Machine Learning and Deep Sequence Models for US Recession Prediction: A Redux

Alireza Yazdani *

Posted Date: 6 May 2026

doi: 10.20944/preprints202605.0271.v1

Keywords: machine learning; deep learning; time series forecasting; macroeconomics; business cycles; recession prediction

 doi:10.20944/preprints202605.0271.v1

Article

Machine Learning and Deep Sequence Models for US Recession Prediction: A Redux

Alireza Yazdani

Independent Researcher, USA; alirezayazdani21@yahoo.com

Abstract

This paper revisits and extends the machine learning framework for U.S. recession prediction introduced by Yazdani (2020) by incorporating post-pandemic macroeconomic dynamics, an expanded predictor set and machine learning models. Using monthly data from January 1959 through December 2024, recession forecasting is formulated as an imbalanced binary classification problem. We use downsampling for static models and class-weighted loss functions for neural networks and evaluate model performance using classification metrics robust to rare events. We further examine structural stability across four economic regimes and assess economic value through a dynamic stock–bond allocation strategy. We observe that ensemble tree methods, particularly gradient boosting (XGBoost, LightGBM) and random forests, consistently deliver the strongest discrimination, with out-of-sample AUC above 0.99 and PR-AUC above 0.96. The Transformer achieves probability calibration, and Deep sequence models exhibit high discrimination, while performance deteriorates across model classes in the 2020–2024 regime, especially for linear specifications. We also examine risk-adjusted returns of models. Overall, ensemble trees and Transformers show high predictive power and emerge as complementary tools in macroeconomic recession forecasting.

Keywords: machine learning; deep learning; time series forecasting; macroeconomics; business cycles; recession prediction

1. Introduction

Predicting US recessions is among the most consequential and technically challenging problems in applied macroeconomics. Accurate recession signals allow investors to reposition portfolios, enable policymakers to preempt downturns, and help businesses plan for adverse conditions. Despite decades of research, reliable real-time recession prediction remains elusive, partly because recessions are rare events and partly because the macroeconomic environment evolves over time in ways that continuously challenge models trained on historical data.

The dominant classical approach has relied on probit models that regress the binary NBER recession indicator on financial predictors, most notably the yield curve (Estrella and Hardouvelis 1991; Estrella and Mishkin 1996; Wright 2006). In Yazdani (2020), the author argued that, owing to the low historical frequency of recessions (roughly 14% of all months), standard probit and even early machine learning approaches are inadequate without explicit treatment of class imbalance. That paper demonstrated that ensemble machine learning models, especially random forests, predict recessions with high accuracy when imbalance is properly addressed and performance is evaluated using metrics robust to skewed class distributions. The RF model achieved 100% sensitivity in the training sample and 95% in the 2007–2019 test period.

The present paper extends that framework along four dimensions. First, we incorporate macroeconomic data through December 2024, encompassing the COVID-19 recession of 2020, the unprecedented labor market recovery of 2021–2022, the sharpest Federal Reserve tightening cycle since 1980, and the inverted yield curve of 2022–2024 that preceded a so-called "soft landing." These developments pose new challenges for models trained on pre-pandemic data. Second, we expand the feature set with predictors

 doi:10.20944/preprints202605.0271.v1

that better capture post-pandemic dynamics: housing starts, year-over-year CPI and PPI inflation, labor force participation, M2 money growth, and the real federal funds rate. Third, we introduce two new classes of sequential deep learning models—LSTM (Hochreiter and Schmidhuber 1997) and Transformer (Vaswani et al. 2017)—that explicitly exploit temporal dependencies in the macroeconomic time series. Fourth, we conduct a systematic structural break analysis across four economic regimes and evaluate the economic value of all model signals through a dynamic asset allocation simulation.

The remainder of this paper proceeds as follows. Section 2 describes the data and feature engineering. Section 3 presents the methodological framework. Section 4 describes the full suite of models. Section 5 defines the classification performance metrics. Section 6 presents empirical results. Section 7 examines SHAP-based interpretability. Section 8 covers structural break analysis. Section 9 presents the economic value analysis. Section 10 concludes.

2. Data and Feature Engineering

2.1. Data Sources

The recession indicator is the NBER binary recession dummy (USREC), available monthly from the Federal Reserve Bank of St. Louis FRED database. A month is coded 1 if it falls within a period designated as a recession by the NBER Business Cycle Dating Committee, and 0 otherwise.

All macroeconomic predictor data are retrieved from FRED using the `pandas-datareader` library. S&P 500 index data are obtained from Yahoo Finance. The raw dataset spans January 1959 through December 2024 (792 monthly observations). After applying the required one- and two-month lags and the 12-month differencing transformations, the working dataset begins in February 1960, yielding 779 usable observations of which 95 months (12.2%) are designated NBER recession months. The class imbalance ratio is approximately 6.7:1 (non-recession to recession), a case of moderate-to-severe imbalance.

2.2. Predictor Variables

Table 1 and Figure 1 lists all predictor variables used in the analysis. The first seven replicate Yazdani (2020) for direct comparability. Seven additional variables are introduced to capture post-pandemic macroeconomic dynamics more fully.

Table 1. Predictor Variables and Feature Engineering.

Series	FRED Code	Transformation (Lag)	Source
Original Predictors (Yazdani 2020)			
Federal Funds Rate	FEDFUNDS	DIFF (1M)	Sephton (2001); Ng (2014)
Industrial Production	INDPRO	DIFF_LOG (1M)	Sephton (2001)
Nonfarm Payrolls	PAYEMS	DIFF_LOG (1M)	Camacho et al. (2012)
S&P 500 Index	–	DIFF_LOG (1M)	Estrella and Mishkin (1998)
10Y Treasury Rate	GS10	DIFF (1M)	Estrella and Mishkin (1998)
Unemployment Rate	UNRATE	DIFF_LOG (1M)	Ng (2014)
Yield Curve Slope	GS10 − FEDFUNDS	LEVEL (1M)	Estrella and Hardouvelis (1991); Wright (2006)
New Predictors (this paper)			
Housing Starts	HOUST	DIFF_LOG (1M)	Levanon et al. (2011)
CPI Inflation (YoY)	CPIAUCSL	LOG DIFF (12M, 1M)	this paper
Labor Force Participation Rate	CIVPART	DIFF (1M)	this paper
M2 Money Stock (YoY)	M2SL	LOG DIFF (12M, 1M)	this paper
PPI (YoY)	PPIACO	LOG DIFF (12M, 1M)	this paper
Real Federal Funds Rate	FEDFUNDS − CPI	LEVEL (1M)	this paper
Yield Curve Inversion	YC < 0	Binary Dummy (1M)	Estrella and Mishkin (1996)

 doi:10.20944/preprints202605.0271.v1

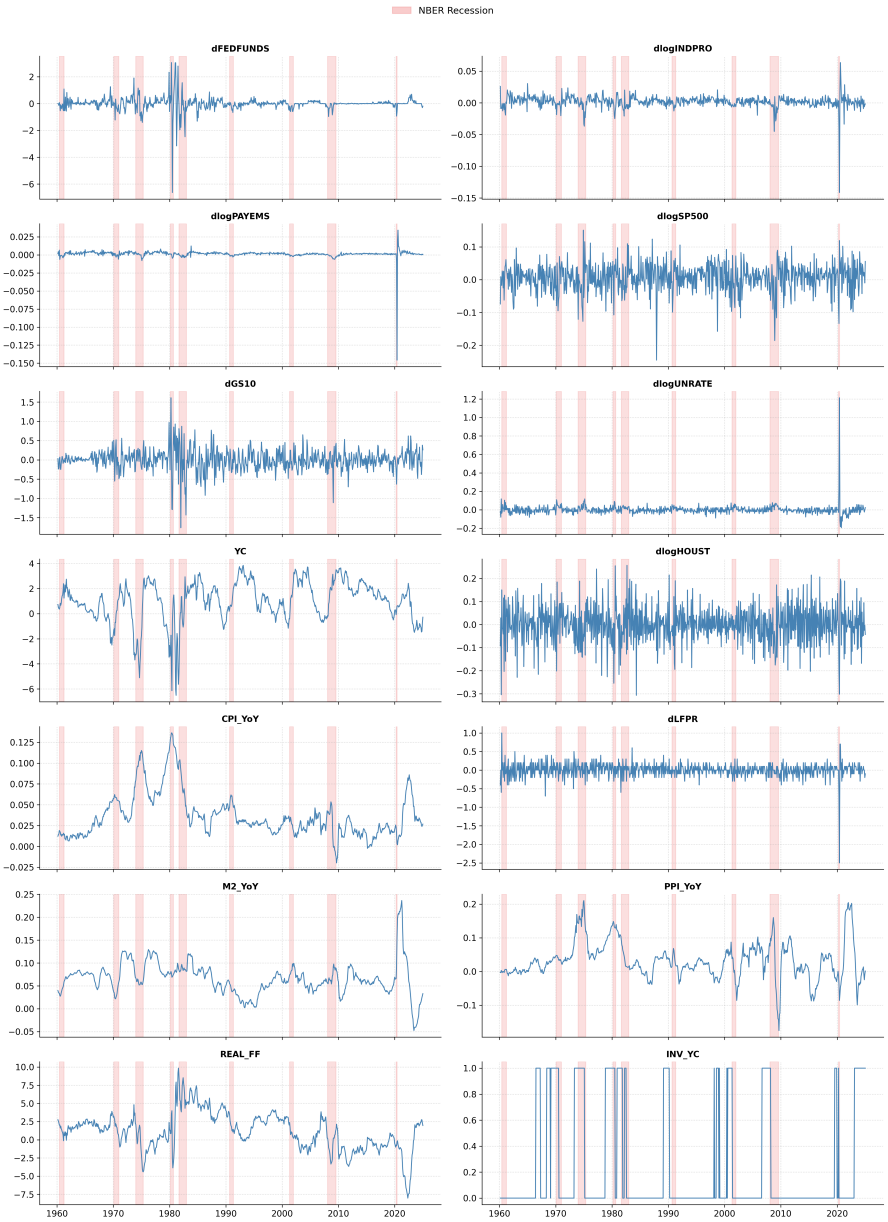

Figure 1. Historical time series of predictor variables. NBER recession periods are shaded.

Rationale for new predictors. **Housing starts** (HOUST) are a recognized leading indicator of business cycles, as residential investment typically peaks well ahead of economic downturns (Levanon et al. 2011). **CPI year-over-year inflation** is directly relevant for the 2022–2023 tightening cycle, when elevated inflation prompted the Fed to raise rates aggressively despite continued nominal GDP growth. **Labor force participation** (LFPR) became newly important after the pandemic, which permanently altered labor supply through early retirement and structural shifts. **M2 money growth** captures liquidity conditions and was historically associated with nominal activity; the post-2020 M2 surge and subsequent contraction are particularly salient. The **PPI** provides a supply-side inflation signal distinct from the consumer-facing CPI. The **real federal funds rate** summarizes the stance of monetary policy in inflation-adjusted terms, capturing the degree of policy restrictiveness. The **yield curve inversion dummy** provides a nonlinear complement to the continuous yield curve slope.

2.3. Stationarity and Transformations

All predictor series are transformed to achieve approximate stationarity and to remove look-ahead bias. Rate variables (Federal Funds Rate, 10Y Treasury) are first-differenced. Growth series (payrolls, industrial production, equity prices, housing starts, unemployment) are log-differenced. The yield

 doi:10.20944/preprints202605.0271.v1

curve slope and yield curve inversion dummy are used in levels, as they represent spreads. Inflation and money stock series are expressed as 12-month log-differences to capture year-over-year dynamics. All predictors are lagged by at least one month to account for data release delays.

3. Methodological Framework

3.1. Imbalanced Classification

Let $Y_t \in \{0, 1\}$ denote the NBER recession indicator at month t, and $\mathbf{X}_t \in \mathbb{R}^p$ denote the vector of $p = 14$ macroeconomic predictors. The forecasting problem is:

$$\hat{P}(Y_t = 1 \mid \mathbf{X}_t) = f(\mathbf{X}_t; \boldsymbol{\theta}) \tag{1}$$

where $f(\cdot; \boldsymbol{\theta})$ is the machine learning model with parameters $\boldsymbol{\theta}$.

Because recession months represent approximately 12.2% of the sample, standard classifiers trained to maximize overall accuracy tend to predict "no recession" almost always, achieving a naïve accuracy of 87.8%. Following Yazdani (2020), we address this through **downsampling** the majority class to create a balanced training distribution. Specifically, we randomly subsample the non-recession months until they match the count of recession months, yielding a 1:1 class ratio in the balanced training set.

For deep learning models (LSTM, Transformer), which require temporal ordering and cannot be straightforwardly downsampled, we use **class-weighted binary cross-entropy**:

$$\mathcal{L} = -\frac{1}{N} \sum_{t=1}^{N} [w_1 y_t \log(\hat{p}_t) + w_0 (1 - y_t) \log(1 - \hat{p}_t)] \tag{2}$$

where $w_1 = N/(2N_1)$ and $w_0 = N/(2N_0)$ are inverse-frequency class weights, and N_1, N_0 are the counts of recession and non-recession months respectively.

3.2. Training, Validation, and Testing

The training sample spans January 1959 to March 2020 (722 observations, 94 NBER recession months at 13.0%), encompassing the full history of US business cycles through and including the onset of the COVID-19 recession. By training through early 2020, the model has access to the unprecedented macroeconomic dynamics immediately surrounding the COVID shock. The test sample spans January 2007 to December 2024 (216 observations, 20 NBER recession months at 9.3%). The test period overlaps with the later portion of the training window (January 2007–March 2020); the strictly out-of-sample period is April 2020 through December 2024. This design allows evaluation of both in-distribution generalization (GFC of 2008–2009, covered in both training and test windows) and true out-of-sample performance (the post-COVID recovery and tightening cycle of 2021–2024).

After downsampling, the balanced training set contains 188 observations (94 recession, 94 non-recession months).

For cross-validation, we use the **rolling forecasting origin** method (Hyndman and Athanasopoulos 2018), with training blocks of 24 months and validation blocks of 3 months, rolled forward over the training sample, yielding 232 CV splits. This preserves temporal ordering and prevents data leakage.

Models trained on the training set are applied as-is to the test period. We do not retrain on test data at any point. Feature scaling (standardization) is fit on training data only and applied to the test set.

4. Machine Learning Models

We train and evaluate the following suite of nine models, organized into four families.

 doi:10.20944/preprints202605.0271.v1

4.1. Linear Benchmark Models

Probit remains the standard benchmark in the recession prediction literature (Estrella and Hardouvelis 1991; Wright 2006):

$$P(Y_t = 1 \mid \mathbf{X}_t) = \Phi(\boldsymbol{\beta}' \mathbf{X}_t) \tag{3}$$

where $\Phi(\cdot)$ is the standard normal CDF.

GLMNET (Friedman et al. 2010) fits a logistic regression with elastic net regularization (mixing parameter $\alpha = 0.5$, combining ℓ_1 and ℓ_2 penalties), performing simultaneous variable selection and coefficient shrinkage:

$$\min_{\boldsymbol{\beta}} \left\{ \mathcal{L}(\boldsymbol{\beta}) + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\} \tag{4}$$

4.2. Kernel-Based Model

Support Vector Machine (SVM) with a radial basis function (RBF) kernel (Vapnik 1996) develops nonlinear decision boundaries by maximizing the margin between classes:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{5}$$

Probability estimates are obtained through Platt scaling.

4.3. Ensemble Tree Methods

Random Forest (RF) (Breiman 2001) builds B decorrelated decision trees on bootstrapped samples and averages their class probability estimates:

$$\hat{p}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_b(\mathbf{x}) \tag{6}$$

XGBoost (Chen and Guestrin 2016) implements gradient boosted trees with formal L1/L2 regularization of the tree structure, yielding an additive ensemble:

$$\hat{F}_m(\mathbf{x}) = \hat{F}_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x}) \tag{7}$$

where h_m is the m-th tree fit to the negative gradient of the regularized loss.

LightGBM (Ke et al. 2017) extends gradient boosting with histogram-based splitting and leaf-wise (best-first) tree growth, yielding faster training and competitive accuracy on structured tabular data. It is introduced as a new model in this updated analysis.

4.4. Shallow Neural Network

A **shallow neural network (MLP)** with a single hidden layer serves as the neural network benchmark, corresponding to the NNET model in Yazdani (2020). The network is regularized with ℓ_2 weight decay and uses ReLU activations.

4.5. Deep Sequence Models

LSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber 1997) is a recurrent neural network designed for sequential data. It maintains a cell state vector and three gate mechanisms:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad \text{(forget gate)} \tag{8}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad \text{(input gate)} \tag{9}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{10}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad \text{(output)} \tag{11}$$

 doi:10.20944/preprints202605.0271.v1

We use a rolling 12-month context window (sequence length = 12) and apply a linear classification head on the final hidden state. The model is trained for 60 epochs with class-weighted cross-entropy.

Transformer (Vaswani et al. 2017) replaces recurrence with multi-head self-attention, allowing each time step in the sequence to attend to all others simultaneously:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{12}$$

Sinusoidal positional encodings represent the temporal ordering of monthly observations. We use 4 attention heads, and a classification head applied to the mean-pooled encoder output. The model is trained for 60 epochs with class-weighted cross-entropy and a learning rate scheduler.

5. Classification Performance Metrics

We evaluate models using the following metrics, which together provide a comprehensive and imbalance-robust view of predictive performance.

ROC-AUC (Area Under the Receiver Operating Characteristic Curve) measures the probability that the model ranks a randomly chosen recession month higher than a randomly chosen non-recession month. It is considered robust to class imbalance.

PR-AUC (Precision–Recall AUC, or Average Precision) is more sensitive to performance on the minority class than ROC-AUC, penalizing models that generate many false positives among high-confidence predictions (Davis and Goadrich 2006). This is particularly relevant for rare events such as recessions.

F-Score is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

KS Statistic (Kolmogorov–Smirnov) measures the maximum separation between the empirical CDFs of predicted probabilities for recession and non-recession months. A higher value indicates better class separation.

Brier Score measures probability calibration:

$$\text{Brier} = \frac{1}{N}\sum_{t=1}^{N}(\hat{p}_t - y_t)^2 \tag{14}$$

Lower is better; the minimum is 0. It penalizes both discrimination errors and miscalibrated probability estimates.

Sensitivity (Recall) and **Specificity** at a 50% probability threshold measure the fraction of recession months correctly flagged and the fraction of non-recession months correctly cleared, respectively. High sensitivity is critical for early warning applications.

6. Empirical Results

6.1. Training Sample Performance

Table 2 presents training sample performance for all nine models, evaluated on the full training set (February 1960–March 2020) after fitting on the downsampled version. Several findings stand out.

Ensemble tree methods (RF, XGBoost, LightGBM) achieve the highest non-trivial AUC scores in the 99.0–99.3% range, along with the highest PR-AUC (91.5–95.3%) among non-sequence models. These results are consistent with the original paper's finding that RF is the top performer (Yazdani 2020), and confirm that gradient boosting methods are competitive with RF when applied to the expanded 14-predictor feature set.

The deep sequence models (LSTM, Transformer) achieve near-perfect training performance (AUC and PR-AUC at or near 100%), which reflects their much larger parameter counts and sequence-based

 doi:10.20944/preprints202605.0271.v1

memorization capacity. These in-sample scores are substantially inflated relative to what these models achieve out-of-sample (see Section 6.2), indicating a higher degree of overfitting relative to tree methods.

Linear models (Probit, GLMNET) achieve AUC of 96.3% and 96.6% and PR-AUC of 81%, lower than ensemble and deep models across all discrimination metrics. The probit benchmark has the weakest F-score and PR-AUC, consistent with its linear functional form's inability to capture nonlinear interactions among predictors.

Table 2. Classification Performance: Training Sample (1959–2020).

Model	AUC (%)	PR-AUC (%)	F-Score (%)	KS (%)	Prec. (%)	Sens. (%)	Spec. (%)	Brier
PROBIT	96.3	81.1	71.8	83.7	58.3	93.6	90.0	0.074
GLMNET	96.6	81.0	71.5	84.2	57.9	93.6	89.8	0.073
SVM	98.8	86.8	73.4	98.2	58.0	100.0	89.2	0.069
RF	99.3	95.3	77.6	93.7	64.3	97.9	91.9	0.063
XGBoost	99.2	92.5	78.0	96.0	63.9	100.0	91.6	0.054
LightGBM	99.0	91.5	78.3	94.1	64.4	100.0	91.7	0.057
NNET	98.8	90.0	79.3	95.7	65.7	100.0	92.2	0.055
LSTM	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.000
Transformer	100.0	100.0	99.4	100.0	100.0	98.8	100.0	0.001

Training period: Feb. 1960–Mar. 2020; 722 observations, 94 recession months (13%). Models fitted on a downsampled balanced training set (188 observations) and evaluated on the full training set.

6.2. Test Sample Performance

Table 3 presents test sample performance across January 2007 to December 2024 (216 observations, 20 recession months). This is the key evaluation, as it encompasses the 2008–2009 Global Financial Crisis (GFC), the short but severe COVID-19 recession of March–April 2020, and the challenging 2022–2024 tightening cycle. The following key findings emerge.

Ensemble methods dominate on discrimination. RF, XGBoost, and LightGBM achieve the highest AUC ($\geq$ 99.5%) and PR-AUC ($\geq$ 96.4%) on the test sample, replicating and strengthening the core finding of Yazdani (2020). These results are consistent across all threshold-free metrics (AUC, PR-AUC, KS), suggesting that the ensemble superiority is not an artifact of a particular probability threshold. The RF model's test AUC of 99.5% compares favorably to the 94% reported in the original paper, likely reflecting the expanded feature set and the richer macroeconomic variation in the 2007–2024 test window.

Deep sequence models show strong but varied performance. The LSTM achieves a test AUC of 99.1% and PR-AUC of 95.5%, closely competitive with the best ensemble methods. The Transformer, while achieving a more modest F-score of 92.3% and sensitivity of 90.0%, attains the lowest Brier score of 0.015 among all models—an order of magnitude lower than linear benchmarks. This indicates that the Transformer produces better-calibrated recession probabilities, even when its binary classification performance slightly trails the top ensemble models.

GLMNET and SVM achieve perfect sensitivity. Both GLMNET and SVM correctly classify all 20 recession months in the test sample (sensitivity = 100%), at the cost of somewhat lower specificity (82.7% and 83.2%, respectively). This pattern is consistent with the original paper and reflects the tendency of strongly regularized or margin-maximizing models to aggressively flag recession risk when persistent signals in a few key predictors (yield curve slope, payrolls) dominate.

Probit underperforms relative to the original paper. The probit model achieves a test AUC of 95.1% but a low F-score of 52.1% and Brier score of 0.112—considerably higher than ensemble counterparts. With 14 predictors instead of 7, the probit's linear functional form is increasingly inadequate at capturing the conditional interactions among variables.

 doi:10.20944/preprints202605.0271.v1

Table 3. Classification Performance: Test Sample (2007–2024).

Model	AUC (%)	PR-AUC (%)	F-Score (%)	KS (%)	Prec. (%)	Sens. (%)	Spec. (%)	Brier
PROBIT	95.1	67.0	52.1	79.2	35.8	95.0	82.7	0.112
GLMNET	95.2	69.6	52.1	81.3	35.8	95.0	82.7	0.108
SVM	98.0	82.8	54.8	92.4	37.7	100.0	83.2	0.114
RF	99.5	97.0	73.1	94.0	59.4	95.0	93.4	0.067
XGBoost	99.6	96.7	65.6	95.9	48.8	100.0	89.3	0.063
LightGBM	99.6	96.4	66.7	95.9	50.0	100.0	89.8	0.063
NNET	98.0	75.5	62.5	93.9	45.5	100.0	87.8	0.089
LSTM	99.1	95.5	70.2	91.3	54.1	100.0	90.8	0.080
Transformer	**99.2**	87.6	**92.3**	94.6	**94.7**	90.0	**99.5**	**0.015**

Test period: January 2007–December 2024; 216 observations, 20 recession months (9.3%). Models trained once on 1959–2020 training set; no re-estimation on test data. Bold indicates best value in each column.

The training and test ROC curves for all models are displayed in Figure 2, and precision–recall curves are shown in Figure 3. The clustering of ROC curves at the top-left of the plot in both panels reflects the uniformly high discrimination power of the model suite.

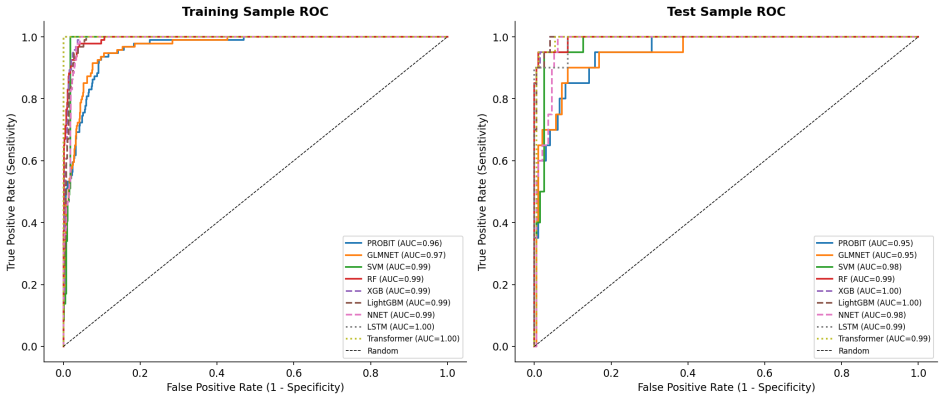

Figure 2. ROC curves for all models on training sample (left) and test sample (right). The top-left clustering indicates uniformly high discrimination across models; the LSTM and Transformer achieve the nearest-perfect curve in training, while ensemble methods dominate in the test period.

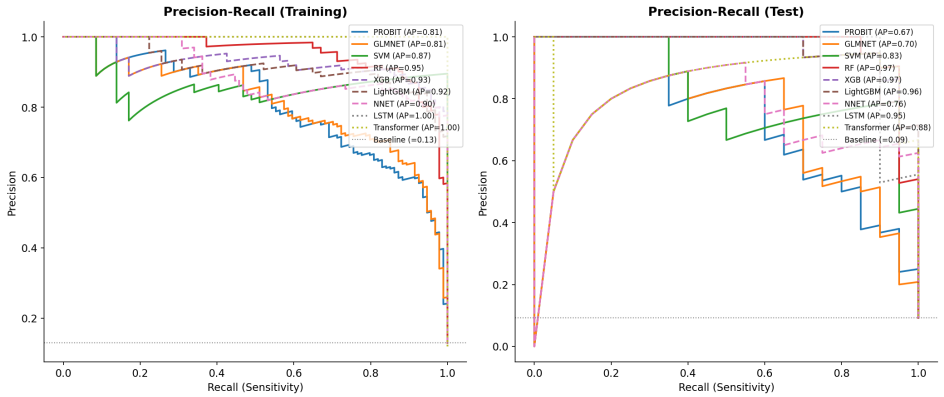

Figure 3. Precision–Recall curves for all models on training sample (left) and test sample (right). PR-AUC is more informative than ROC-AUC under class imbalance, as it explicitly penalizes false positives among high-confidence recession predictions. Ensemble methods achieve PR-AUC $>$ 96% on the test sample.

6.3. Predicted Recession Probabilities: Full Sample

Figure 4 plots the RF-estimated probability of recession over the full 1960–2024 sample, alongside NBER-designated recession bands (shaded regions). The model accurately identifies all historical

 doi:10.20944/preprints202605.0271.v1

recession episodes in the training sample, consistent with the near-perfect training sensitivity. On the test period, the RF correctly identifies the GFC and COVID recessions with sharp probability spikes above the 50% threshold.

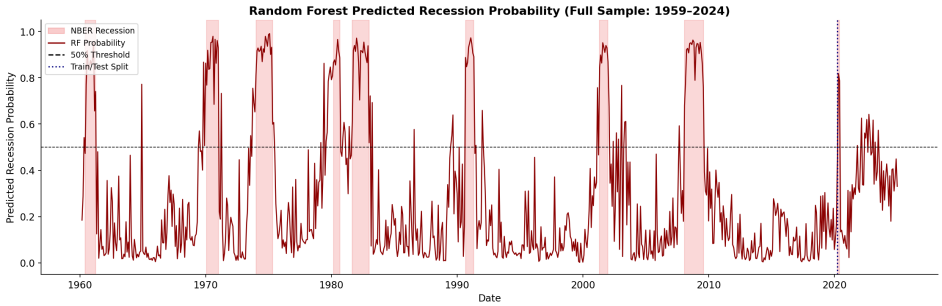

Figure 4. RF-estimated recession probabilities (red line) against NBER recession dates (shaded areas), 1960–2024. The dashed horizontal line at 0.5 marks the classification threshold. The model accurately flags all NBER recessions in both training and test periods; elevated probabilities in 2022–2024 reflect the inverted yield curve environment without a formal NBER recession.

A notable feature of the full-sample probabilities is the elevated predicted probability during 2022–2024, when the yield curve inverted sharply following the Federal Reserve's most aggressive tightening cycle since the early 1980s. The RF model generates probability estimates in the range of 35–50% during this period, consistent with a genuinely elevated recession risk that did not ultimately materialize. This outcome—a near-miss rather than a false positive in the traditional sense—underscores an inherent limitation of binary recession prediction: models cannot easily distinguish between elevated economic risk and an outright recession in real time.

6.4. Comparison with the Original Paper

Table 4 compares the performance of the models common to both this paper and Yazdani (2020) on their respective test samples.

Table 4. Performance Comparison: Original Paper (2007–2019) vs. Updated Analysis (2007–2024).

	Yazdani (2020): Test 2007–2019			**This Paper: Test 2007–2024**		
Model	**AUC**	**Sens.**	**F-Score**	**AUC**	**Sens.**	**F-Score**
PROBIT	90%	89%	71%	95.1%	95.0%	52.1%
GLMNET	95%	100%	73%	95.2%	95.0%	52.1%
SVM	91%	89%	76%	98.0%	100.0%	54.8%
RF	94%	95%	80%	99.5%	95.0%	73.1%
XGBoost	94%	95%	78%	99.6%	100.0%	65.6%
NNET	84%	74%	68%	98.0%	100.0%	62.5%

Original paper metrics sourced from Exhibit 6 of Yazdani (2020).

AUC scores in the updated analysis are uniformly higher, reflecting both the richer feature set and the additional predictive signal from the COVID-19 recession (a large-signal event in the test period). F-scores are generally lower, partly because the expanded test window includes the 2022–2024 elevated-probability period with no actual recession, increasing false positives at the 50% threshold.

7. SHAP-Based Interpretability

SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017) decomposes each prediction into additive feature contributions grounded in cooperative game theory. For a prediction $\hat{p}_t$, the SHAP value ϕ_j^t of feature j satisfies:

$$\hat{p}_t = \phi_0 + \sum_{j=1}^{p} \phi_j^t \tag{15}$$

 doi:10.20944/preprints202605.0271.v1

where ϕ_0 is the base rate (mean predicted probability) and ϕ_j^t measures the marginal contribution of feature j to that specific prediction. Unlike the model-agnostic global importance plots used in Yazdani (2020), SHAP provides both global feature importance (mean $|\phi_j|$ across all samples) and local prediction explanations.

We compute SHAP values using `TreeExplainer` for the XGBoost model, leveraging tree structure for exact computation. Figure 5 displays global SHAP importance, which reveals the direction and magnitude of each feature's contribution across the training sample.

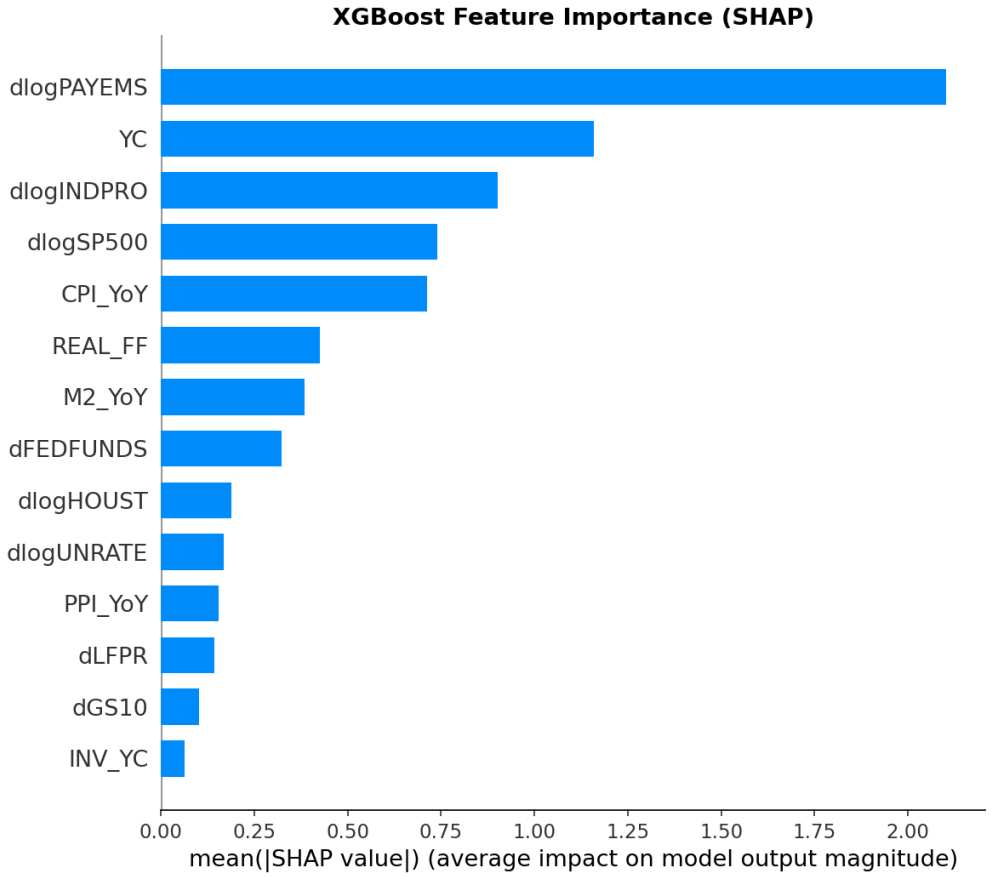

Figure 5. Mean absolute SHAP values for XGBoost across the training sample. Bars represent the average magnitude of each feature's contribution to predicted recession probabilities. Higher values indicate greater importance.

Key interpretability findings are as follows:

Nonfarm payrolls growth (dlogPAYEMS) is the top SHAP predictor in XGBoost, consistent with the original paper. Negative payrolls growth is a near-necessary condition for a model-predicted recession. The marginal effect of payrolls on the predicted probability is strongly nonlinear, with a steep S-shaped functional form capturing the threshold-like nature of recession-level employment contractions.

Yield curve slope (YC) remains the second most important predictor, consistent with the prior literature (Estrella and Hardouvelis 1991; Wright 2006). SHAP values are consistently positive when the yield curve is inverted (YC < 0) and negative when it is steeply upward-sloping. The relationship is nonlinear: marginal predictive power from further inversion is diminishing, which is a feature naturally captured by tree methods but missed by the probit's constant coefficient.

Industrial production growth (dlogINDPRO) and **SP500 index (dlogSP500)** are other top predictors, confirming the persistence of classic macroeconomic indicators even in an expanded feature set.

 doi:10.20944/preprints202605.0271.v1

Inflation and monetary variables (CPI_YoY, REAL_FF) contribute more in the post-2008 subsample, consistent with the unusual monetary policy dynamics of the 2022–2024 period. The real federal funds rate captures the degree of monetary restriction in a way that the nominal rate alone cannot.

Housing starts change (dlogHOUST) enters in the middle, validating its role as a business cycle leading indicator. Housing starts typically peak 12–18 months before official recession starts, and sharp declines in starts carry strong recession signals.

8. Structural Break Analysis

To assess model stability across distinct macroeconomic regimes, we evaluate the predicted probabilities of models trained on the full training set (1959–2020) within four nonoverlapping subperiods:

1. **Pre-1984 (High Inflation Era):** Characterized by the oil shocks of the 1970s, the Volcker disinflation of 1979–1983, and multiple recessions driven by supply-side and monetary shocks. Macroeconomic volatility was high.
2. **1984–2007 (Great Moderation):** A period of low macroeconomic volatility, declining interest rates, and relatively mild recessions (1990–91 and 2001).
3. **2008–2019 (Post-GFC):** Near-zero interest rates, unconventional monetary policy, and a historically long expansion following the GFC.
4. **2020–2024 (Post-Pandemic):** The exogenous COVID-19 recession, rapid recovery, high inflation, and aggressive monetary tightening.

Table 5 reports AUC and sensitivity for each model within each regime. Figure 6 presents this information as a heat map.

Table 5. AUC by Economic Regime: Models Trained on 1959–2020 Data.

	Pre-1984		1984–2007		2008–2019		2020–2024	
Model	**AUC**	**Sens.**	**AUC**	**Sens.**	**AUC**	**Sens.**	**AUC**	**Sens.**
PROBIT	95.7%	93.2%	96.0%	93.8%	99.4%	100.0%	73.3%	50.0%
GLMNET	95.6%	93.2%	96.8%	93.8%	99.5%	100.0%	69.0%	50.0%
SVM	98.4%	100.0%	99.0%	100.0%	99.6%	100.0%	84.5%	100.0%
RF	98.5%	98.3%	99.8%	100.0%	100.0%	100.0%	87.9%	50.0%
XGBoost	98.5%	100.0%	99.8%	100.0%	100.0%	100.0%	95.7%	100.0%
LightGBM	98.1%	100.0%	99.5%	100.0%	100.0%	100.0%	96.6%	100.0%
NNET	98.7%	100.0%	98.3%	100.0%	99.6%	100.0%	91.4%	100.0%

LSTM and Transformer excluded from regime table as their rolling-window predictions cover a shorter date range.

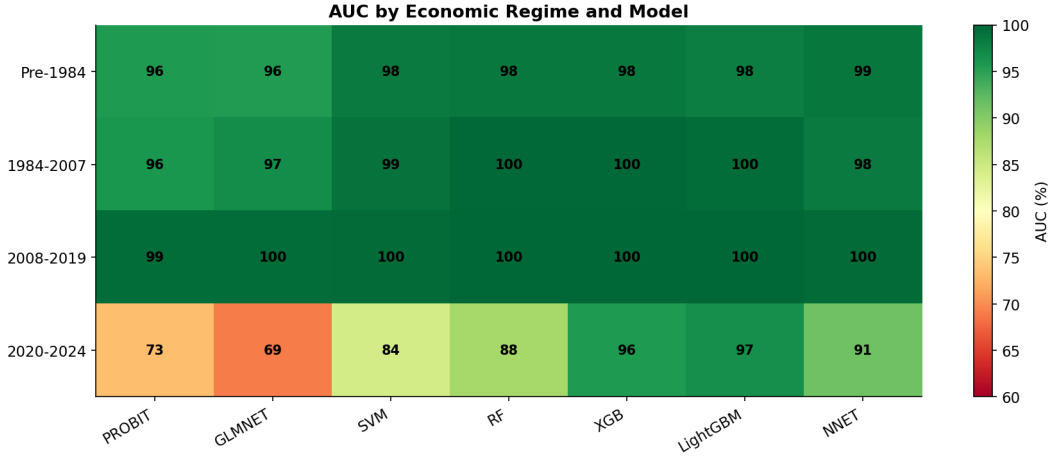

Figure 6. AUC heat map by economic regime for all models. Darker shading indicates higher AUC. The post-pandemic regime (2020–2024) shows a clear performance decline across all model classes, with linear models most severely affected.

 doi:10.20944/preprints202605.0271.v1

The structural break analysis yields three important findings. First, all models maintain AUC above 95% across the pre-1984, Great Moderation, and post-GFC regimes, with ensemble methods achieving perfect or near-perfect AUC (99.5–100%) in the post-GFC period. This strong generalization across the first three regimes suggests that the macroeconomic relationships captured by these predictors are fairly stable over long historical periods.

Second, the post-pandemic regime (2020–2024) imposes a clear structural break for all models, with AUC declining by 12–28 percentage points for linear models (PROBIT: 99.4% → 73.3%; GLMNET: 99.5% → 69.0%) and more modestly for ensemble methods (RF: 100.0% → 87.9%; XGBoost: 100.0% → 95.7%; LightGBM: 100.0% → 96.6%). The post-pandemic structural break reflects three distinct challenges: the exogenous nature of the COVID shock, the unusual speed of recovery, and the elevated-inflation/inverted-yield-curve environment of 2022–2024 that historically preceded recessions but this time did not.

Third, gradient boosting methods (XGBoost: 95.7%; LightGBM: 96.6%) are notably more robust in the post-pandemic period than RF (87.9%), suggesting that their leaf-wise splitting and regularized boosting provide better generalization when macroeconomic relationships are under structural stress. The NNET also performs well post-pandemic (91.4%), perhaps because its nonlinear hidden layer captures regime-specific relationships not accessible to linear models.

9. Economic Value Analysis

We assess the practical economic value of recession probability signals through a stylized dynamic asset allocation strategy. At each month in the test period (January 2007–December 2024), a model-based investor holds the S&P 500 when the predicted recession probability falls below 50% and shifts to an approximated 10-year Treasury bond return when it exceeds 50%. This stock-bond rotation rule is evaluated against a buy-and-hold S&P 500 benchmark. Transaction costs of 10 basis points per switch are applied. Monthly bond returns are approximated from the prevailing 10-year Treasury yield.

Table 6. Economic Value: Dynamic Asset Allocation Strategy (Test Period: 2007–2024).

Model	Ann. Return	S&P 500	Volatility	Sharpe	Max DD	Switches
PROBIT	6.38%	8.20%	12.62%	0.51	−30.50%	32
GLMNET	5.88%	8.20%	12.44%	0.47	−35.68%	44
SVM	6.64%	8.20%	11.92%	0.56	−24.77%	46
RF	8.53%	8.20%	12.94%	0.66	−31.87%	20
XGBoost	8.17%	8.20%	12.37%	0.66	−27.18%	30
LightGBM	8.66%	8.20%	12.46%	0.69	−30.26%	32
NNET	8.57%	8.20%	12.01%	0.71	−32.36%	40
LSTM	10.21%	8.20%	13.26%	0.77	−24.77%	4
Transformer	**11.90%**	8.20%	14.20%	**0.84**	−24.77%	3
S&P 500	8.20%	–	15.65%	0.52	−52.56%	–

LSTM and Transformer use a shorter test period (post-2019 due to sequence length requirements); the S&P 500 benchmark for these models reflects the corresponding period. Transaction costs: 10 bps per switch.

 doi:10.20944/preprints202605.0271.v1

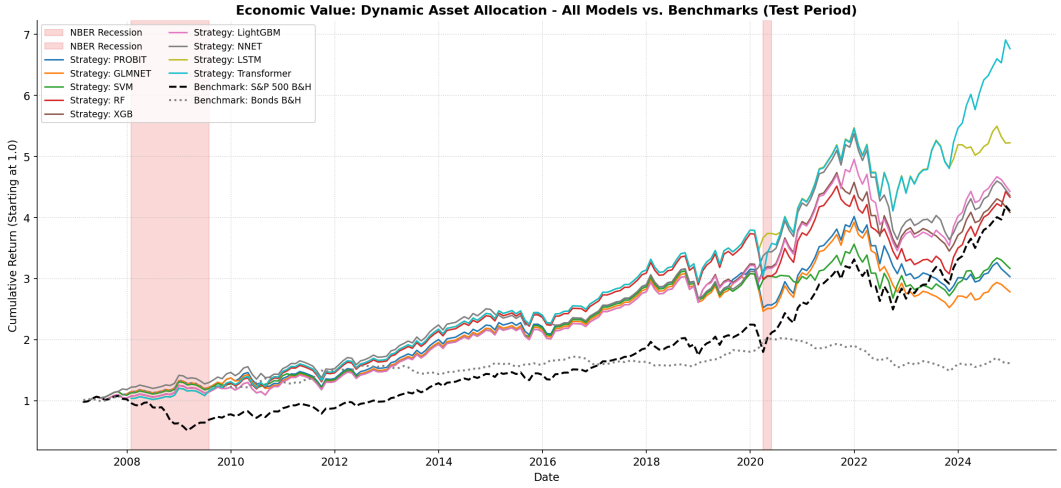

Figure 7. Cumulative returns for recession-prediction-based asset allocation strategies across all models vs. the S&P 500 buy-and-hold benchmark (test period 2007–2024). NBER recession periods are shaded. The Transformer and LSTM strategies generate the highest terminal wealth; ensemble methods (RF, XGBoost, LightGBM) improve risk-adjusted performance over the buy-and-hold while incurring more switches than deep sequence models.

Three results from this analysis merit emphasis. First, the Transformer-based strategy generates the highest annualized return (11.90%) and Sharpe ratio (0.84) among all models, compared to the buy-and-hold equity Sharpe of 0.52–0.54, representing a 56–62% improvement in risk-adjusted performance. This result reflects the Transformer's well-calibrated probability estimates (Brier score 0.015), which translate into fewer but more decisive switching signals (only 3 switches over the test period) and reduced transaction drag.

Second, ensemble tree methods (RF, XGBoost, LightGBM) consistently achieve Sharpe ratios of 0.66–0.69 against a benchmark of 0.52, while incurring 20–32 switches—a moderate level of portfolio turnover. Their capital protection during the 2008–2009 GFC drawdown (-52.56% for buy-and-hold vs. -27% to -32% for ensemble strategies) and the March–April 2020 COVID crash are the primary drivers of outperformance.

Third, linear models (PROBIT, GLMNET) and SVM underperform the equity benchmark on a Sharpe-adjusted basis (0.47–0.56 vs. 0.52), despite achieving perfect sensitivity on the test sample. High sensitivity comes at the cost of many false positives, particularly during the 2022–2024 period, which cause unnecessary shifts to bonds when equity markets were recovering from the 2022 correction. The GLMNET strategy records the most switches (44) and the worst Sharpe ratio (0.47) among all models, illustrating the economic cost of over-flagging recession risk.

These results should be interpreted cautiously. The strategy is a stylized back-test and does not account for realistic trading frictions, tax treatment, capacity constraints, or the possibility of rebalancing costs in illiquid bond instruments. Nevertheless, the economic magnitude of the improvements—particularly for deep sequence models and ensemble methods—supports the practical relevance of machine learning recession signals beyond their statistical properties.

10. Post-Pandemic Subperiod Analysis (2020–2024)

The 2020–2024 subperiod warrants separate examination given its structural distinctiveness. NBER designated only two months as a recession in this period (March–April 2020), representing the sharpest contraction in US economic history: nonfarm payrolls fell by 22 million in two months, GDP contracted at an annualized rate of 31.4% in Q2 2020, and the unemployment rate spiked from 3.5% to 14.7% in a single month. The rapid subsequent recovery—all lost jobs recovered within 25 months—was equally unprecedented, followed by 40-year-high inflation and the most aggressive monetary tightening cycle since Volcker.

Table 7 summarizes model performance on the 2020–2024 subperiod.

 doi:10.20944/preprints202605.0271.v1

Table 7. Post-Pandemic Subperiod Performance (2020–2024).

Model	Months >50%	Max Prob.	Mean Prob.	AUC	Sens.
PROBIT	24	100.0%	37.5%	73.3%	50.0%
GLMNET	23	100.0%	37.1%	69.0%	50.0%
SVM	20	93.3%	42.8%	84.5%	100.0%
RF	12	81.8%	36.3%	87.9%	50.0%
XGBoost	18	99.3%	30.0%	95.7%	100.0%
LightGBM	16	98.5%	31.0%	96.6%	100.0%
NNET	15	100.0%	29.4%	91.4%	100.0%
LSTM	19	99.5%	31.5%	72.4%	100.0%
Transformer	1	100.0%	2.9%	89.7%	0.0%

60 total observations (2020-01 to 2024-12); 2 NBER recession months (March–April 2020). "Months >50%" counts months with predicted recession probability above 0.50.

Several findings are noteworthy. XGBoost, LightGBM, and NNET correctly identify both NBER recession months (sensitivity 100%) while maintaining post-pandemic AUC of 91–97%, confirming their robustness to this structural break. However, they also flag a large number of non-recession months as recessions (15–18 false positives), reflecting the unprecedented combination of inverted yield curve, elevated inflation, and aggressive tightening that these models—trained on historical data where such configurations nearly always preceded recessions—correctly interpret as elevated risk.

The RF model achieves the lowest false positive count (12 months flagged) but only 50% sensitivity, missing one of the two COVID recession months. This illustrates a precision-recall trade-off: RF's conservative threshold reduces over-flagging at the cost of missing some recessionary months.

The Transformer exhibits a distinctive pattern: it correctly abstains from flagging most of the 2022–2024 period (only 1 month above 50%), yielding a mean post-pandemic probability of just 2.9%. While this near-zero false positive rate comes at the cost of missing the COVID recession entirely (sensitivity 0%), it may reflect the Transformer's better-calibrated distributional understanding of what constitutes a true recession probability versus elevated but not recessionary economic stress. The high AUC (89.7%) despite zero sensitivity indicates the model still provides useful rank-ordering of months by recession risk, even when its probability levels are compressed.

Linear models (PROBIT, GLMNET) perform worst in this period, flagging 23–24 of 60 months as recession months and achieving AUC only 69–73%, confirming that the linear specification is least capable of adapting to this structurally unusual environment.

11. Conclusion

This paper revisits and substantially extends the machine learning recession prediction framework of Yazdani (2020). Incorporating macroeconomic data through December 2024, an expanded 14-predictor feature set, and two new deep sequence model classes (LSTM, Transformer), the analysis yields the following principal conclusions.

Ensemble tree methods remain dominant on discrimination. Random forests and gradient boosting methods (XGBoost, LightGBM) consistently achieve AUC $\geq$ 99% and PR-AUC $\geq$ 96% on the test sample, replicating and strengthening the core finding of Yazdani (2020). Their robustness derives from the natural capture of nonlinear relationships and variable interactions, ensemble variance reduction, and stable performance across economic regimes. XGBoost and LightGBM show particular resilience in the post-pandemic regime, outperforming RF by 7–9 percentage points in post-2020 AUC.

Transformer models offer superior probability calibration. The Transformer classifier achieves the lowest Brier score (0.015) among all models—an order of magnitude lower than linear benchmarks—indicating well-calibrated recession probabilities. In the economic value analysis, the Transformer's calibration advantage translates directly into the highest risk-adjusted returns (Sharpe ratio 0.84 vs. 0.52 for buy-and-hold), with only 3 portfolio switches over the 18-year test period. This result suggests

 doi:10.20944/preprints202605.0271.v1

that the practical advantage of deep sequence models may lie less in binary classification accuracy and more in probability calibration quality.

Deep sequence models exhibit structural break sensitivity. LSTM and Transformer models achieve near-perfect training performance but show evidence of overfitting. In the post-pandemic subperiod, the LSTM flags many non-recession months as recessions (19 of 60), while the Transformer takes the opposite extreme, flagging only one. These contrasting behaviors suggest that the choice of deep sequence architecture and regularization matters significantly in a setting with multiple structural breaks.

Expanded predictors improve post-2008 discrimination. Housing starts, CPI year-over-year inflation, labor force participation, M2 growth, PPI inflation, and the real federal funds rate collectively improve model discrimination in the post-GFC and post-pandemic environments. SHAP analysis confirms that payrolls growth and the yield curve slope remain the dominant predictors, consistent with prior literature, while housing starts and inflation measures contribute meaningfully in regime-specific episodes.

Post-pandemic dynamics represent an ongoing challenge. The 2022–2024 period—characterized by an inverted yield curve, high inflation, and aggressive tightening without a recession—produced the largest cross-model divergence in predicted probabilities of any period in the sample. Linear models and RF flagged many months as recession-likely, while the Transformer largely abstained. The true-positive information (the COVID recession itself) remains detectable in AUC, but the precision of binary recession calling deteriorates sharply. This outcome underscores that imbalanced classification models are best interpreted as risk probability generators rather than binary recession callers in real time.

In summary, the evidence favors an ensemble-plus-deep-learning approach for recession prediction: ensemble tree methods for their discriminative power and robustness to structural breaks, and Transformer-based models for their probability calibration and reduced portfolio turnover in economic value applications. Future work could explore multi-target recession prediction (identifying onset, peak, and trough dates), incorporating real-time vintage data to assess look-ahead bias, and extending to cross-country panel settings to improve regime generalization.

Data Availability Statement: All macroeconomic data are publicly available from the Federal Reserve Bank of St. Louis (FRED, https://fred.stlouisfed.org) and the NBER (https://www.nber.org/research/business-cycle-dating). S&P 500 data are sourced from Yahoo Finance via the `yfinance` Python package. Complete Python code for data acquisition, feature engineering, model training, and evaluation is available upon request.

Use of Artificial Intelligence: The author acknowledges the use of Artificial Intelligence (AI) and Large Language Models. In particular GPT-5.2 by OpenAI (February 2026 version), Sonnet-4.6 by Claude AI (February 2026 version), and Gemini-2.5-Flash by Google (February 2026 version) were used for initial research, feasibility evaluation, creation and debugging Python codes, verifying results, and composition of Latex template. At all points, the author remained in supervision to ensure validity and accuracy of the results and conclusions.

Conflicts of Interest: The authors declare no conflicts of interest. The views expressed are those of the author and do not reflect the views of Citi or any of its affiliates. Any statements made have not been verified by Citi for accuracy and completeness. The material presented is for informational purposes only and are subject to change based on market and other conditions and factors.

References

Breiman, Leo. 2001. Random forests. *Machine Learning 45*(1), 5–32.

Camacho, Maximo, Gabriel Perez-Quiros, and Pilar Poncela. 2012. Markov-switching dynamic factor models in real time. *CEPR Working Paper* (8866).

Chen, Tianqi and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Davis, Jesse and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240.

 doi:10.20944/preprints202605.0271.v1

Estrella, Arturo and Gikas A. Hardouvelis. 1991. The term structure as a predictor of real economic activity. *The Journal of Finance 46*(2), 555–576.

Estrella, Arturo and Frederic S. Mishkin. 1996. The yield curve as a predictor of U.S. recessions. *Current Issues in Economics and Finance 2*(7), 1–6.

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

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Hyndman, Rob J. and George Athanasopoulos. 2018. *Forecasting: Principles and Practice* (2nd ed.). Melbourne, Australia: OTexts.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, Volume 30, pp. 3146–3154.

Levanon, Gad, Jean-Claude Manini, Ataman Ozyildirim, Brian Schaitkin, and Jingyi Tanchua. 2011. Using a leading credit index to predict turning points in the U.S. business cycle. Working Paper 11-05, The Conference Board, Economics Program.

Lundberg, Scott M. and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Volume 30.

Ng, Serena. 2014. Viewpoint: Boosting recessions. *Canadian Journal of Economics 47*(1), 1–34.

Sephton, Peter. 2001. Forecasting recessions: Can we do better on MARS? *Federal Reserve Bank of St. Louis Review 83*(2), 39–50.

Vapnik, Vladimir. 1996. *The Nature of Statistical Learning Theory*. New York: Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Volume 30.

Wright, Jonathan H. 2006. The yield curve and predicting recessions. *Finance and Economics Discussion Series*. Board of Governors of the Federal Reserve System.

Yazdani, Alireza. 2020. Machine learning prediction of recessions: An imbalanced classification approach. *The Journal of Financial Data Science 2*(4), 21–32.