IMES DISCUSSION PAPER SERIES

Forecasting Recessions Using Machine Learning on Text Data and Mixed-Frequency Predictors

Yusuke Oh and Mototsugu Shintani

Discussion Paper No. 2026-E-7

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

2-1-1 NIHONBASHI-HONGOKUCHO
CHUO-KU, TOKYO 103-8660
JAPAN

IMES Discussion Paper Series 2026-E-7
March 2026

Forecasting Recessions Using Machine Learning on Text Data and Mixed-Frequency Predictors

Yusuke Oh* and Mototsugu Shintani**

Abstract

We forecast Japanese recessions by integrating machine learning methods, mixed-frequency data, and text-based indicators within an unrestricted mixed data sampling (U-MIDAS) framework. The model combines monthly macroeconomic variables with weekly financial indicators and newspaper-based text indicators. A pseudo-real-time forecasting exercise over three decades shows that machine learning models consistently outperform traditional logit benchmarks. The model confidence set (MCS) suggests horizon dependence: Text indicators are more informative at short horizons, while financial variables are more informative at longer horizons. To improve interpretability, we apply sparse principal component analysis (Sparse PCA) to the text indicators and identify three economic narratives: 'Corporate Distress,' 'Financial Distress,' and 'Deflationary Pressure.' Furthermore, SHAP (SHapley Additive exPlanations) analysis indicates that different recession episodes are associated with different combinations of these narratives, underscoring the heterogeneous nature of economic downturns.

Keywords: business cycles; mixed data sampling; model confidence set; text analysis; recession forecasting

JEL classification: C32, C53, E37, O53

*Deputy Director, Institute for Monetary and Economic Studies, Bank of Japan (E-mail: yuusuke.ou@boj.or.jp)
**The University of Tokyo (E-mail: shintani@e.u-tokyo.ac.jp)

The views expressed in this paper are those of the authors and do not necessarily reflect the official views of the Bank of Japan. The authors thank Shin-ichi Fukuda, Keiichi Goshima, Masahiro Higo, Daisuke Ikeda, Shosei Sakaguchi, Toshiaki Watanabe, Fan Dora Xia, and the participants of the 45th International Symposium on Forecasting, The 8th International Conference on Econometrics and Statistics, and the 2025 BOK/ERI- BOJ/IMES Joint Research Workshop.

1 Introduction

Economic forecasting is an essential part of conducting monetary policy, which is necessarily forward-looking. Forecasting inflation has gained attention as many forecasters, including central banks, failed to project the inflation surge that materialized after the COVID-19 pandemic. However, forecasting recessions is arguably even more challenging, given that recessions occur only intermittently over the business cycle. While an extensive literature exists on predicting U.S. recessions, corresponding research on Japan remains limited.

More importantly, while the literature on recession forecasting has explored promising avenues, advancements have mainly occurred in three parallel, largely independent streams. First, the application of machine learning methods to recession forecasting has expanded rapidly, with Vrontos et al. (2021) demonstrating that these techniques can outperform traditional approaches. Second, Galvão and Owyang (2022) show that incorporating timely information through mixed-frequency data, particularly via the mixed data sampling (MIDAS) framework, significantly improves prediction accuracy. Third, Pierdzioch and Gupta (2020) highlight the value of text-based indicators in capturing market sentiment and external shocks. However, efforts to integrate these three developments into a unified framework remain scarce.

Our paper aims to fill this gap by constructing a unified framework to forecast recessions. We synthesize three recent streams of innovation into a coherent analytical approach: (1) methodologically, by applying a wide range of machine learning models combined with the unrestricted MIDAS (U-MIDAS) framework; (2) empirically, by incorporating a comprehensive set of predictors including financial cycle measures and text-based indicators; and (3) evaluatively, by using the model confidence set (MCS) procedure (Hansen et al., 2011) for robust model comparison. This integrated approach allows us to systematically test the marginal benefit of modern techniques and novel data sources, addressing multiple-testing concerns inherent in such a large-scale comparison.

We apply the unified forecasting framework to the Japanese data. Our empirical analysis yields several notable findings. First, in line with the international evidence,

machine learning models consistently outperform traditional logistic regression in forecasting recessions across various time horizons. Second, term spreads and financial variables remain crucial predictors, particularly for longer forecast horizons. Third, text-based indicators materially improve model performance, especially at shorter horizons, demonstrating their ability to capture rapidly evolving economic conditions. Fourth, we find that the marginal benefits of adding weekly predictors are limited when sufficiently informative monthly predictors are already included in the model. Finally, by applying sparse principal component analysis (Sparse PCA) to the text data, we uncover interpretable underlying components, such as 'Corporate Distress,' 'Financial Distress,' and 'Deflationary Pressure,' that offer valuable insights into the drivers of recession risk.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on recession forecasting. Section 3 explains the U-MIDAS framework and machine learning models. Section 4 describes our dataset, including traditional economic indicators and text-based metrics. Section 5 describes the methods used to evaluate the out-of-sample forecasting performance of the competing models. Section 6 presents our main findings and their implications. Section 7 provides an in-depth interpretation of the text-based indicators through Sparse PCA. Finally, Section 8 concludes.

2 Literature Review

The literature on forecasting recessions using financial and economic indicators has a rich history dating back several decades. This section reviews the key developments in this field, focusing on seminal contributions, methodological advancements, and variable extensions that have shaped our understanding of recession forecasting.

2.1 Earlier Work on Yield Curve and Recession Forecasting

The influential work on using the yield curve to forecast recessions was conducted by Estrella and Mishkin (1996, 1998). These studies demonstrated that the term spread, particularly the difference between 10-year and 3-month Treasury rates, is a robust pre-

dictor of recessions up to eight quarters ahead in the United States. Their research showed that the shape of the yield curve reflects underlying cyclical economic conditions, with an inverted yield curve (where short-term rates exceed long-term rates) typically preceding economic downturns.

This relationship between yield curve inversion and subsequent recessions can be explained through several channels. First, monetary policy tightening tends to raise short-term rates while having less impact on long-term rates, leading to a flattening or inversion of the yield curve and subsequently slowing economic growth. Second, the term spread contains information about market expectations of future economic conditions, with inversions reflecting pessimistic outlooks among market participants.

2.2 Methodological Extensions

Building on these foundational insights into the predictive power of the yield curve, substantial methodological advances have been made in recent years in recession forecasting techniques, particularly through the application of machine learning algorithms and mixed-frequency data methods.

Vrontos et al. (2021) made a notable contribution by introducing various machine learning techniques for U.S. recession forecasting. Their study found that machine learning models, particularly tree-based methods such as Random Forests and Gradient Boosting, consistently outperform traditional logit and probit models. The superior performance of these methods stems from their ability to capture nonlinear interactions between predictors and to adapt to complex, dynamic economic environments without requiring a priori specification of functional forms.

In parallel, Galvão and Owyang (2022) developed a MIDAS-Probit model that leverages high-frequency (weekly) term spread data to forecast recessions. Their approach demonstrated substantial improvements in prediction accuracy compared to models using only monthly data. By employing the MIDAS framework, they were able to incorporate the most current market information while avoiding the noise often present in high-frequency financial data, resulting in more timely and accurate recession forecasts.

Given the proliferation of models and methods, a key methodological challenge is how to rigorously compare their performance. When comparing numerous model configurations, the risk of false discovery increases substantially. The MCS framework (Hansen et al., 2011) addresses this issue by providing a set of models constructed to contain the best model with a given confidence level, rather than selecting a single "best" model. The MCS procedure sequentially eliminates models that are found to have significantly inferior predictive ability, using a test for equal predictive ability and an elimination rule. The resulting set size reflects data informativeness: less informative data yield a larger MCS, while more informative data yield a smaller one. This procedure provides model-specific p-values and helps mitigate multiple-comparison concerns by eliminating models only when there is statistically significant evidence of inferior predictive ability.

2.3 Variable Extensions in Recession Forecasting

Beyond methodological innovations, research has also expanded the range of predictors used for recession forecasting, moving beyond the traditional focus on the yield curve alone.

Borio et al. (2020) demonstrated that financial cycle measures, particularly the debt service ratio (DSR), provide stronger predictive power than the term spread across both advanced and emerging economies. The DSR, calculated as the ratio of interest payments to income, effectively captures the financial stress that can precede economic downturns. Their work highlighted the importance of incorporating dynamic financial variables to better understand cyclical fluctuations and market risks.

The integration of text-based indicators represents another important extension. Pierdzioch and Gupta (2020) showed that volatility indicators derived from news articles (NVIX) are powerful predictors for U.S. recessions. These text-based metrics capture market sentiment and external shocks that may not be immediately reflected in traditional numerical indicators. Their research emphasized the value of incorporating qualitative information from text sources to complement conventional quantitative predictors.

2.4 Research on Japanese Recessions

While international research has grown rapidly, studies focusing specifically on Japan remain limited. Japan's unique economic environment, characterized by prolonged periods of near-zero interest rates and deflationary pressures, presents distinct challenges for recession forecasting that differ from other advanced economies.

Early work by Hirata and Ueda (1998) used a probit model to evaluate the predictive power of term spreads for Japanese recessions, finding them useful but less effective than in the U.S. context. Similarly, Bernard and Gerlach (1998) examined term-spread predictability across advanced economies and found Japan's term spread to be only weakly predictive, possibly due to the country's strict financial regulations.

Another strand of literature employs sophisticated econometric models to estimate recessions. Watanabe (2003) effectively applied a Markov-switching dynamic factor model to detect turning points using components of the CI. As we discuss in Section 4, while the focus was commonly on identifying turning points using CIs, the use of the LI is often more appropriate for forecasting purposes. For instance, Miyazaki (2016) used a similar Markov-switching framework but applied it to the LI specifically to test its viability as an early-warning signal for recessions.

More recent studies have explored alternative indicators for Japanese recessions. Okimoto and Takaoka (2017), for instance, found that credit spreads of medium-grade corporate bonds have stronger predictive power for the coincident index than government bond yields, albeit not for the binary recession event itself.

Despite these advances, existing research on Japanese recession forecasting has several important limitations. First, studies have typically focused on individual methodologies or predictors in isolation, without systematically comparing the performance of modern machine learning techniques against traditional approaches. Second, the integration of text-based indicators, which have shown promise in international studies, remains largely unexplored for Japan. Third, the potential benefits of using mixed-frequency data in the Japanese context have not been thoroughly investigated. Finally, most existing studies have not incorporated the recent developments in financial cycle measures alongside

traditional predictors. Our paper addresses these gaps by providing a comprehensive, unified framework that integrates modern methodologies, diverse data sources, and rigorous model comparison procedures specifically tailored to the Japanese economy.

3 Methodology

This section outlines our methodological approach to forecasting Japanese recessions. We explain the U-MIDAS framework that allows us to effectively incorporate mixed-frequency predictors, and describe the machine learning models employed in our analysis. Specific implementation details, including lag structures, forecast horizons, and validation procedures, are provided in Section 5.

3.1 U-MIDAS Framework

Mixed-frequency data presents a fundamental challenge in economic forecasting: how to effectively incorporate information from variables sampled at different frequencies. Traditional approaches often involve aggregating higher-frequency data to match the lowest frequency in the dataset, which can lead to a loss of potentially valuable information. The MIDAS framework, developed by Ghysels et al. (2005), addresses this challenge by allowing direct use of data at different frequencies within the same model.

While the original MIDAS approach uses parametric weighting functions to handle high-frequency lags, we employ the unrestricted MIDAS (U-MIDAS) variant proposed by Foroni et al. (2015). The U-MIDAS approach is particularly well-suited for cases where the frequency mismatch is not too large (e.g., monthly and weekly data, as in our case) and offers greater flexibility by not imposing a specific functional form on the lag weights.

For binary recession forecasting, we incorporate a link function into the U-MIDAS specification:

$$P(Y_{t+h} = 1|\Omega_t) = G\left(\alpha + \sum_{j=0}^{p_m-1} \beta_j X_{t-j}^{(m)} + \sum_{k=0}^{p_w-1} \gamma_k Z_{s(t)-k}^{(w)}\right) \tag{1}$$

where Y_{t+h} is a binary indicator equal to 1 if the economy is in recession in month $t+h$ and 0 otherwise. The function $G(\cdot)$ denotes the logistic link function, Ω_t denotes the information set available at time t, $X_t^{(m)}$ and $Z_t^{(w)}$ denote monthly and weekly predictors, respectively, and the superscripts indicate the sampling frequency. Moreover, $s(t)$ denotes the index of the last week ending in month t. This specification avoids approximating the number of weeks in a month by a fixed constant (e.g., 4.3) and instead aligns weekly observations with the monthly target by anchoring them to the end of the month. Unlike parametric MIDAS, U-MIDAS allows the data to determine the lag weights without imposing a functional-form restriction on how they decay over time.

The U-MIDAS framework offers several key advantages for our analysis. First, it preserves the rich information contained in high-frequency data, which is particularly valuable for financial variables that may react quickly to changing economic conditions. Second, it provides natural compatibility with machine learning methods, as the unrestricted approach simply expands the feature space with additional lags, which machine learning algorithms are generally well-equipped to handle through their built-in regularization mechanisms. Third, it allows us to directly test the marginal contribution of high-frequency information by comparing models with and without weekly predictors.

3.2 Machine Learning Models

Traditional linear probability models, such as logistic regression, may fail to capture the complex, nonlinear relationships relevant to recession forecasting. In contrast, machine learning models offer a more flexible alternative by adapting to dynamic economic relationships without requiring pre-specified functional forms.

We consider a diverse set of machine learning models, largely following Vrontos et al. (2021). In particular, we use three penalized linear classifiers (Lasso, ridge, and elastic net), support vector machines with RBF kernels, two tree-based ensembles (random forests and LightGBM), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and shallow neural networks (with one or two hidden layers). The specific hyperparameter configurations and model training procedures are provided in Section 5. In addition

to the nine machine learning methods described above, we also consider a traditional logistic regression as a benchmark, resulting in a total of ten forecasting models.

4 Data

We construct a comprehensive set of predictors for Japanese recessions by combining traditional economic variables with text-based indicators. Our monthly dataset spans from January 1992 to December 2024. This sample covers more than three decades, including significant episodes such as the Lost Decade, the Global Financial Crisis, the Great East Japan Earthquake, and the COVID-19 pandemic.

4.1 Recession Dating and Leading Economic Indicators

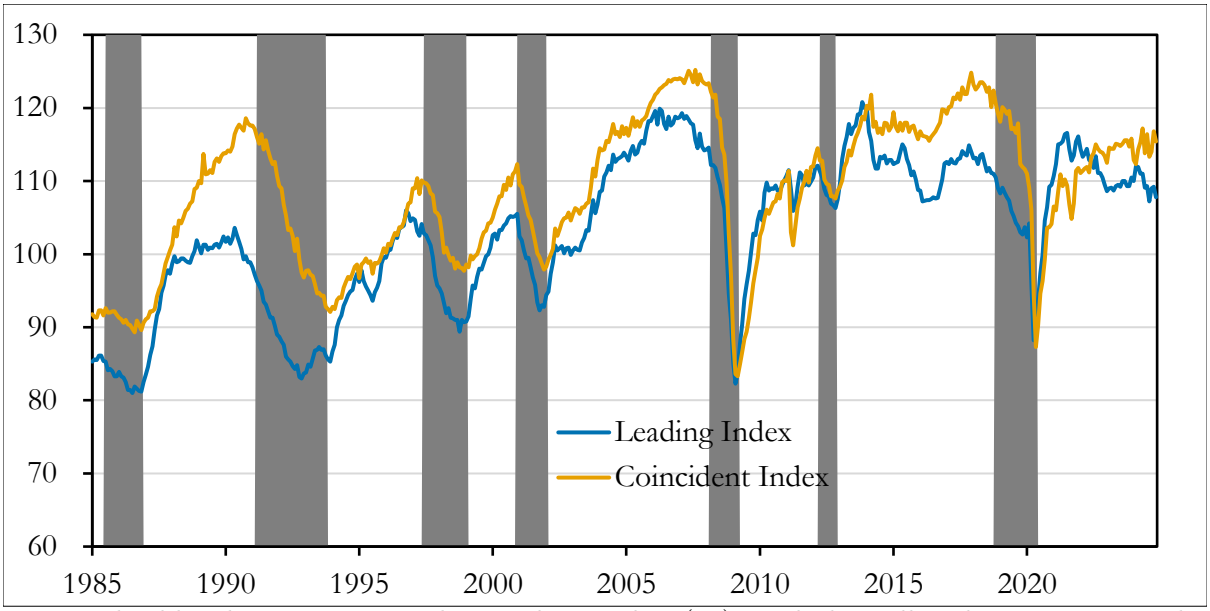
The target variable Y_{t+h} is a binary indicator for recession periods, based on the official business cycle chronology determined by the Cabinet Office's Economic and Social Research Institute (ESRI). These dates are established primarily through the Coincident Composite Index (CI), which tracks current economic conditions.

Alongside the CI, the Cabinet Office maintains the Leading Composite Index (LI) as part of the Business Conditions Index System, explicitly designed to provide advance signals of business cycle turning points. This predictive orientation has made the LI a natural choice for forecasting applications. As shown in Figure 1, the LI generally leads the CI by several months, with its turning points preceding the peaks and troughs of the business cycle. This lead relationship provides the empirical foundation for using LI components as predictive features.

We adopt the 11 disaggregated components of the LI (variables 1-11 in Table 1) as our baseline predictor set. These components capture diverse aspects of the Japanese economy, including production and inventories (1-2), labor markets (3), investment (4-5), confidence indicators (6, 10-11), market prices (7, 9), and monetary aggregates (8).

Using individual components rather than the composite index is well-suited to our machine learning approach. Modern machine learning methods can flexibly learn opti-

Figure 1. Trends of the Leading and Coincident Indexes



Note: The blue line represents the Leading Index (LI), and the yellow line represents the Coincident Index (CI). Shaded regions indicate official recession periods as determined by the Cabinet Office.
Source: Cabinet Office, Government of Japan.

mal weightings across high-dimensional predictor sets, potentially improving upon the fixed aggregation scheme of the composite index while allowing us to assess the relative importance of different economic dimensions. All LI components are available at monthly frequency only and form the foundation of our predictor set across all specifications.

4.2 Term Spreads and Financial Variables

Following the literature on recession forecasting, we include several term spread measures (variables 12-14 in Table 1) calculated from Japanese government bond yields. Specifically, we construct three term spreads: 10-year minus 1-year, 5-year minus 1-year, and 3-year minus 1-year yields. These data are obtained from the Japanese Ministry of Finance and are available at both monthly and weekly frequencies, making them suitable candidates for our U-MIDAS framework.

We also incorporate the Debt Service Ratio (DSR) (variable 15 in Table 1), calculated as the ratio of interest expenses to operating income based on the Business Outlook Survey conducted by the Ministry of Finance. This measure aims to capture the corporate interest payment burden, which can signal financial stress preceding economic downturns

as demonstrated by Borio et al. (2020). The DSR is available at monthly frequency only.

Additionally, we include the realized volatility of the Nikkei 225 stock index (variable 16 in Table 1) as a measure of market uncertainty, calculated as the standard deviation of daily returns within each month or week. This variable is available at both monthly and weekly frequencies.

4.3 Text-Based Indicators

A distinctive feature of our dataset is the inclusion of text-based indicators (variables 17-19 in Table 1) constructed from Japanese newspapers. We extract these metrics from the business section of Mainichi Shimbun, one of Japan's major daily newspapers, covering the period from January 1992 to December 2024 at a daily frequency.

We construct these text-based indicators using methodologies developed in prior studies:

- **Macroeconomic Sentiment (Sentiment-M):** Based on the methodology of Goshima et al. (2022), this indicator uses a domain-specific dictionary of 874 words related to economic conditions, each assigned a sentiment score of +1 (positive) or -1 (negative). The daily sentiment score is calculated as the net balance of positive and negative words appearing in articles, normalized by the total word count.

- **Financial Market Sentiment (Sentiment-F):** Following Ito et al. (2018), this measure employs a larger dictionary of 19,630 words specifically related to financial market analysis, with continuous sentiment scores ranging from -1 to +1. This dictionary was constructed through machine learning methods applied to financial news and captures more nuanced market sentiment than binary classification.

- **Economic Policy Uncertainty (EPU):** Based on the methodology of Baker et al. (2016), this index quantifies uncertainty regarding economic policy by counting the frequency of articles containing terms related to economic policy uncertainty. Specifically, we count articles that simultaneously mention terms related to (1) the

economy, (2) policy, and (3) uncertainty. The raw counts are then normalized and scaled to have a mean of 100 over the sample period.

These text-based indicators are aggregated to both monthly and weekly frequencies to align with our mixed-frequency modeling approach. Monthly values are calculated as averages over all days in each month, while weekly values are averages over seven-day periods. As shown in Figure 2, all three indicators exhibit pronounced movements during recession episodes, though with distinct patterns. Sentiment-M and Sentiment-F tend to decline sharply during recessions, reflecting deteriorating tone in economic and financial reporting. The EPU index, by contrast, spikes upward during periods of policy uncertainty, with particularly large increases during the Global Financial Crisis and the COVID-19 pandemic.

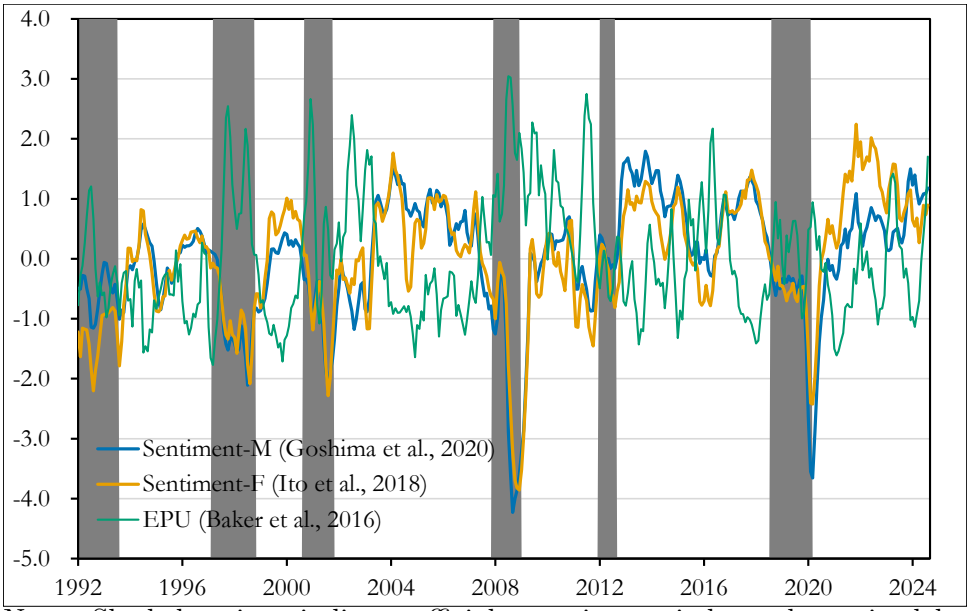
Table 1 summarizes all predictors used in our analysis, indicating which variables are available at weekly frequency (denoted W) and which are available only at monthly frequency (denoted M). This distinction is crucial for our U-MIDAS implementation, as only variables with weekly availability can contribute high-frequency information to our forecasting models. The last column (Transform) indicates the transformation applied to each variable.

Table 1. Overview of Predictors

	Variable	Frequency	Transform
	Baseline Predictors (Leading Index Components)		
1	Final Demand Goods Inventory Ratio Index	M	Δ
2	Producer Goods Inventory Ratio Index	M	Δ
3	New Job Offers (excluding new graduates)	M	Δ
4	Real Machinery Orders (Manufacturing)	M	Δ
5	New Housing Construction Floor Area	M	Δ
6	Consumer Confidence Index†	M	Δ
7	Nikkei Commodity Index	M	Δ
8	Money Stock (M2)	M	Δ
9	Tokyo Stock Price Index (TOPIX)	M	Δ
10	Investment Environment Index (Manufacturing)	M	Δ
11	Small Business Sales Forecast DI†	M	Δ
	Term Spreads		
12	Term Spread (10Y-1Y)	W	Level
13	Term Spread (5Y-1Y)	W	Level
14	Term Spread (3Y-1Y)	W	Level
	Financial Variables		
15	Debt Service Ratio (DSR)	M	Δ
16	Nikkei 225 Realized Volatility	W	Level
	Text-based Indicators		
17	Macroeconomic Sentiment	W	Level
18	Financial Market Sentiment	W	Level
19	Economic Policy Uncertainty (EPU) Index	W	Level

Note: M denotes monthly series; W denotes weekly series. Δ denotes simple 3-month level differencing: $\Delta x_t = x_t - x_{t-3}$, applied uniformly to all LI components and DSR, including price indices such as TOPIX and the Nikkei Commodity Index. "Level" indicates the variable enters the model without differencing. †DI-based series with bounded scales. A logistic transformation $100 \times \ln[(U - x)/(x - L)]$ is applied before differencing to map the bounded scale to an unbounded range, where U and L are the upper and lower bounds of the original scale (Vermeulen, 2012).

Figure 2. Time Series of Text-Based Indicators



Note: Shaded regions indicate official recession periods as determined by ESRI. All series are standardized to have zero mean and unit variance, and smoothed using a 90-day moving average for display purposes only. The original daily data are aggregated to monthly and weekly frequencies for use in our forecasting models.

5 Empirical Design

This section outlines our empirical strategy for evaluating the performance of different models and data combinations in forecasting Japanese recessions. We detail our model specification approach, forecast evaluation framework, and the statistical procedures used to identify superior forecasting models.

5.1 Model Specifications and Variable Selection

Existing research on recession forecasting has typically examined individual methodological or data innovations in isolation. For instance, some studies compare the predictive power of different variables within a single model framework (e.g., probit models), while others assess the benefits of mixed-frequency data using a specific modeling approach or evaluate machine learning techniques using data at a single frequency. While these studies have advanced our understanding, they leave open the question of which elements

(variable selection, modeling approach, or data frequency) contribute most to forecasting performance unconditionally, and whether their effects interact. To address this gap, we adopt a comprehensive experimental design that systematically controls for the presence or absence of each element, allowing us to isolate and evaluate the marginal contribution of each innovation.

To systematically assess the contribution of different predictor sets and methodologies, our strategy involves the following dimensions:

1. **Base predictor set:** We always include the 11 disaggregated components of the LI (described in Section 4) as our foundation, representing the standard approach to Japanese economic forecasting.

2. **Term spread inclusion:** We evaluate models both with and without the three term spread measures (variables 12-14 in Table 1).

3. **Financial variables inclusion:** Independently, we evaluate models both with and without additional financial variables, namely the Debt Service Ratio (DSR) and Nikkei 225 realized volatility (variables 15-16 in Table 1).

4. **Text indicator selection:** We examine four possible configurations: (1) no text-based indicator, or options (2) through (4), each using one of our three text-based metrics (Sentiment-M, Sentiment-F, or EPU). We do not include multiple text indicators simultaneously to maintain model parsimony and avoid multicollinearity.

5. **Data frequency:** We compare pure monthly models against those incorporating weekly data through the U-MIDAS framework (described in Section 3). Note that weekly frequency is only available for term spreads, realized volatility, and text-based variables (indicated by W in Table 1), not for the LI components or the DSR.

This approach yields 32 distinct feature combinations per model type ($2 \times 2 \times 4 \times 2 = 32$). With one benchmark logistic regression and nine machine learning models (described

in Section 3), we evaluate a total of 320 model configurations ($32 \times 10 = 320$) for each prediction horizon.

For lag structure, we use 2 lags for monthly variables and 8 weekly lags (equivalent to approximately 2 monthly lags) for weekly variables in the U-MIDAS framework, maintaining temporal consistency across frequencies. We evaluate predictive performance across three forecast horizons:

- **Short-term:** 3 months ahead ($h = 3$)
- **Medium-term:** 6 months ahead ($h = 6$)
- **Long-term:** 12 months ahead ($h = 12$)

These horizons correspond to typical monetary policy transmission lags and business planning cycles, allowing us to assess both near-term and medium-term forecasting ability.

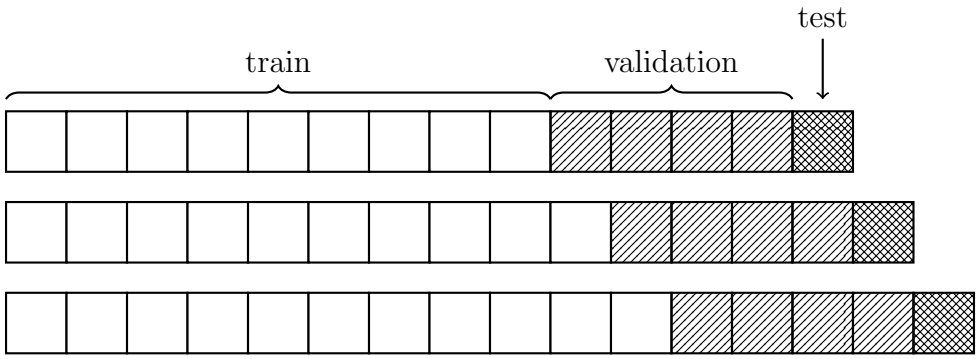
5.2 Model Training and Validation Procedure

Our sample period spans from January 1992 to December 2024. To ensure robust model training while preventing look-ahead bias, we implement a pseudo-real-time forecasting scheme with quarterly model re-estimation. We focus our evaluation on out-of-sample forecasts for the period from January 2003 to December 2024, which includes several recession episodes associated with the Global Financial Crisis in 2008, the Great East Japan Earthquake in 2011, and the COVID-19 pandemic in 2020.

Figure 3 illustrates our research design, which features an expanding training window, a rolling validation window for hyperparameter tuning, and a test period for out-of-sample evaluation. The forecasting procedure can be described as follows:

1. **Initial training:** Models are first estimated using data from January 1992 to December 2002.
2. **Re-estimation of the model:** We re-estimate all models on the first day of each quarter (January, April, July, October) using only information available at

Figure 3. Research Design Overview



Note: The white region represents the expanding training window, the light gray region with diagonal lines represents the 72-month rolling validation window used for hyperparameter tuning, and the dark gray region with cross-hatching represents the test period for out-of-sample evaluation. Models are re-estimated quarterly, with the training window expanding and the validation window rolling forward at each point of forecast.

the point of prediction. It should be noted that, due to the limited availability of real-time vintages for Japanese economic data, we rely on final revised series for the hard data. In contrast, text-based indicators are never revised and thus accurately reflect the real-time information set.

3. **Hyperparameter tuning:** The training set spans from the start of the sample up to 72 months before the prediction point, and the validation set comprises the subsequent 72 months.[1] Based on the validation set, hyperparameters are selected to minimize the logarithmic(log) loss function given by

$$-\frac{1}{T}\sum_{t=1}^{T}\left[Y_{t+h}\log(\hat{p}_t)+(1-Y_{t+h})\log(1-\hat{p}_t)\right] \tag{2}$$

where Y_{t+h} is the recession indicator (0 or 1) and $\hat{p}_t = \hat{P}(Y_{t+h} = 1|\Omega_t)$ is the predicted probability from the model.[2] During this step, all non-binary predictors are standardized (to zero mean and unit variance) using the mean and variance of the training set.

[1]The average duration of business cycles in postwar Japan has been approximately 50 months. Thus, our choice of a 72-month window is sufficient to cover at least one cycle, ensuring that the binary recession indicator take a value of 1. See the Cabinet Office's analysis of postwar Japanese business cycles for more details (https://www5.cao.go.jp/keizai3/2004/1219nk/04-00201.html).

[2]The log loss penalizes confident misclassifications heavily and provides a proper scoring rule for probabilistic forecasts.

4. **Out-of-sample forecast evaluation:** Based on selected hyperparameters, the final model is estimated from the training and validation data combined. For this final estimation and subsequent out-of-sample evaluation, predictors are standardized using the mean and variance of the combined sample. Hyperparameter selection and final model estimation are conducted separately for each forecast horizon. We repeat these steps by expanding the training sample until the end of the test period reaches December 2024.

6 Results

This section presents our empirical findings on the performance of various models and data combinations in forecasting Japanese recessions across different time horizons. We first present the top-performing specifications, and then provide a more comprehensive assessment using the model confidence set procedure.

6.1 Top-Performing Models based on AUC

Let us first evaluate the model in terms of the **area under the receiver operating characteristic curve (AUC)**. The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate across all possible classification thresholds, and the AUC summarizes this relationship in a single metric. The AUC ranges from 0.5 (random classification) to 1.0 (perfect classification), with higher values indicating better discriminatory power. The AUC is particularly appropriate for recession forecasting because it is threshold-independent, robust to class imbalance (recessions are rare events), and has a clear interpretation: the probability that the model ranks a randomly chosen recession month higher than a randomly chosen non-recession month.

Table 2 presents the top three model specifications for each forecast horizon based on out-of-sample AUC, along with the benchmark logistic regression for comparison. Machine learning models consistently outperform the benchmark logistic regression, with sizable AUC improvements across all horizons. Several notable patterns emerge: KNN

Table 2. Top Three Model Specifications by Forecast Horizon

Horizon	Model	Term Spread	Financial Variables	Mixed Frequency	Text	AUC
3-Month	KNN	×	×	✓	Sentiment-M	0.93
	LightGBM	×	×	✓	Sentiment-M	0.91
	KNN	✓	✓	×	Sentiment-M	0.91
	Logit (Benchmark)	×	×	×	×	*0.66*
6-Month	LightGBM	✓	✓	×	Sentiment-M	0.87
	KNN	✓	✓	✓	Sentiment-M	0.86
	KNN	✓	✓	×	Sentiment-F	0.86
	Logit (Benchmark)	×	×	×	×	*0.55*
12-Month	LightGBM	✓	✓	×	×	0.88
	KNN	✓	✓	×	×	0.86
	KNN	✓	×	✓	×	0.86
	Logit (Benchmark)	×	×	×	×	*0.55*

Note: Term Spread includes variables 12–14; Financial Variables include DSR and realized volatility (variables 15–16); Mixed Frequency indicates whether weekly data is incorporated via U-MIDAS (✓) or only monthly data is used (×); Text indicates which text-based indicator (if any) is included. All models include the 11 LI components as baseline predictors. The benchmark logistic regression model uses only these components without additional predictors or regularization.

performs best for short-term forecasts while LightGBM leads in longer horizons; text indicators (particularly Sentiment-M) appear consistently in top models for short and medium-term forecasts but are absent from long-term forecasts; conversely, term spreads and financial variables are largely absent from short-term models but appear in all top long-term models. These horizon-dependent patterns suggest fundamentally different predictive dynamics across time scales, which we examine more rigorously through the model confidence set procedure.

6.2 Model Confidence Set Analysis

To identify superior models while addressing the multiple testing problem inherent in comparing 320 model configurations, we employ the **model confidence set (MCS)** procedure developed by Hansen et al. (2011).

The MCS procedure constructs a set of models that contains the best model with a given level of confidence, rather than attempting to identify a single "best" model. The procedure sequentially tests whether all remaining models have equal expected loss and

eliminates the worst-performing model if the null is rejected, continuing until the null hypothesis cannot be rejected. The resulting MCS contains all models that cannot be statistically distinguished from the best model. This approach offers several advantages: it controls the family-wise error rate across multiple comparisons, accounts for the variability in relative performance across different time periods, and provides a statistically principled way to identify multiple competitive models. The size of the MCS itself is informative: a smaller MCS indicates more decisive evidence about model superiority.

Table 3 summarizes the characteristics of selected models that remained in the 95% model confidence set (i.e., at the 5% significance level).

Table 3. Models included in the MCS

Horizon	Total	Term	Financial	Text		Mixed
	(out of 320)	Spread	Variables	Any	Sentiment-M	Frequency
3-Month	27	44%	44%	96%	85%	48%
6-Month	34	68%	50%	62%	53%	53%
12-Month	24	75%	58%	25%	17%	46%

Note: The significance level for the MCS procedure is set to 5% at each stage, using the log loss function. Percentages indicate the share of remaining models that include each feature category. "Any" denotes the share of remaining models that include at least one text indicator (Sentiment-M, Sentiment-F, or EPU). "Sentiment-M" specifically reports the inclusion rate of the macroeconomic sentiment indicator. "Mixed Frequency" indicates the share of models using weekly data via U-MIDAS for applicable variables.

Machine Learning vs. Benchmark The benchmark logistic regression is excluded from the MCS for any forecast horizon, indicating strong evidence in favor of machine learning models. This result motivates our focus on a broad set of model classes and suggests that more flexible approaches better capture the dynamics relevant for Japanese recessions.

Term Spreads and Financial Variables Both term spreads and financial variables show clear monotonic increases in inclusion rates across horizons. Term spreads rise from 44% (3-month) to 68% (6-month) to 75% (12-month), while financial variables increase from 44% to 50% to 58%. This progression provides strong statistical support for their role as long-leading indicators. The low inclusion rates for short-term forecasts suggest these

variables contribute little to near-term prediction accuracy, while their high inclusion at long horizons confirms they are essential for identifying recessions several quarters ahead (Estrella and Mishkin, 1996; Borio et al., 2020). The similar patterns suggest they capture partially overlapping but distinct dimensions of longer-term recession risk.

Text-Based Indicators Sentiment-M exhibits the opposite pattern, with inclusion rates falling from 85% (3-month) to 53% (6-month) and 17% (12-month). This result suggests that Sentiment-M is essential for near-term dynamics while being less informative at longer horizons. The sharp decline confirms a clear horizon-dependent pattern in the effectiveness of news sentiment.

Mixed Frequency Data Mixed frequency inclusion rates remain relatively stable (48%, 53%, 46%) across horizons, showing no systematic advantage over monthly-only specifications. Combined with the limited appearance in top models, this result suggests that when comprehensive monthly indicators are available, the U-MIDAS framework provides at best modest benefits. While Galvão and Owyang (2022) find that weekly-sampled term spreads improve recession forecasts over monthly-sampled spreads for the U.S., their models employ a relatively small set of predictors, primarily the Chicago Fed National Activity Index and a few financial variables. In contrast, our models already incorporate 11 LI components alongside additional economic and financial predictors at monthly frequency. This richer monthly information set may reduce the marginal value of exploiting higher-frequency variation, suggesting that mixed-frequency approaches provide greater benefits when monthly predictors are limited rather than comprehensive.

Key Findings In summary, the MCS procedure yields three key insights. First, machine learning models outperform traditional benchmarks, which fail to remain in the MCS at any forecast horizon. Second, predictor importance is strongly horizon-dependent: Sentiment-M consistently ranks among the best for short-term forecasts (85% inclusion rate) but its relative importance diminishes at long horizons (17%), while term spreads and financial variables show the opposite pattern, with inclusion rates rising from

44% to 75% as the horizon extends. Third, mixed-frequency data provides limited systematic benefits when comprehensive monthly indicators are available. These findings suggest that optimal recession forecasting requires horizon-specific model configurations rather than a one-size-fits-all approach.[3]

6.3 Robustness and Model Diagnostics

While the superior performance of machine learning models is evident, understanding the drivers of these predictions is crucial for policy implementation. A particular concern in the Japanese context is the potential impact of unconventional monetary policy on predictor reliability. Since the Bank of Japan launched Quantitative and Qualitative Monetary Easing (QQE) in April 2013 and subsequently introduced Yield Curve Control, the prolonged compression of the yield curve may have fundamentally altered the information content of term spreads—one of the most widely used recession predictors in the literature.

To investigate the stability of predictor relationships across monetary policy regimes and to identify the sources of prediction errors, we provide a detailed SHAP-based analysis in Appendix C. There, we demonstrate that while traditional yield curve signals have indeed diminished substantially in the post-QQE period, text-based indicators have gained relative importance, further justifying their inclusion in our unified framework.

7 Interpretation of Text-Based Indicators

While our results demonstrate the predictive value of text-based indicators, particularly for short-horizon forecasts, the interpretation of these metrics poses challenges due to their high dimensionality and complex construction. In this section, we apply sparse principal component analysis (Sparse PCA) to extract interpretable components from our text data and examine their relationship with recession episodes.

[3]We also applied the MCS procedure separately within individual model classes. The results for LightGBM and KNN—the two best-performing algorithms—are reported in Appendix D. The model-specific inclusion rates are broadly consistent with the patterns documented here.

7.1 Sparse Principal Component Analysis

We apply Sparse PCA to the 874-dimensional daily time series of macroeconomic sentiment scores (Sentiment-M) to extract interpretable components.[4] Unlike standard PCA, Sparse PCA imposes an L1 penalty that forces many loadings to be exactly zero, resulting in components that load on a small subset of meaningful terms. This sparsity greatly enhances interpretability while preserving the key information contained in the text-based indicators.

We extract three principal components using Sparse PCA, with implementation details provided in Appendix A. This choice balances interpretability with information retention, allowing us to capture the main narrative structures in the text data.

We selected Sparse PCA over other dimension reduction techniques such as Latent Dirichlet Allocation (LDA) for several reasons. First, Sparse PCA better preserves sentiment polarity, which is crucial for our recession forecasting application. Second, it yields more stable components over time compared to topic models. Third, it does not require distributional assumptions about the underlying data generation process.

Table 4. Terms and loadings for each principal component

PC1: Corporate Distress		PC2: Financial Distress		PC3: Deflationary Pressure	
Term	Loading	Term	Loading	Term	Loading
Bankruptcy	0.58	Financial Crisis	0.54	Deflation	0.70
Non-performing Loans	0.56	Deterioration	0.48	Deflationary Pressure	0.65
Business Failure	0.47	Worsening	0.40	Stock Price Decline	0.29
Layoffs	0.37	Bankruptcy	0.33		
Unrealized Losses	0.02	Recession	0.32		
		Production Cut	0.19		
		Crisis	0.01		
		Stabilization	-0.07		

Note: This table lists all terms with non-zero loadings in each principal component. See Appendix A for technical details on the Sparse PCA implementation.

Table 4 presents the terms and their loadings for the three leading components extracted from our macroeconomic sentiment data. Based on these loadings, we can meaningfully interpret each component:

[4]Component loadings are derived from the full sample to ensure semantic continuity for interpretation. Recursive estimation would update loadings at each step, changing the definition of components over time and making consistent economic interpretation impossible.

- **Component 1: Corporate Distress** - This component loads heavily on terms related to business failures, bankruptcy, non-performing loans, layoffs, and unrealized losses. It captures the financial distress experienced by Japanese corporations during economic downturns.

- **Component 2: Financial Distress & Recession** - This component captures terms related to systemic financial stress and broader economic deterioration, including financial crisis, deterioration, worsening, and recession. The loading on "Production Cut" indicates that this component reflects not only financial sector distress itself but also its transmission to the real economy through production adjustments.

- **Component 3: Deflationary Pressure** - This component loads primarily on terms related to deflation, deflationary pressure, and stock price declines, reflecting Japan's unique experience with persistent deflationary pressures.

Importantly, when we replaced the aggregate Sentiment-M index with these three sparse principal components (PCs) in our machine learning models, we achieved comparable out-of-sample performance, with AUC values of 0.91, 0.87, and 0.82 for short, medium, and long-term predictions, respectively.[5] This result suggests that these interpretable components capture the essential information contained in the broader sentiment metrics.

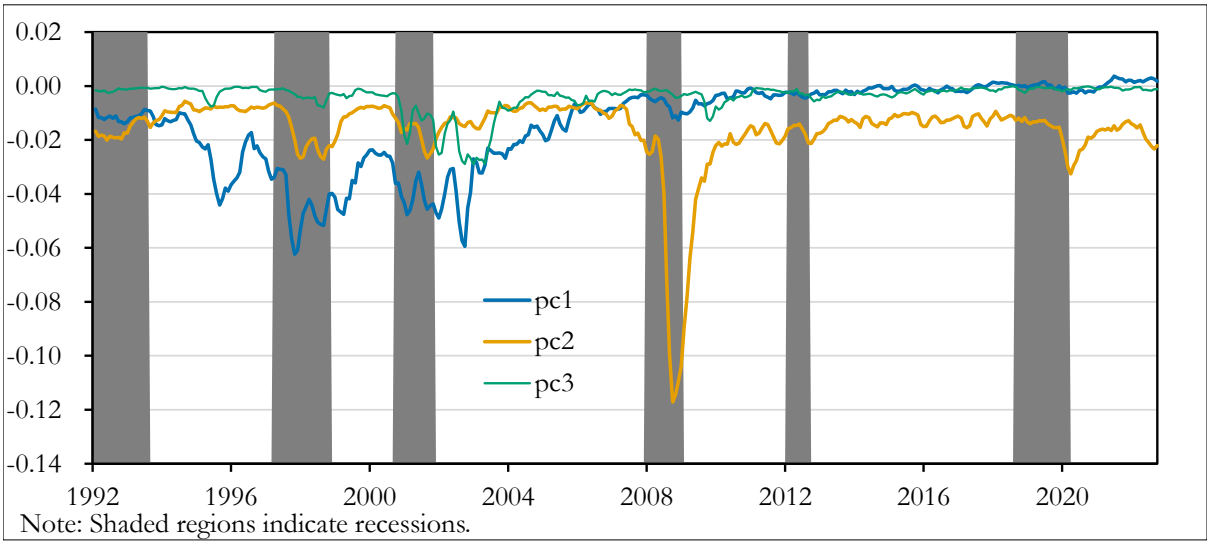
7.2 Component Dynamics Across Recession Episodes

Figure 4 displays the time series of our three sparse PCs across different recession episodes. Each component exhibits distinct temporal dynamics: the Corporate Distress component (PC1) shows sustained elevations during the early 2000s and again following the 2011

[5]We also conduct the MCS procedure in the same manner as in Section 6. The inclusion rates for Sentiment-M and sparse PCs are 33% and 33% at the 3-month horizon, 33% and 33% at the 6-month horizon, and 19% and 43% at the 12-month horizon, respectively. For this comparison, the other predictors are fixed (term spread = yes, financial variables = yes, mixed frequency = monthly only), with three text options: none, Sentiment-M, and sparse PC. Together with 10 model types (including Logit and machine-learning methods), there are total of $1 \times 1 \times 3 \times 1 \times 10 = 30$ specifications.

earthquake, while the Financial Distress component (PC2) displays a sharp spike concentrated in 2008–2009. The Deflationary Pressure component (PC3) peaks during the Global Financial Crisis but also registers modest elevations during earlier deflationary episodes.

Figure 4. Time series plot of sparse principal components



To quantify the contribution of each component to recession predictions, we applied analysis to our LightGBM model, with implementation details provided SHAP (SHapley Additive exPlanations)in Appendix B. Figure 5 shows the resulting probability contributions over time, decomposed by component.

Several notable patterns emerge from this analysis. The Corporate Distress component (PC1) shows particularly strong signals during the early 2000s recession and the period following the 2011 Tohoku earthquake. The former period corresponds to the acceleration of non-performing loan disposals, exemplified by high-profile failures such as the bankruptcy of retail giant Mycal in 2001. The latter period captures the post-disaster industrial shakeout, most notably the collapse of Elpida Memory in 2012, which marked the largest manufacturing bankruptcy in Japan's postwar history.

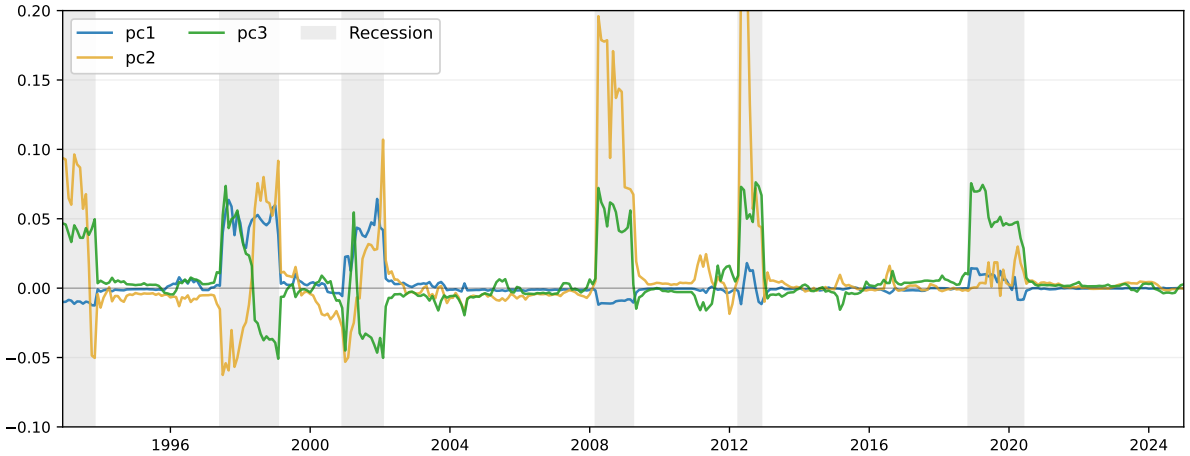
The Financial Distress component (PC2) dominates during the Global Financial Crisis of 2008-2009 and the European sovereign debt crisis period. This component captures the contagion from international financial market disruptions to the Japanese economy.

The Deflationary Pressure component (PC3) shows its strongest signal during the

Global Financial Crisis, when deflationary concerns intensified amid a severe demand shock. This component highlights Japan's persistent challenge with deflationary pressures, which has been a distinctive feature of its economic cycles.

These findings suggest that text-based indicators can provide not only predictive power but also interpretable signals about the nature of different recession episodes. The varying importance of different text components across recessions suggests that economic downturns in Japan have had distinct underlying drivers, ranging from corporate distress to financial market disruptions to deflationary spirals.

Figure 5. Component contributions during different recession periods



Note: The plot shows the probability contributions of each principal component to recession predictions, derived from SHAP (SHapley Additive exPlanations) values transformed by a logistic function.

8 Conclusion

This paper develops a unified framework for forecasting Japanese recessions that integrates machine learning methods, mixed-frequency data, and text-based indicators. While international research has advanced these approaches separately (machine learning for recession forecasting (Vrontos et al., 2021), mixed-frequency data handling (Galvão and Owyang, 2022), and text-based indicators (Pierdzioch and Gupta, 2020)), a comprehensive evaluation for Japan has been lacking. Through pseudo real-time evaluation over three decades, we assess the relative contribution of these innovations to forecasting performance.

Machine learning models outperform traditional benchmarks across forecast horizons. The model confidence set results further favor machine learning models: The benchmark logit models are not included at any horizon, while multiple machine-learning specifications remain in the confidence set. This outcome suggests that flexible modeling approaches may better capture recession dynamics in the Japanese context.

Our analysis extends insights from the U.S.-focused literature to Japan. Text-based indicators are particularly effective for short-term forecasts but their relative importance diminishes for longer horizons. Conversely, term spreads and financial variables, consistent with Estrella and Mishkin (1998) and Borio et al. (2020), show limited short-term value but become increasingly important as the forecast horizon extends. These contrasting patterns reveal strong horizon dependence in predictor importance, suggesting fundamentally different information structures underlying near-term versus long-term recession signals. Mixed-frequency data offers limited marginal benefits when comprehensive monthly indicators are available, suggesting that the information gain from higher-frequency observations depends critically on the richness of the baseline predictor set. For practical implementation, this result implies that unless real-time immediacy is the absolute priority, resources may be more effectively allocated to expanding the breadth of monthly indicators and refining machine learning models, rather than investing heavily in the infrastructure required for higher-frequency (e.g., weekly) data updates. Sparse PCA offers one approach to interpreting the text-based indicators, identifying components related to corporate failures, financial disruptions, and deflationary pressures, though the relative importance of these dimensions varies across recession episodes.

For policymakers navigating Japan's post-zero-interest-rate environment, these findings suggest that monitoring near-real-time indicators such as news sentiment, alongside traditional data and within flexible machine learning frameworks, could facilitate earlier identification of recession risks and enable more timely countercyclical responses. Future research could explore alternative text sources such as central bank communications, investigate forecast combination strategies across model types and frequencies, and extend the analysis to cross-country comparisons to assess whether our findings reflect Japan-

specific dynamics or broader patterns in advanced economies. As policymakers face the challenges of policy normalization, the methods developed in this paper may offer useful tools for recession forecasting.

References

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Bernard, H. and S. Gerlach (1998). Does the term structure predict recessions? The international evidence. *International Journal of Finance & Economics 3*(3), 195–215.

Borio, C., M. Drehmann, and F. D. Xia (2020). Forecasting recessions: the importance of the financial cycle. *Journal of Macroeconomics 66*, 103258.

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

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

Foroni, C., M. Marcellino, and C. Schumacher (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 57–82.

Galvão, A. B. and M. Owyang (2022). Forecasting low-frequency macroeconomic events with high-frequency data. *Journal of Applied Econometrics 37*(7), 1314–1333.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2005). There is a risk-return trade-off after all. *Journal of Financial Economics 76*(3), 509–548.

Goshima, K., M. Shintani, and H. Takamura (2022). Sentiment dictionary for business cycle analysis and its applications (in Japanese). *Journal of Natural Language Processing 29*, 1233–1253.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hirata, H. and K. Ueda (1998). The yield spread as a predictor of Japanese recessions. Discussion Paper Series 98-E-6, Institute for Monetary and Economic Studies, Bank of Japan.

Ito, T., H. Sakaji, K. Tsubouchi, K. Izumi, and T. Yamashita (2018). Text-visualizing neural network model: Understanding online financial textual data. In D. Phung, V. S. Tseng, G. Webb, B.-N. Ho, M. Ganji, and L. Rashidi (Eds.), *Advances in Knowledge Discovery and Data Mining*, Volume 10939 of *Lecture Notes in Computer Science*, Cham, pp. 247–259. Springer.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4765–4774. arXiv:1705.07874.

Miyazaki, T. (2016). Estimation of Japan's recession probability using the composite index (CI) :An approach using Markov-switching models (in Japanese). JCER Discussion Paper 145, Japan Center for Economic Research.

Okimoto, T. and S. Takaoka (2017). The term structure of credit spreads and business cycle in japan. *Journal of the Japanese and International Economies 45*, 27–36.

Pierdzioch, C. and R. Gupta (2020). Uncertainty and forecasts of U.S. recessions. *Studies in Nonlinear Dynamics & Econometrics 24*(4), 20190004.

Vermeulen, P. (2012). Quantifying the qualitative responses of the output purchasing managers index in the US and the Euro area. ECB Working Paper 1417, European Central Bank.

Vrontos, S. D., J. Galakis, and I. D. Vrontos (2021). Modeling and predicting U.S. recessions using machine learning techniques. *International Journal of Forecasting 37*(3), 647–671.

Watanabe, T. (2003). Measuring business cycle turning points in Japan with a dynamic Markov switching factor model. *Monetary and Economic Studies 21*(1), 35–68.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.

Appendix

A Sparse Principal Component Analysis Implementation

A.1 Methodology

Sparse principal component analysis (Sparse PCA) extends traditional PCA by incorporating sparsity constraints on the component loadings. Originally proposed by Zou et al. (2006), Sparse PCA addresses a fundamental limitation of standard PCA: While PCA achieves optimal dimension reduction in terms of variance explanation, each principal component is typically a linear combination of all original variables, making interpretation challenging in high-dimensional settings. By imposing L1 regularization, Sparse PCA produces components that load on only a subset of variables, dramatically improving interpretability while maintaining the dimension reduction benefits of PCA. This approach has proven particularly valuable in text analysis and genomics, where the number of features often exceeds the number of observations and where identifying the most relevant features is as important as dimension reduction itself.

We implement Sparse PCA to extract interpretable components from the 874-dimensional sentiment term matrix. Sparse PCA solves the following optimization problem:

$$(\mathbf{U}^*, \mathbf{V}^*) = \arg\min_{\mathbf{U},\mathbf{V}} \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \alpha\|\mathbf{V}\|_1 \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{n\times 874}$ is the centered sentiment term matrix, $\mathbf{U}$ contains the component scores, $\mathbf{V}$ contains the sparse loadings, and α is the L1 regularization parameter controlling sparsity.

A.2 Parameter Selection

Number of Components We extract three principal components (`n_components = 3`). Attempts to include a fourth component or more yielded only negligible oscillations

without interpretable economic content. The three-component solution captures the essential narrative structures while maintaining clear economic interpretation (Corporate Distress, Financial Distress, and Deflationary Pressure).

Regularization Parameter We set the regularization parameter $\alpha = 10$ after extensive experimentation. Lower values ($\alpha \leq 9$) resulted in components that loaded on too many terms—sometimes exceeding 100 terms per component—making economic interpretation difficult. At $\alpha = 10$, we achieved an optimal balance with each component loading on 5 to 8 terms, allowing for clear economic interpretation while retaining essential information. Higher values ($\alpha \geq 11$) produced excessive sparsity that led to information loss and reduced predictive performance.

A.3 Implementation

We use scikit-learn's `SparsePCA` implementation with `n_components=3` and `alpha=10`, keeping all other parameters at their default values. The resulting loadings matrix achieves 99.2% sparsity (21 non-zero entries out of 2,622), enabling the clear interpretation presented in Table 4.

When these three sparse components replace the full 874-dimensional sentiment vector in our forecasting models, we maintain comparable out-of-sample performance (AUC values of 0.95, 0.94, and 0.96 for 3-, 6-, and 12-month horizons, respectively), confirming that the dimensionality reduction preserves essential predictive information.

B SHAP Analysis for Model Interpretation

B.1 Methodology

SHAP (SHapley Additive exPlanations) is a unified framework for interpreting machine learning model predictions based on game-theoretic Shapley values. Introduced by Lundberg and Lee (2017), SHAP assigns each feature an importance value for a particular prediction, satisfying several desirable properties including local accuracy, missingness,

and consistency. The key insight is that Shapley values from cooperative game theory provide a principled way to distribute the "payout" (the model's prediction) among the "players" (the input features).

For a given prediction $f(x)$, the SHAP value for feature i is:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{4}$$

where F is the set of all features, S is a subset of features, and $f_S(x_S)$ represents the model's expected prediction when only features in S are known. This formulation captures the average marginal contribution of feature i across all possible feature combinations.

B.2 Application to LightGBM Models

For tree-based models like LightGBM, SHAP provides an efficient algorithm called TreeSHAP that exactly computes Shapley values in polynomial time rather than the exponential time required by the general definition. TreeSHAP leverages the tree structure to efficiently evaluate feature contributions along decision paths, making it computationally feasible even for ensemble models with hundreds of trees.

In our analysis, we apply SHAP to decompose recession probability predictions from our LightGBM models. For each time point t, we obtain SHAP values for all features, with particular focus on the three sparse principal components. The SHAP value $\phi_{i,t}$ represents the contribution of component i to the log-odds of recession at time t. We then transform these values through the logistic function to obtain probability contributions shown in the main text.

B.3 Implementation

We use the `shap` Python library with the `TreeExplainer` for our LightGBM models. The implementation follows these steps: (1) Train the LightGBM model on the training set; (2) Create a TreeExplainer object fitted to the trained model; (3) Calculate SHAP values

for all observations in the test set; (4) Transform log-odds contributions to probability scale for visualization.

The resulting SHAP values provide both global feature importance (averaged across all predictions) and local explanations (for specific time periods), enabling us to understand how different text components contribute to recession predictions during various economic episodes. This decomposition reveals that recession episodes in Japan are characterized by distinct combinations of text components—some driven primarily by corporate distress signals, others by financial market turmoil or deflationary pressures—providing interpretable structure beyond aggregate prediction accuracy.

C Detailed SHAP Analysis of Recession Prediction Models

This appendix uses SHAP values to explore the internal logic of our best-performing models. We focus on three key aspects: (1) the sign conditions of predictors, (2) the structural shift in variable importance before and after the introduction of QQE, and (3) the decomposition of prediction errors.

The results in Section 6 demonstrate that machine learning models significantly outperform traditional benchmarks and that predictor importance is horizon-dependent. However, the analysis does not address the sign conditions linking each predictor to recession probability, the extent to which the predictive power of term spreads has diminished under unconventional monetary policy, or the sources of model errors. This appendix addresses these aspects using SHAP analysis applied to our LightGBM models.

Unless otherwise noted, SHAP values in Sections C.1–C.2 are computed in-sample from the final estimation point (full training data through 2024) to ensure smooth, interpretable patterns. Section C.3 uses out-of-sample SHAP from the pseudo real-time forecasting scheme to analyze genuine prediction errors.

C.1 Sign Conditions and Structural Stability

To characterize the direction and magnitude of each predictor's influence on recession probability, we compute the standardized SHAP–feature relationship coefficient β, defined as:

$$\beta = \rho \times \sigma(\text{SHAP}) \tag{5}$$

where ρ is the Pearson correlation between feature values and their SHAP contributions, and $\sigma(\text{SHAP})$ is the standard deviation of the SHAP values for that feature. This coefficient measures the change in recession log-odds per one standard deviation increase in the feature, capturing both the direction (sign of ρ) and magnitude (variability of SHAP) of the predictor's influence.

Table 5 reports β for all predictors, split by the Pre-QQE period (before April 2013) and Post-QQE period (April 2013 onward), at forecast horizons of 3, 6, and 12 months. The QQE split is motivated by the Bank of Japan's adoption of large-scale asset purchases in April 2013, which fundamentally altered the yield curve dynamics that underpin term spread predictability.

Several patterns emerge from Table 5. First, sign conditions are largely consistent with economic priors. Improvements in labor market conditions (New Job Offers, $\beta = -0.81$) and rising machinery orders ($\beta = -0.31$) push away from recession, while rising inventory ratios ($\beta = +0.27$) and deteriorating confidence indicators push toward recession. Sentiment-M exhibits the largest absolute β (-2.55), consistent with its strong short-horizon predictive power documented in Section 6.

Second, and more importantly, term spreads exhibit a striking structural break across the QQE divide. The 10-year minus 1-year spread has $\beta = -0.41$ Pre-QQE but shrinks to $\beta = -0.03$ Post-QQE, representing a 93% decline in magnitude. The 3-year minus 1-year spread shows a similar pattern ($-0.26 \rightarrow -0.03$). This finding provides direct quantitative evidence that the Bank of Japan's yield curve control policy has substantially eroded the term spread's informational content for recession forecasting, consistent with the broader literature questioning the yield curve's predictive power in low-rate

Table 5. Standardized SHAP–Feature Relationship (β)

		$h = 3$M			$h = 6$M			$h = 12$M		
Group	Variable	Pre	Post	All	Pre	Post	All	Pre	Post	All
Leading Index (LI) components (variables 1–11)										
	New Job Offers	−0.99	−0.48	−0.81	−0.99	−0.47	−0.81	−1.00	−0.46	−0.81
	Consumer Confidence Index	+0.92	+0.53	+0.73	+0.93	+0.53	+0.73	+0.94	+0.51	+0.73
	Small Business Sales Forecast DI	+0.41	+0.19	+0.32	+0.41	+0.19	+0.32	+0.41	+0.19	+0.32
	Machinery Orders	−0.28	−0.36	−0.31	−0.28	−0.36	−0.31	−0.28	−0.36	−0.31
	Inventory Ratio (Producer Goods)	+0.30	+0.22	+0.27	+0.30	+0.22	+0.27	+0.30	+0.22	+0.27
	Nikkei Commodity Index	−0.27	−0.15	−0.24	−0.27	−0.15	−0.24	−0.27	−0.16	−0.24
	Housing Floor Area	+0.18	+0.06	+0.15	+0.18	+0.06	+0.15	+0.18	+0.07	+0.15
	Investment Environment Index	−0.05	−0.48	−0.17	−0.05	−0.47	−0.17	−0.05	−0.46	−0.17
	Inventory Ratio (Final Demand)	+0.17	+0.08	+0.13	+0.17	+0.08	+0.13	+0.12	+0.14	+0.13
	Money Stock (M2)	−0.06	−0.05	−0.05	−0.06	−0.05	−0.05	−0.05	−0.05	−0.05
	TOPIX	−0.03	+0.11	+0.01	−0.03	+0.11	+0.01	−0.04	+0.11	+0.01
Term Spreads (variables 12–14)										
	Term Spread (10Y−1Y)	−0.41	−0.03	−0.24	−0.43	−0.04	−0.24	−0.42	+0.00	−0.24
	Term Spread (5Y−1Y)	−0.21	−0.18	−0.13	−0.22	−0.18	−0.13	−0.22	−0.17	−0.13
	Term Spread (3Y−1Y)	−0.26	−0.03	−0.19	−0.26	−0.03	−0.19	−0.26	−0.03	−0.19
Financial (variables 15–16)										
	Debt Service Ratio (DSR)	−0.16	−0.17	−0.16	−0.16	−0.17	−0.16	−0.15	−0.22	−0.16
	Realized Volatility	−0.09	−0.20	−0.13	−0.09	−0.20	−0.13	−0.08	−0.20	−0.13
Text-based (variable 17)										
	Macroeconomic Sentiment	−2.07	−2.47	−2.55	−2.09	−2.52	−2.55	−2.14	−2.60	−2.55
Other										
	Recession (lagged)	+0.52	+0.54	+0.53	+0.51	+0.55	+0.53	+0.52	+0.55	+0.53

Note: $\beta = \rho \times \sigma(\text{SHAP})$, where ρ is the Pearson correlation between feature values and SHAP values, and $\sigma(\text{SHAP})$ is the standard deviation of SHAP contributions. Units are log-odds per one standard deviation of the feature. $\beta > 0$: a one SD increase pushes toward recession; $\beta < 0$: pushes away from recession. "Pre" denotes the period before April 2013 (QQE); "Post" denotes April 2013 onward. Variable numbers refer to Table 1, which also documents transformations applied to each variable.

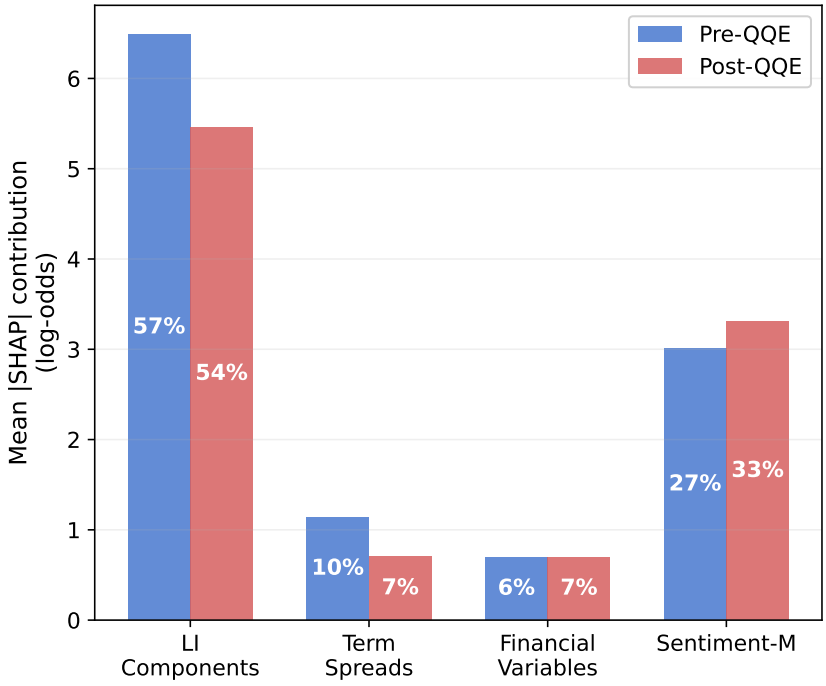
environments (Borio et al., 2020).

C.2 Structural Shift in Variable Group Importance

The preceding analysis of individual β coefficients reveals sign conditions and structural breaks at the variable level. We now aggregate to the group level to quantify how the relative importance of different predictor categories has shifted across policy regimes.

Figure 6 compares the mean absolute SHAP contribution (in log-odds) of each variable group between the Pre-QQE and Post-QQE periods. For each group, we sum the absolute SHAP values across all constituent variables and average over observations within each period. The percentages inside each bar indicate the group's share of the total model reliance.

Figure 6. SHAP-based variable group importance: Pre- vs. Post-QQE



Note: Each bar represents the mean absolute SHAP contribution (log-odds) for a variable group. Percentages indicate each group's share of total importance. "Leading Index" comprises 11 components; "Term Spreads" comprises 3 yield curve measures; "Financial" includes DSR and realized volatility. Pre-QQE: before April 2013; Post-QQE: April 2013 onward.

The figure reveals two notable shifts. First, the LI's share of total importance declines from approximately 57% Pre-QQE to 54% Post-QQE. While the decline is moderate in

percentage terms, it suggests that the LI's lead time over the CI has diminished in recent decades, as the composition of Japan's business cycles has shifted. Second, and more notably, the Text-based group's share rises from 27% Pre-QQE to 33% Post-QQE, a relative increase of roughly 20%. This shift is particularly notable given that the Text-based group comprises a single indicator (Sentiment-M), whereas the LI group comprises 11 variables. These results support the view that text-based indicators have become increasingly important for recession forecasting in an environment where traditional macroeconomic signals have weakened.

Conversely, term spreads show the sharpest decline in absolute importance (from 1.14 to 0.71 in mean |SHAP|), reinforcing the conclusion from the β analysis that yield curve information has been substantially diminished under QQE.

C.3 Error Analysis

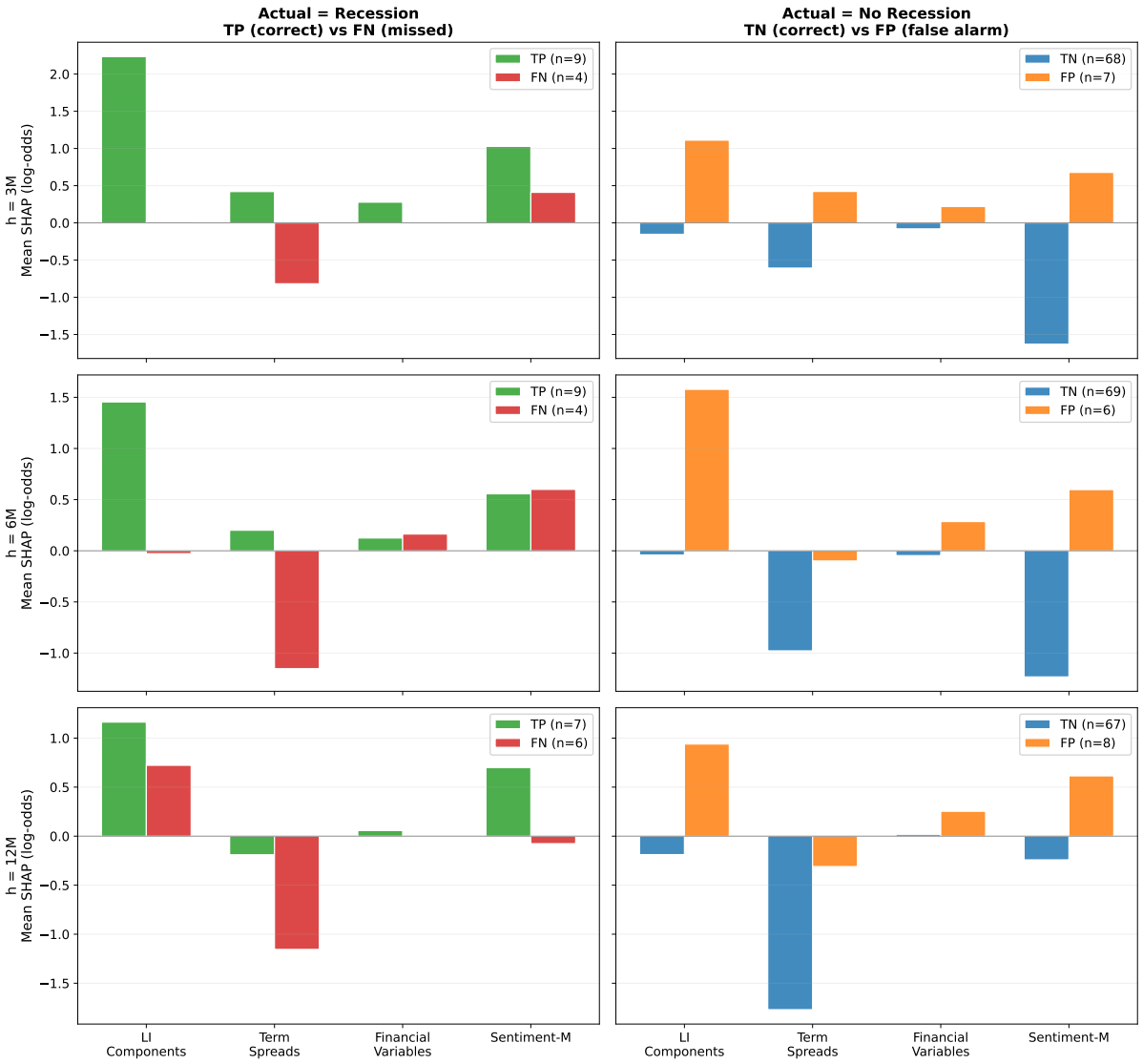
This section examines the model's prediction errors to identify the sources of false negatives (missed recessions) and false positives (false alarms). In contrast to our strategy in the proceeding sections, we now use out-of-sample SHAP values from the pseudo real-time forecasting scheme, ensuring that each observation's SHAP decomposition reflects a genuine ex-ante prediction.

We classify each out-of-sample prediction into one of four categories—True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).[6] Figure 7 compares the mean SHAP contribution of each variable group across these categories.

The decomposition reveals a pattern that is most pronounced at shorter horizons. False negatives (missed recessions) are primarily driven by the LI failing to signal recession: at $h = 3\,\mathrm{M}$, the LI's mean SHAP drops from $+2.2$ in TP cases to near zero ($+0.01$) in FN cases. Notably, term spreads actively mislead the model during FN episodes, contributing large negative SHAP values (-0.82 at $h = 3\,\mathrm{M}$) despite an actual recession being underway—by contrast, in TP cases term spreads contribute positively ($+0.42$),

[6]The classification threshold is the OOS unconditional recession frequency ($13/88 \approx 0.15$), i.e. a quarter is classified as "predicted recession" when the model deems it more likely than the historical base rate.

Figure 7. SHAP error decomposition by variable group



Note: Each bar represents the mean SHAP contribution (log-odds) for a variable group, computed over out-of-sample observations classified as TP, FN, FP, or TN. The classification threshold is the OOS unconditional recession frequency ($13/88 \approx 0.15$). Left panels show actual recession periods (TP vs. FN); right panels show non-recession periods (TN vs. FP). Rows correspond to forecast horizons $h = 3, 6, 12$ months.

suggesting that correct detections coincide with episodes where the yield curve still inverts in the traditional manner. Indeed, the total SHAP for FN cases is negative (-0.40), indicating that the model actively pushed these predictions *away* from recession. This result is consistent with the structural break documented in Section C.1: in the Post-QQE environment, prolonged unconventional monetary easing has compressed the yield curve, preventing the traditional pre-recession inversion.

At shorter horizons, text-based indicators partially compensate for these failures. In FN cases, Sentiment-M contributes a positive SHAP of $+0.41$ at $h = 3\,\mathrm{M}$ and $+0.60$ at $h = 6\,\mathrm{M}$, correctly pointing toward recession; at $h = 12\,\mathrm{M}$, however, this compensating effect vanishes (-0.08). Even where present, the text signal is insufficient to overcome the LI's silence and term spreads' misdirection. This finding highlights a potential avenue for improving model performance: assigning greater weight to text signals when traditional macroeconomic and financial indicators send conflicting or muted signals.

For false positives (false alarms), the pattern is reversed. At $h = 3\,\mathrm{M}$, the LI contributes $+1.1$ in SHAP (pushing toward recession despite no actual downturn), while text contributes $+0.68$. These false alarms appear to be driven by temporary deteriorations in leading indicators that resolve without culminating in a recession.

D Model-Specific MCS Results

To complement the cross-model MCS analysis in Section 6, we apply the MCS procedure separately within individual model classes. In practice, a forecaster may have already selected a preferred algorithm based on institutional constraints, computational resources, or interpretability requirements. For such a user, the relevant question is not which algorithm to choose, but which input configuration performs best within the chosen model class. We therefore conduct within-model MCS for the two best-performing algorithms—LightGBM and KNN—each comprising 32 input configurations. Tables 6 and 7 report the results.

Within both model classes, the Sentiment-M indicator maintains its distinct horizon-

Table 6. Models included in the MCS (LightGBM)

Horizon	Total	Term	Financial	Text		Mixed
	(out of 32)	Spread	Variables	Any	Sentiment-M	Frequency
3-Month	21	43%	57%	81%	48%	43%
6-Month	29	55%	55%	80%	28%	48%
12-Month	31	52%	52%	78%	26%	48%

Note: The MCS procedure is applied at the 5% significance level using log loss as the loss function. "Any" denotes the share of remaining models that include at least one text indicator (Sentiment-M, Sentiment-F, or EPU).

Table 7. Models included in the MCS (KNN)

Horizon	Total	Term	Financial	Text		Mixed
	(out of 32)	Spread	Variables	Any	Sentiment-M	Frequency
3-Month	15	60%	47%	93%	53%	47%
6-Month	11	64%	64%	81%	36%	45%
12-Month	31	52%	52%	78%	26%	48%

Note: See Table 6 for details on the MCS procedure.

dependent pattern, with high inclusion rates at shorter horizons that diminish as the horizon extends, confirming its robustness as a near-term signal. However, the monotonic increase in importance for term spreads and financial variables observed in the cross-model MCS (Table 3) is less evident in these model-specific results. This discrepancy is likely attributable to the reduced discriminatory power of the MCS at longer horizons: nearly all configurations remain at the 12-month horizon (31 out of 32 for both algorithms). This high inclusion rate reflects the inherent difficulty of long-term forecasting, which makes it statistically challenging to distinguish between specifications and consequently obscures the contributions of long-leading indicators.