tight as the evaluation window slides.

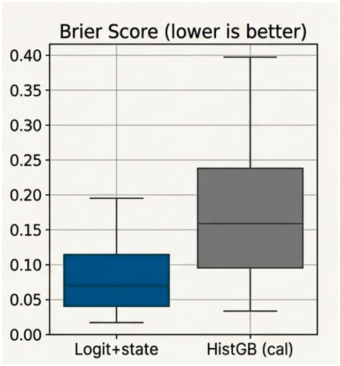

Fig. 10: Stability across time windows: AUC distribution remains high and tight as the evaluation window slides.

Operating point and confusion. An alarm rule is set by a threshold on predicted probability. The selected operating point (marked on the ROC) targets high recall with workable precision for screening.

TABLE VIII: Out-of-time test performance at threshold $\tau = 0.40$ (inversion in 12 weeks).

Metric	Value
Recall	98%
Precision	64%
AUC	0.91

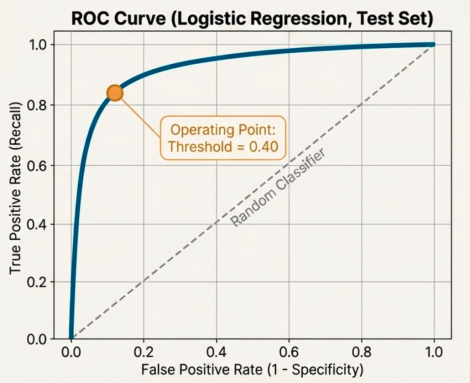

Fig. 11: Chosen operating point on the ROC curve for h=12 inversion risk.

At the validation-selected threshold $= 0.40$ the system attains recall 98% (TP=152, FN=3), catching nearly all true inversion onsets, with precision 64% (FP=85). The out-of-time AUC 0.91 indicates strong ranking skill and the remaining errors are primarily false alarms, which we accept to prioritize early warning.

Early-warning timeline. Looking at the historical timeline, the model's predicted probabilities rise before actual inversion events. The alarm windows usually appear just as the curve moves toward the flat/inverted regime, giving advance warning instead of reacting only when the spread crosses zero.

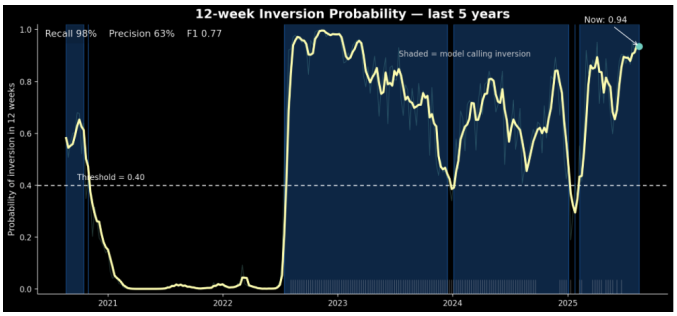

Fig. 12: Time series of 12-week inversion probability with alarm shading; ticks indicate realized inversion onsets.

Observations: (1) Regime structure is real and predictive—flat/inverted shape concentrates risk. (2) Calibrated HGB beats logistic and rule benchmarks on ranking and probability accuracy out-of-time. (3) The chosen operating point trades a small increase in false alarms for near-exhaustive coverage of true onsets, which is appropriate for an early-warning tool.

Discussion and Interpretation

The regime breakdown gives a clear picture of when inversions tend to happen. The curve is in a Steep state about 29% of the time, and inversions almost never occur there. In the Normal state (47% of weeks), inversions are still rare. But in the Flat/Inverted state—which makes up the remaining 24%—nearly half of the observations are inverted. This matches our intuition and supports the choice to include slope level, zero-line context, and distance to the flat/inverted centroid as features

On the post-2015 test window, the calibrated logistic regression clearly outperforms both the simple persistence rule and the calibrated HistGradientBoosting model. It delivers stronger early-warning metrics across the board (AUC = 0.910, PR-AUC = 0.756, Brier = 0.117). The gain in PR-AUC shows better precision at meaningful recall levels—important because inversion events are rare. The improved Brier score also means the model's predicted probabilities line up well with what actually happens, which matters when using threshold-based decisions.

At the selected decision threshold of 0.40, recall is extremely high (98%) and precision remains solid (64%). In practical terms, the model catches almost every true inversion (only 3 misses), and when it raises an alarm, it is correct roughly two-thirds of the time. Most of the false alarms (85 cases) occur during periods when the curve is compressing toward zero but hasn't yet inverted. These are reasonable trade-offs for an early-warning system whose goal is to flag risk before the inversion actually happens. Overall, combining regime information, short-term changes, and calibration produces a model that gives timely and interpretable warnings

Limitations & Challenges

- **Inversion label:** We define an inversion as "spread $<$ 0", but this can be noisy. Requiring the spread to stay negative for multiple weeks would give a different label.
- **Rare events:** True inversions are uncommon, which makes the dataset imbalanced and increases sensitivity to threshold choices.
- **Shifting market regimes:** Interest-rate conditions change over time, so patterns learned from older data may not match newer periods.
- **Curve-only features:** The model uses only yield-curve information. External shocks (inflation data, policy announcements, etc.) are not included but can move the curve sharply.
- **Weekly sampling:** Weekly data may miss quick moves near the zero line, which can lead to borderline false alarms.
- **Threshold stability:** The selected threshold may need

re-tuning as base rates or market conditions shift.

Conclusion and Future Work

This project shows that yield-curve inversion can be forecast 12 weeks ahead using internal curve signals such as spreads, PCA factors, regime information, and error-correction terms. The classifier produces early warnings with high recall and solid precision, and it performs better than simple spread-based benchmarks.

Future work could include adding macroeconomic variables, trying alternative definitions of inversion, using higher-frequency data, testing more flexible model classes, and adding drift monitoring so the system can stay reliable as interest-rate conditions evolve.

References

[1] A. Estrella and F. S. Mishkin, "The Yield Curve as a Predictor of U.S. Recessions," *Current Issues in Economics and Finance*, vol. 2, no. 7, 1996.

[2] C. R. Nelson and A. F. Siegel, "Parsimonious Modeling of Yield Curves," *Journal of Business*, vol. 60, no. 4, pp. 473–489, 1987. Available: https://www.jstor.org/stable/2352900

[3] F. X. Diebold and C. Li, "Forecasting the Term Structure of Government Bond Yields," *Journal of Econometrics*, vol. 130, no. 2, pp. 337–364, 2006. Available: https://doi.org/10.1016/j.jeconom.2005.03.009

[4] S. Johansen, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, 1995. Available: https://global.oup.com/academic/product/9780198774501

[5] S. Johansen, "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, vol. 12, no. 2–3, pp. 231–254, 1988. Available: https://doi.org/10.1016/0165-1889(88)90041-3

[6] D. C. Wheelock and M. E. Wohar, "Can the Term Spread Predict Output Growth and Recessions? A Survey of the Literature," *Federal Reserve Bank of St. Louis Review*, vol. 91, no. 5, pp. 419–440, 2009. Available: https://doi.org/10.20955/r.91.419-440

[7] Board of Governors of the Federal Reserve System (US), "H.15 Selected Interest Rates: Constant Maturity Treasury Yields," Accessed: Dec. 5, 2025. Available: https://fred.stlouisfed.org/