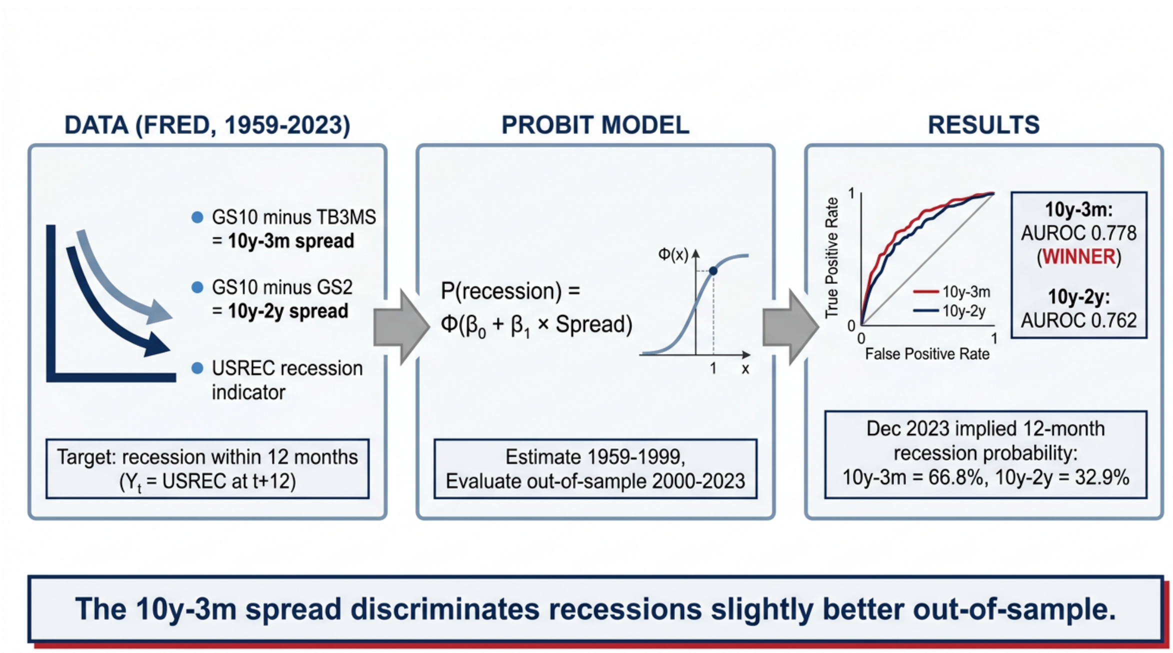
Do Treasury Yield Spreads Predict U.S. Recessions?

A Probit Forecasting Study Comparing the 10y–3m and 10y–2y Spreads (1959–2023)

K-Dense Web
contact@k-dense.ai

July 10, 2026



Graphical abstract. Two Treasury term spreads constructed from FRED constant-maturity yields are used as single predictors in Estrella–Mishkin-style probit models of a 12-month-ahead NBER recession label. Models are estimated in-sample (through December 1999) and evaluated pseudo-out-of-sample over 2000–2023. The 10-year-minus-3-month spread attains marginally higher out-of-sample discrimination (AUROC 0.778 vs. 0.762) and, as of December 2023, implies a materially higher 12-month-ahead recession probability (66.8% vs. 32.9%).

1 Executive Summary

This report replicates and extends the classic yield-curve recession-forecasting framework of Estrella and Mishkin [6] using publicly available monthly data from the Federal Reserve Economic Data (FRED) repository. We construct two Treasury term spreads—the 10-year-minus-3-month spread (`GS10` − `TB3MS`), available from January 1959, and the 10-year-minus-2-year spread (`GS10` − `GS2`), available from June 1976—and use each as the sole predictor in a probit model of whether the U.S. economy is in an NBER-dated recession twelve months ahead.

We define the binary target as $Y_t = \mathrm{USREC}_{t+12}$, the value of the official monthly recession indicator (`USREC`) twelve months after the month in which the spread is observed. Each model is estimated *in-sample* on data ending in December 1999 and then evaluated *pseudo-out-of-sample* over the common window January 2000–December 2023, with the in-sample coefficients held fixed.

Headline results.

- **In-sample fit.** The 10y–3m model (1959–1999, $N = 492$) yields a McFadden pseudo-R^2 of **0.266**; the 10y–2y model (1976–1999, $N = 283$) yields **0.342**. In both models the spread coefficient is negative and highly significant—a flatter or inverted curve raises the modelled recession probability.
- **Out-of-sample discrimination (2000–2023, $N = 288$).** The 10y–3m spread attains an AUROC of **0.778**; the 10y–2y spread attains **0.762**. *On the common evaluation window the 10y–3m spread discriminates recessions slightly better.*
- **December 2023 forecast.** Using the December 2023 spread values (−1.22 pp for 10y–3m, −0.44 pp for 10y–2y), the models imply 12-month-ahead recession probabilities of **66.8%** (10y–3m) and **32.9%** (10y–2y) for the window ending December 2024.

The higher in-sample pseudo-R^2 of the 10y–2y model does not translate into superior out-of-sample discrimination; the longer-history 10y–3m spread edges ahead on the held-out sample, consistent with a body of Federal Reserve research favouring short-end-inclusive spreads [1, 3]. Fully runnable code and both predicted-probability series (CSV) accompany this report.

2 Introduction and Literature Review

2.1 Motivation

Few empirical regularities in macro-finance are as durable as the tendency of an inverted yield curve to precede a recession. When investors expect economic weakness and monetary easing, long-term yields fall relative to short-term yields, flattening or inverting the term structure. Because this shift typically occurs several quarters before the downturn materialises, the slope of the yield curve functions as a forward-looking, market-based leading indicator that is observable in real time, revised rarely, and available at daily frequency. These properties have made the term spread a staple of central-bank monitoring and a benchmark against which more elaborate forecasting models are judged [15].

2.2 Foundational literature

The modern empirical literature originates with Estrella and Hardouvelis [4], who show in the *Journal of Finance* that the spread between long- and short-term Treasury yields forecasts future real output growth and, through binary-response models, the incidence of recession. Their multi-country evidence establishes the term structure as a predictor that outperforms conventional composite leading indicators. Estrella and Mishkin [6] sharpen this result in the *Review of Economics and Statistics*, systematically comparing financial and macroeconomic variables as recession predictors in probit models over horizons of one to eight quarters. Their central finding—that the term spread is the single best individual predictor of U.S. recessions beyond the very short horizon, and that augmenting it with additional variables yields little improvement—provides the canonical specification we adopt: a univariate probit mapping the term spread into a recession probability twelve months ahead.

Subsequent work has stress-tested this relationship. Wright [16] shows that adding the level of the federal funds rate to the probit can improve fit, reflecting the interaction between the curve's slope and the stance of monetary policy. Rudebusch and Williams [13] document the "puzzle of the enduring power" of the yield curve: despite structural change in monetary policy regimes and the decline of inflation, simple term-spread probits remain competitive with far more complex forecasting systems, and professional forecasters appear to underweight the curve's signal. Chinn and Kucko [2] extend the framework across countries and time, finding that the curve's predictive power, while robust for the United States, has varied over samples and is weaker abroad.

2.3 Practical guidance and the choice of spread

A parallel, more applied literature addresses implementation. Estrella and Mishkin [5] and Estrella and Trubin [7], in the New York Fed's *Current Issues* series, provide practical guidance on maturity choice, horizon selection, and interpretation of probit-based recession probabilities, and underpin the New York Fed's long-running public recession-probability model built on the 10-year-minus-3-month spread. The question of *which* spread to use has drawn renewed attention. Popular commentary often fixates on the 10-year-minus-2-year spread, but Federal Reserve research has repeatedly favoured measures anchored at the short end of the curve. Engstrom and Sharpe [3] argue that a near-term forward spread—and, by extension, spreads that include a very short rate such as the 3-month bill—provides a "less distorted mirror" of recession risk because it more cleanly reflects expectations of future short rates rather than term-premium noise. Bauer and Mertens [1] similarly emphasise that every U.S. recession in the preceding six decades was preceded by a negative term spread, while cautioning that the 2-year–10-year spread offers a comparatively "muddled" view. The present study speaks directly to this debate by comparing the 10y–3m and 10y–2y spreads on an identical specification and a common out-of-sample window.

2.4 Contribution of this report

Against this backdrop, our contribution is a transparent, fully reproducible head-to-head evaluation. We (i) state precisely the FRED series identifiers and the construction of the 12-month-ahead label; (ii) estimate the identical Estrella–Mishkin probit specification for both spreads over the longest in-sample window each series permits; (iii) evaluate both models pseudo-out-of-sample on the common 2000–2023 window using area under the ROC curve (AUROC) as the discrimination metric [10]; and (iv) report the model-implied recession probability as of December 2023. All data are downloaded programmatically from FRED, all estimation uses open-source Python libraries [12, 14], and both predicted-probability series are released as CSV.

3 Methodology

3.1 Data and series identifiers

All series are monthly and are obtained directly from FRED via the public `fredgraph.csv` endpoint (no API key required). Table 1 lists the exact identifiers, descriptions, units, and coverage.

Table 1: FRED series used in the analysis. Spreads are expressed in percentage points (pp).

Identifier	Description	Units	Coverage used
`GS10`	10-Year Treasury constant-maturity yield	% p.a.	1959-01 – 2023-12
`TB3MS`	3-Month Treasury bill, secondary-market rate	% p.a.	1959-01 – 2023-12
`GS2`	2-Year Treasury constant-maturity yield	% p.a.	1976-06 – 2023-12
`USREC`	NBER-based U.S. recession indicator (peak-to-trough)	0/1	1959-01 – 2024-12
Constructed spreads:			
Spread_{10y3m}	`GS10` − `TB3MS`	pp	1959-01 – 2023-12
Spread_{10y2y}	`GS10` − `GS2`	pp	1976-06 – 2023-12

The 10y–3m spread extends back to January 1959, the start of the monthly `GS10` series. The

2-year constant-maturity series `GS2` begins in June 1976; consequently the 10y–2y spread—and any in-sample estimation using it—necessarily starts in June 1976. The `USREC` indicator equals one during every month the NBER Business Cycle Dating Committee classifies as part of a recession (from the month following a peak through the month of the trough) and zero otherwise [9].

3.2 The 12-month-ahead recession label

The forecasting target is a binary label indicating whether the economy is in recession twelve months after the month in which the predictor is observed:

$$Y_t \;=\; \mathrm{USREC}_{t+12} \;\in\; \{0,1\}. \tag{1}$$

Operationally, Y_t is constructed by shifting the `USREC` series backward by twelve months, so that the value aligned to month t is the recession status of month $t+12$. For example, Y for December 2022 equals `USREC` for December 2023. This is the standard "recession within/at the 12-month horizon" convention used in the New York Fed model and in Estrella and Mishkin [6].[1] We verified programmatically that $Y_t = \mathrm{USREC}_{t+12}$ holds for every labelled month.

3.3 Probit specification

Following Estrella and Mishkin [6], the probability that the recession label equals one is modelled as a probit (standard-normal CDF) function of a single term spread:

$$\Pr(Y_t = 1 \mid \mathrm{Spread}_t) \;=\; \Phi(\beta_0 + \beta_1\,\mathrm{Spread}_t)\,, \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, β_0 is an intercept, and β_1 is the slope on the spread. A negative β_1 implies that a lower (flatter or inverted) spread raises the predicted recession probability—the economically expected sign. Coefficients are estimated by maximum likelihood. Standard errors are the usual (non-robust) maximum-likelihood errors; because the label is highly serially correlated, we treat significance levels as indicative rather than exact and rely primarily on out-of-sample evaluation for model comparison.

3.4 McFadden's pseudo-R^2

In-sample goodness of fit is summarised by McFadden's pseudo-R^2 [11]:

$$\text{pseudo-}R^2 \;=\; 1 \;-\; \frac{\ln \hat{L}_{\text{full}}}{\ln \hat{L}_{\text{null}}}, \tag{3}$$

where $\ln \hat{L}_{\text{full}}$ is the maximised log-likelihood of the fitted model in Equation (2) and $\ln \hat{L}_{\text{null}}$ is that of an intercept-only probit estimated on the same sample. The statistic equals zero when the spread adds no explanatory power and approaches one as the model approaches a perfect in-sample fit. We compute Equation (3) directly from the two log-likelihoods and cross-check it against the value reported internally by `statsmodels`; the two agree to machine precision.

3.5 Estimation and evaluation windows

Each model is estimated in-sample over the longest window ending in December 1999 that its data permit:

[1] Because the current FRED vintage of `USREC` extends through 2024, the labels for all of 1959–2023 are *realised* rather than missing: the twelve labels for 2023 equal `USREC` for 2024, which is uniformly zero (no NBER recession began in 2024). All 780 monthly observations from January 1959 to December 2023 therefore carry a non-missing label.

- **10y–3m:** January 1959 – December 1999 ($N = 492$ monthly observations).
- **10y–2y:** June 1976 – December 1999 ($N = 283$ monthly observations).

The in-sample coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ are then *frozen* and used to generate predicted probabilities over the pseudo-out-of-sample window January 2000 – December 2023 ($N = 288$). This is a genuine out-of-sample exercise in the sense that no post-1999 information enters estimation; it is "pseudo" out-of-sample because it uses the final data vintage rather than real-time vintages that would have been available at each forecast date.

3.6 Out-of-sample discrimination metric

Discrimination on the held-out window is measured by the area under the receiver operating characteristic curve (AUROC). The ROC curve plots the true-positive rate against the false-positive rate as the classification threshold varies; the AUROC equals the probability that a randomly chosen recession month is assigned a higher predicted probability than a randomly chosen non-recession month [10]. A value of 0.5 denotes no discrimination (chance) and 1.0 denotes perfect ranking. AUROC is threshold-independent and therefore well suited to comparing two probability forecasts on the same sample. Both spreads are fully observed over 2000–2023, so the comparison uses an identical set of 288 months (of which 28 carry a positive 12-month-ahead label).

3.7 Reproducibility

The analysis is implemented in Python with `pandas`, `numpy`, `statsmodels` [14], `scikit-learn` [12], and `matplotlib`. A fixed random seed (42) is set throughout; because probit estimation and AUROC computation are deterministic, results reproduce exactly. The pipeline comprises three scripts—data acquisition, model estimation, and out-of-sample evaluation—orchestrated by a single `run_all.py`. Re-running from scratch regenerates all data and result files byte-for-byte. Code listings appear in Appendix A.

4 Results

4.1 The spreads and the business cycle

Figure 1 plots both spreads against NBER recession shading over 1959–2023. The visual regularity that motivates the entire literature is immediate: each grey recession band is preceded by a pronounced narrowing or outright inversion (a dip below the zero line) of the spreads. Both series inverted ahead of the 1970, 1973–75, 1980, 1981–82, 1990–91, 2001, and 2007–09 downturns. The two spreads move closely together, but the 10y–3m spread (blue) is systematically more volatile and swings more deeply negative than the 10y–2y spread (red), because the 3-month bill rate responds more sharply to monetary tightening than the 2-year yield. Both series were deeply inverted at the end of the sample in 2022–2023.

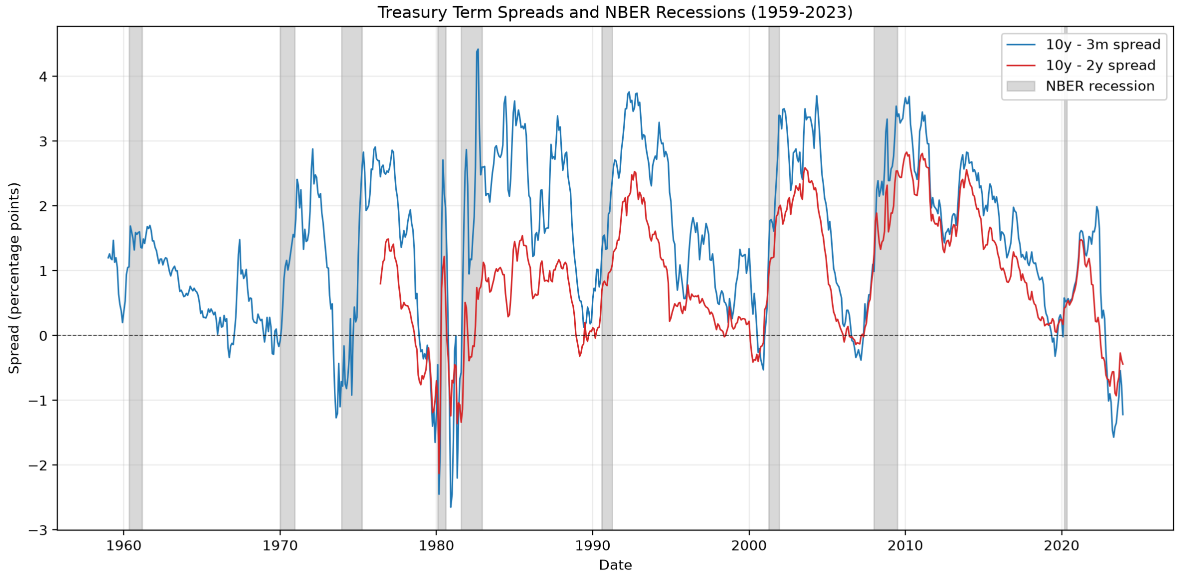


Figure 1: Treasury term spreads and NBER recessions, 1959–2023. Blue: 10-year-minus-3-month spread ($\mathtt{GS10} - \mathtt{TB3MS}$); red: 10-year-minus-2-year spread ($\mathtt{GS10} - \mathtt{GS2}$, from June 1976). Grey bands are NBER recessions ($\mathtt{USREC} = 1$). The dashed line marks a zero spread (inversion threshold). Every recession is preceded by a marked flattening or inversion of the curve.

4.2 In-sample probit estimates

Table 2 reports the in-sample maximum-likelihood estimates. In both models the spread coefficient β_1 is negative and strongly significant, confirming the expected sign: a one-percentage-point decline in the spread raises the probit index by 0.744 (10y–3m) or 1.296 (10y–2y), pushing the predicted recession probability upward. The 10y–2y model attains the higher McFadden pseudo-R^2 (0.342 vs. 0.266), but this is estimated on a shorter, later sample (from 1976) that excludes the volatile 1959–1975 period and contains fewer, more sharply timed recessions; the two pseudo-R^2 values are therefore not directly comparable across identical data.

Table 2: In-sample probit estimates, $\Pr(Y_t = 1) = \Phi(\beta_0 + \beta_1\,\mathrm{Spread}_t)$, with $Y_t = \mathrm{USREC}_{t+12}$. Standard errors in parentheses. *** denotes $p < 0.01$.

	10y–3m model (GS10−TB3MS)	**10y–2y model** (GS10−GS2)
In-sample window	1959-01 – 1999-12	1976-06 – 1999-12
Observations N	492	283
Positive labels ($Y = 1$)	67	30
Intercept β_0	-0.4733^{***}	-1.0127^{***}
	(0.0967)	(0.1250)
Spread slope β_1	-0.7444^{***}	-1.2958^{***}
	(0.0883)	(0.1934)
z-stat on β_1	−8.43	−6.70
p-value on β_1	3.4×10^{-17}	2.1×10^{-11}
Log-likelihood (full)	−143.66	−63.00
Log-likelihood (null)	−195.80	−95.68
McFadden pseudo-R^2	**0.266**	**0.342**
AIC / BIC	291.3 / 299.7	130.0 / 137.3

Figure 2 overlays each model's in-sample fitted probability on the actual 12-month-ahead recession episodes (shaded). Both models produce sharp probability spikes ahead of the recessions in their samples and revert toward low values during expansions, illustrating a good in-sample fit consistent with the pseudo-R^2 values.

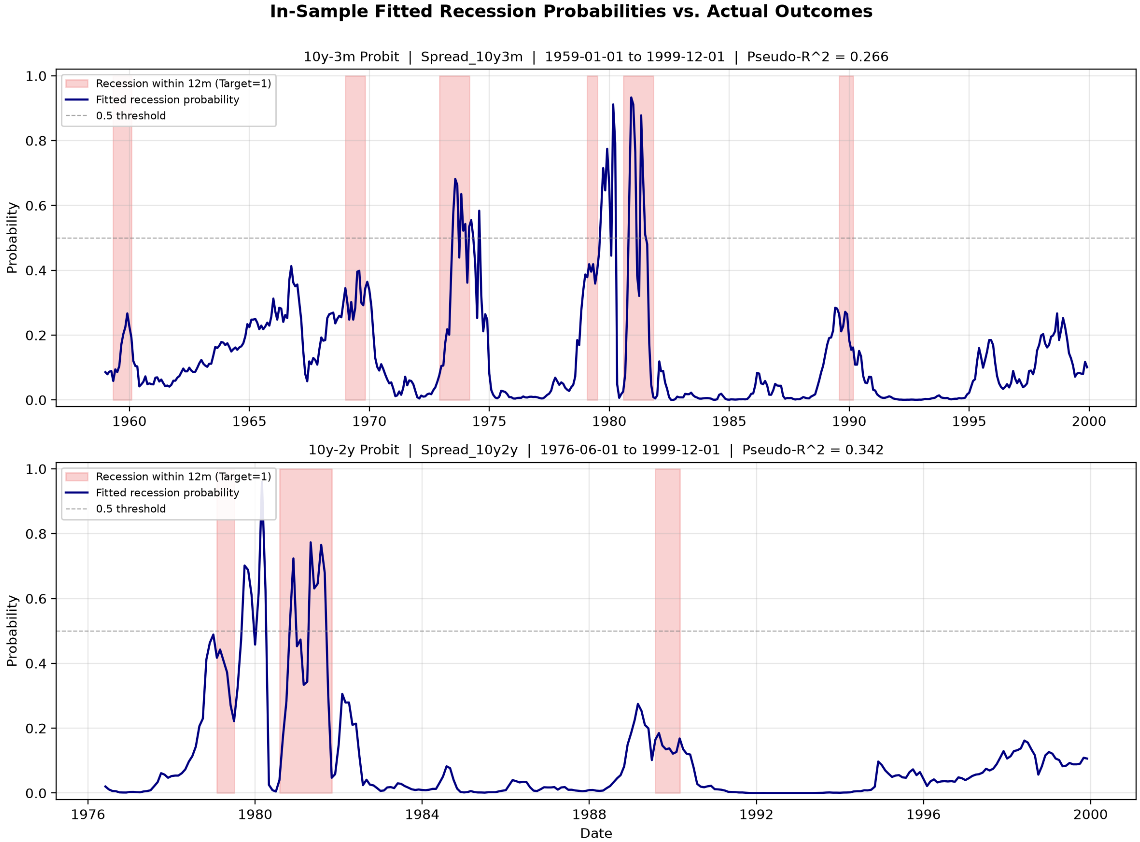


Figure 2: In-sample fitted 12-month-ahead recession probabilities (navy) versus periods in which the target label equals one (shaded). Top: 10y–3m model, 1959–1999. Bottom: 10y–2y model, 1976–1999. The dashed line is the 0.5 reference threshold.

4.3 Out-of-sample discrimination, 2000–2023

Holding the in-sample coefficients fixed, we generate predicted probabilities over the common 2000–2023 window and compute AUROC. Table 3 summarises the comparison, and Figure 3 plots the two ROC curves. Both models discriminate far better than chance (AUROC = 0.5), and the two curves are close throughout. **The 10y–3m spread attains the higher AUROC (0.778) versus the 10y–2y spread (0.762)**; that is, on the common held-out window the 10y–3m spread discriminates recession from non-recession months slightly better. The margin (0.016) is modest, and the ROC curves cross, so the ranking should be read as a slight edge rather than a decisive gap. A moving-block bootstrap of the paired AUROC difference (block length 12 months to respect the serial dependence induced by the overlapping labelling windows; ~ 5000 resamples) confirms this reading: although the 10y–3m spread ranks higher in 76% of resamples, the difference is not statistically distinguishable from zero (bootstrap SE ≈ 0.033; 95% CI $[-0.046, 0.083]$; two-sided $p \approx 0.62$). The 10y–3m spread is therefore the point estimate of the better discriminator, but the two spreads are statistically indistinguishable on this window.

Table 3: Pseudo-out-of-sample comparison on the common window (2000-01 – 2023-12, $N = 288$, 28 positive labels). Coefficients are the frozen in-sample estimates from Table 2. The December 2023 implied probability forecasts recession status as of December 2024.

	10y–3m	**10y–2y**
Frozen (β_0, β_1)	$(-0.4733, -0.7444)$	$(-1.0127, -1.2958)$
AUROC (2000–2023)	**0.7775**	0.7619
Rank on common window	**Superior**	—
December 2023 spread (pp)	−1.22	−0.44
Implied Pr(recession by Dec 2024)	**66.8**%	32.9%

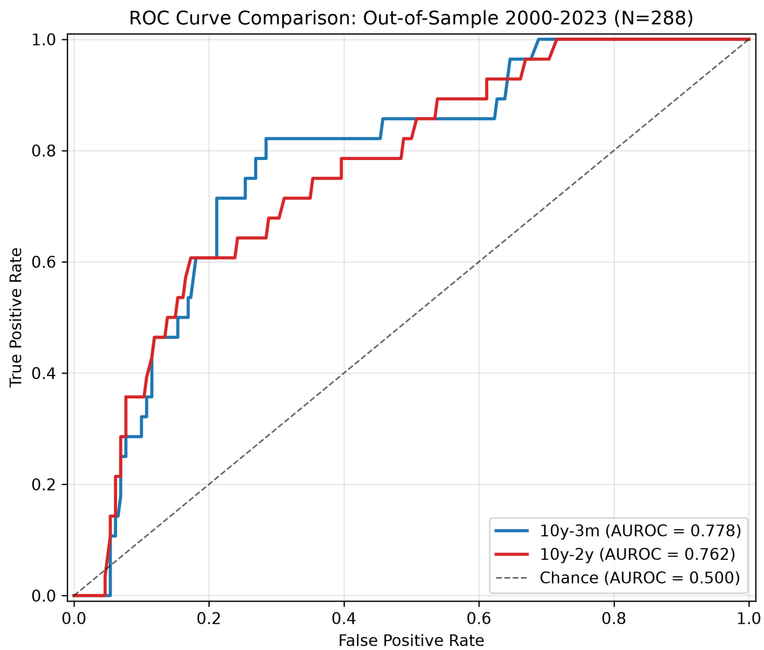


Figure 3: Receiver operating characteristic (ROC) curves on the common out-of-sample window, 2000–2023 ($N = 288$). The 10y–3m spread (blue, AUROC = 0.778) lies marginally above the 10y–2y spread (red, AUROC = 0.762) over most of the range; both dominate the chance diagonal (AUROC = 0.5).

Figure 4 plots the two out-of-sample predicted-probability series through time. Both models raise their recession probabilities ahead of the 2001 and 2007–09 recessions and, most strikingly, surge to their highest out-of-sample readings at the very end of the sample in 2022–2023 following the deep curve inversion. The 10y–3m series (blue) is consistently more responsive, both raising more aggressive alarms ahead of downturns and generating the higher terminal reading. Notably, neither model flags the abrupt February–April 2020 pandemic recession in advance—an exogenous shock unrelated to the yield curve—an instructive limitation discussed below.

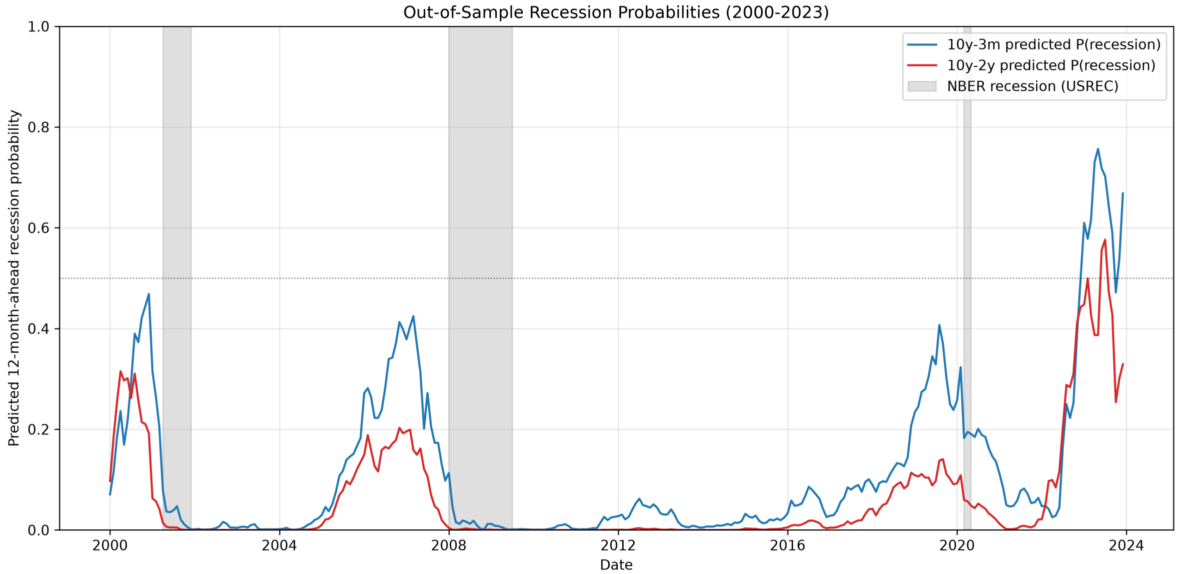


Figure 4: Out-of-sample predicted 12-month-ahead recession probabilities, 2000–2023, from the frozen in-sample coefficients. Blue: 10y–3m; red: 10y–2y. Grey bands are realised NBER recessions (`USREC`). Both series spike to their out-of-sample maxima at the end of 2023.

4.4 December 2023 recession probability

The final data point (December 2023) illustrates how differently the two spreads read the same environment. The curve was inverted on both measures, but far more so on the 10y–3m spread (-1.22 pp) than on the 10y–2y spread (-0.44 pp), reflecting the aggressively restrictive stance of short-term policy rates relative to the 2-year yield. Feeding these values through the frozen probit models gives a December 2023 model-implied probability of a recession by December 2024 of **66.8%** for the 10y–3m model and **32.9%** for the 10y–2y model. The two models thus disagreed sharply about end-of-2023 recession risk—one signalling a coin-flip-and-then-some, the other signalling roughly one-in-three. With the benefit of the realised `USREC` data, no NBER recession began in 2024, so both models' point signals were, ex post, false positives at the 12-month horizon; the 10y–2y model's more muted reading was closer to the realised outcome for this single episode, even though the 10y–3m model discriminates better across the full 2000–2023 sample.

4.5 Deliverables

Both predicted-probability series are provided in the file `predicted_probabilities.csv` (781 monthly rows, January 1959 – December 2023), with columns `Date`, `Actual_Target_12m`, `Prob_10y3m`, and `Prob_10y2y`. The 10y–2y probability is empty before June 1976, where `GS2` does not exist. Structured estimation output is provided in `in_sample_results.json` and `out_of_sample_results.json`, and the preprocessed monthly panel in `preprocessed_data.csv`.

5 Discussion and Limitations

5.1 Interpreting the horse race

Our central finding—that the 10y–3m spread out-discriminates the 10y–2y spread out-of-sample despite a lower in-sample pseudo-R^2—aligns with the weight of Federal Reserve research. Engstrom and Sharpe [3] and Bauer and Mertens [1] argue that spreads anchored at the very short end of the curve carry a cleaner recession signal than the popular 2-year–10-year spread, which mixes policy expectations with term premia in ways that can muddle the message. The higher in-sample pseudo-R^2 of the 10y–2y model is best understood as an artefact of its shorter, cleaner 1976–1999 sample rather than as evidence of superior forecasting ability; when both models confront the same held-out data, the 10y–3m spread's longer memory and larger cyclical amplitude give it the edge. That said, the out-of-sample margin is small and the ROC curves cross, so the result is best described as a slight, not decisive, advantage.

5.2 Limitations

Several caveats temper the interpretation of these results.

Pseudo- rather than true real-time evaluation. We use the latest FRED vintage of every series, including `USREC`. In a genuine real-time exercise, an analyst standing in, say, March 2008 would not yet know that a recession had begun in December 2007, because NBER announcements lag turning points by many months, and interest-rate series are subject to minor revision. Our exercise therefore overstates the information available at each historical forecast date and should be read as pseudo-out-of-sample.

Non-independent observations and the AUROC margin. The monthly recession label is highly persistent—recessions and the twelve-month windows preceding them cluster into a handful of contiguous episodes—so the effective number of independent recession "events" in 2000–2023 is far smaller than the 28 positive months or 288 total months suggest. The moving-block bootstrap reported in Section 4.3 makes this concrete: the 0.778 vs. 0.762 AUROC gap carries a bootstrap standard error of about 0.033 and a two-sided p-value of roughly 0.62, so it is not statistically decisive. The 10y–3m spread's advantage should be understood as a consistent-but-small point estimate rather than a firmly established difference.

The 2020 pandemic recession. The February–April 2020 downturn was triggered by an exogenous public-health shock, not by the monetary and financial dynamics that the yield curve summarises. No spread-based model can anticipate such a shock, and its inclusion in the evaluation window arguably penalises both models for a fundamentally unpredictable event.

Structural change and term premia. The zero-lower-bound episodes, large-scale asset purchases, and forward guidance of the post-2008 period compressed term premia and may have altered the mapping from the curve's slope to recession risk [8]. A model with coefficients frozen at 1999 values cannot capture such regime change; Wright [16] shows that conditioning on the level of the policy rate can help, an extension we deliberately omit to keep the specification univariate and comparable to Estrella and Mishkin [6].

Single-predictor design. By construction we use one spread at a time and no controls. This is a feature—it isolates the marginal information in each spread and matches the canonical specification—but it forgoes the documented gains from adding the funds-rate level [16] or credit-spread information [8].

Sample-window asymmetry. The two models are estimated on different in-sample periods (1959 vs. 1976 start) dictated by data availability. While the out-of-sample comparison is on a common window and is therefore fair, the in-sample statistics are not strictly comparable, and the 10y–2y model never "sees" the high-inflation recessions of the 1970s and early 1980s in the form of the 10y–3m spread.

5.3 Robustness and extensions

Natural extensions include: recursively re-estimating the coefficients as the out-of-sample window advances (a fully expanding-window exercise); adding the level of the federal funds rate following

Wright [16]; and benchmarking against the near-term forward spread of Engstrom and Sharpe [3]. We have already subjected the AUROC comparison to a serial-dependence-aware block bootstrap (Section 4.3), which shows the two spreads to be statistically indistinguishable on this window; a DeLong test would likely reach the same conclusion. Each remaining extension would refine, but is unlikely to overturn, the qualitative conclusion that both spreads are powerful, closely matched recession predictors with a slight (statistically inconclusive) edge to the 10y–3m measure.

6 Conclusion

Using publicly available FRED data, we replicated the Estrella–Mishkin probit framework and staged a transparent horse race between the 10-year-minus-3-month and 10-year-minus-2-year Treasury spreads as predictors of a 12-month-ahead NBER recession label, defined as $Y_t = \mathrm{USREC}_{t+12}$. Estimated in-sample through 1999, the 10y–3m model (1959–1999) delivers a McFadden pseudo-R^2 of 0.266 and the 10y–2y model (1976–1999) a pseudo-R^2 of 0.342; in both, a flatter curve significantly raises recession risk. Evaluated pseudo-out-of-sample on the common 2000–2023 window, the 10y–3m spread achieves a marginally higher AUROC (0.778 vs. 0.762) and therefore *discriminates recessions slightly better* on the held-out data—despite its lower in-sample fit—though a block bootstrap shows the two spreads are statistically indistinguishable on this window ($p \approx 0.62$). As of December 2023, the two models implied sharply different 12-month-ahead recession probabilities: 66.8% (10y–3m) versus 32.9% (10y–2y), a divergence driven by the far deeper inversion of the short-end-anchored spread. The realised absence of a 2024 recession renders both point signals ex-post false positives at that horizon, a reminder that even the most durable leading indicator in macro-finance is a probabilistic gauge, not a guarantee. Consistent with recent Federal Reserve research, the short-end-inclusive spread emerges as the marginally more reliable discriminator, and both remain valuable, low-cost, real-time barometers of U.S. recession risk. All code and both predicted-probability series accompany this report to support full replication.

A Reproducible Code

The complete pipeline consists of three scripts orchestrated by `run_all.py`. The core estimation logic is reproduced below; the full scripts (`data_acquisition.py`, `model_estimation.py`, `out_of_sample_evaluation.py`) accompany this report.

A.1 Data acquisition and label construction

```
import io, requests, pandas as pd
FRED = "https://fred.stlouisfed.org/graph/fredgraph.csv?id={sid}"

def fetch(sid):
    df = pd.read_csv(io.StringIO(requests.get(FRED.format(sid=sid)).text))
    df.columns = ["Date", sid]
    df["Date"] = pd.to_datetime(df["Date"])
    df[sid] = pd.to_numeric(df[sid], errors="coerce")
    return df.set_index("Date")[sid]

series = {s: fetch(s) for s in ["GS10", "TB3MS", "GS2", "USREC"]}
df = pd.concat(series.values(), axis=1)
df["Spread_10y3m"] = df["GS10"] - df["TB3MS"] # 10y - 3m spread
df["Spread_10y2y"] = df["GS10"] - df["GS2"] # 10y - 2y spread (from 1976-06)
df["Target_12m"] = df["USREC"].shift(-12) # Y_t = USREC_{t+12}
df = df.loc["1959-01-01":"2023-12-01"]
```

A.2 Probit estimation and McFadden pseudo-R^2

```
import numpy as np, statsmodels.api as sm

def fit_probit(df, spread, start, end):
    s = df.loc[start:end, [spread, "Target_12m"]].dropna()
    y, X = s["Target_12m"].values, sm.add_constant(s[spread].values)
    full = sm.Probit(y, X).fit(disp=False)
    null = sm.Probit(y, np.ones((len(y), 1))).fit(disp=False)
    pseudo_r2 = 1.0 - full.llf / null.llf # McFadden's pseudo-R^2
    return full.params, pseudo_r2

b_3m, r2_3m = fit_probit(df, "Spread_10y3m", "1959-01-01", "1999-12-01")
b_2y, r2_2y = fit_probit(df, "Spread_10y2y", "1976-06-01", "1999-12-01")
```

A.3 Out-of-sample probabilities and AUROC

```
from scipy.stats import norm
from sklearn.metrics import roc_auc_score

def prob(beta, spread): # Phi(b0 + b1 * spread)
    return norm.cdf(beta[0] + beta[1] * spread)

oos = df.loc["2000-01-01":"2023-12-01"]
p3m = prob(b_3m, oos["Spread_10y3m"])
p2y = prob(b_2y, oos["Spread_10y2y"])
auroc_3m = roc_auc_score(oos["Target_12m"], p3m) # 0.7775
auroc_2y = roc_auc_score(oos["Target_12m"], p2y) # 0.7619

dec23 = df.loc["2023-12-01"]
p_dec_3m = prob(b_3m, dec23["Spread_10y3m"]) # 0.668 (spread -1.22)
p_dec_2y = prob(b_2y, dec23["Spread_10y2y"]) # 0.329 (spread -0.44)
```

B Data Dictionary for `predicted_probabilities.csv`

Column	Type	Description
`Date`	YYYY-MM-DD	First of the month, 1959-01 to 2023-12 (781 rows)
`Actual_Target_12m`	0/1/empty	$Y_t = \mathrm{USREC}_{t+12}$, the realised 12-month-ahead label
`Prob_10y3m`	[0,1]	10y–3m model predicted $\Pr(Y_t = 1)$
`Prob_10y2y`	[0,1]/empty	10y–2y model predicted $\Pr(Y_t = 1)$ (empty before 1976-06)

References

[1] Michael D. Bauer and Thomas M. Mertens. Economic forecasts with the yield curve. Federal Reserve Bank of San Francisco Economic Letter 2018-07, March 2018.

[2] Menzie D. Chinn and Kavan J. Kucko. The predictive power of the yield curve across countries and time. *International Finance*, 18(2):129–156, 2015. doi: 10.1111/infi.12064.

[3] Eric C. Engstrom and Steven A. Sharpe. The near-term forward yield spread as a leading indicator: A less distorted mirror. Finance and Economics Discussion Series 2018-055, Board of Governors of the Federal Reserve System, Washington, DC, 2019.

[4] Arturo Estrella and Gikas A. Hardouvelis. The term structure as a predictor of real economic activity. *The Journal of Finance*, 46(2):555–576, 1991. doi: 10.1111/j.1540-6261.1991.tb02673.x.

[5] Arturo Estrella and Frederic S. Mishkin. The yield curve as a predictor of U.S. recessions. *Current Issues in Economics and Finance, Federal Reserve Bank of New York*, 2(7):1–6, 1996.

[6] Arturo Estrella and Frederic S. Mishkin. Predicting U.S. recessions: Financial variables as leading indicators. *The Review of Economics and Statistics*, 80(1):45–61, 1998. doi: 10.1162/003465398557320.

[7] Arturo Estrella and Mary R. Trubin. The yield curve as a leading indicator: Some practical issues. *Current Issues in Economics and Finance, Federal Reserve Bank of New York*, 12(5):1–7, 2006.

[8] Giovanni Favara, Simon Gilchrist, Kurt F. Lewis, and Egon Zakrajšek. Updating the recession risk and the excess bond premium. Feds notes, Board of Governors of the Federal Reserve System, Washington, DC, 2016.

[9] Robert Hall, Martin Feldstein, Ben Bernanke, Jeffrey Frankel, Robert Gordon, and Victor Zarnowitz. The NBER's business-cycle dating procedure. Business cycle dating committee report, National Bureau of Economic Research, Cambridge, MA, 2003.

[10] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[11] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Glenn D. Rudebusch and John C. Williams. Forecasting recessions: The puzzle of the enduring power of the yield curve. *Journal of Business & Economic Statistics*, 27(4):492–503, 2009. doi: 10.1198/jbes.2009.07213.

[14] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

[15] James H. Stock and Mark W. Watson. Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829, 2003. doi: 10.1257/002205103322436197.

[16] Jonathan H. Wright. The yield curve and predicting recessions. Finance and Economics Discussion Series 2006-07, Board of Governors of the Federal Reserve System, Washington, DC, 2006.