Recession Prediction

Gordon Oboh & Joanne Chen

Department of Data Science, Monroe College, King Graduate School

CS655-151HY: Machine Learning

Professor Eugene Adjei Djan

July 30, 2025

Recession Prediction

Introduction

Context and Background

A recession is generally defined as a significant decline in economic activity that lasts for a few months. This decline is typically visible in real GDP, real income, employment, industrial production, and wholesale-retail sales (National Bureau of Economic Research, n.d.) Translating this qualitative definition into a quantitative forecasting problem has led researchers to examine the predictive content of financial indicators. The study by Bauer and Mertens (2018) highlights the predictive power of the term spread (10-year minus 1-year Treasury rates), asserting that inverted yield curves have preceded all U.S. recessions over the past 60 years, with only one false positive. Their empirical analysis confirms that a negative spread reliably signals elevated recession probability, even after accounting for potential structural changes in the interest rate environment. Similarly, Aramonte and Xia (2019) reinforce the historical reliability of the 10-year minus 3-month Treasury yield spread, noting that every U.S. recession since 1973 was preceded by an inversion of this spread. Nonetheless, both studies converge on a key point: despite evolving macro-financial conditions, the inversion of the yield curve remains one of the most consistent and early indicators of impending recessions (Aramonte & Xia, 2019; Bauer & Mertens, 2018)

Problem Statement

Despite the extensive body of research linking the yield curve to recession forecasting, several critical gaps remain. First, there is a limited focus on alternative yield spreads; the majority of studies concentrate on the 10-year minus 3-month or 1year Treasury spread, overlooking other combinations such as GS10–DGS2 or GS10–DGS3MO that may offer more stable predictive signals. Second, the use of high-frequency data such as daily or weekly averages is relatively rare, as most prior work relies on monthly or quarterly averages, potentially missing high-frequency dynamics crucial for near real-time forecasting. Third, there is an insufficient integration of machine learning with macroeconomic datasets. Fourth, many studies neglect the issue of class imbalance inherent in recession data. Recession periods are rare relative to expansions, yet performance metrics are often reported without adjustment for this imbalance, potentially leading to misleading conclusions. Finally, there is a lack of comparative evaluation across modeling techniques; few studies evaluate both classical econometric models and modern machine learning approaches within the same empirical framework. This leaves open the question of which approaches are most robust across different data conditions and policy regimes

Objectives

This study is grounded in the theoretical relationship between the yield curve and the business cycle. According to the Expectations Hypothesis of the term structure of interest rates, long-term interest rates reflect the average of expected future short-term interest rates. When investors anticipate economic slowdowns, they expect future short-term rates to fall. This shift in expectations can cause long-term yields to fall below short-term rates, resulting in an inverted yield curve. Historically, such inversions have preceded U.S. recessions, making the yield curve a widely recognized leading indicator of economic downturns (Estrella & Mishkin, 1998). This study aims to use several modeling approaches; Logistic Regression a traditional econometric model. In addition, tree-based ensemble models such as Balanced Random Forest and Easy Ensemble Classifier are employed to account for non-linearities and imbalanced class distributions inherent in recession data. Finally, the study incorporates various architectures of Long Short-Term Memory (LSTM) networks.

Literature Review

Relevant Studies

According to National Bureau of Economic Research (n.d.) A recession is generally defined as a significant decline in economic activity that lasts for an few months. This decline is typically visible in real GDP, real income, employment, industrial production, and wholesale-retail sales.

Translating this qualitative definition into a quantitative forecasting problem has led researchers to examine the predictive content of financial indicators. The study by Bauer and Mertens (2018) highlights the predictive power of the term spread (10-year minus 1-year Treasury rates), they assert that inverted yield curves have preceded all U.S. recessions over the past 60 years, with only one false positive. Their empirical analysis confirms that a negative spread reliably signals elevated recession probability, even after accounting for potential structural changes in the interest rate environment. Similarly, Aramonte and Xia (2019) reinforce the historical reliability of the 10-year minus 3-month Treasury yield spread, noting that every U.S. recession since 1973 was preceded by an inversion of this spread. Nonetheless, both studies converge on a key point: despite evolving macro-financial conditions, the inversion of the yield curve remains one of the most consistent and early indicators of impending recessions (Aramonte & Xia, 2019; Bauer & Mertens, 2018).

Gaps in Research

Despite the extensive body of research linking the yield curve to recession forecasting, several critical gaps remain. First, there is a limited focus on alternative yield spreads; the majority of studies concentrate on the 10-year minus 3-month or 1-year Treasury spread, overlooking other combinations such as GS10–DGS2 or GS10–DGS3MO that may offer more stable predictive signals. Second, the use of high-frequency data such as daily or weekly averages is relatively rare, as most prior work relies on monthly or quarterly averages, potentially missing high-frequency dynamics crucial for near real-time forecasting. Third, there is an insufficient integration of machine learning with macroeconomic datasets. Fourth, many studies neglect the issue of class imbalance inherent in recession data. Recession periods are rare relative to expansions, yet performance metrics are often reported without adjustment for this imbalance, potentially leading to misleading conclusions. Finally, there is a lack of comparative evaluation across modeling techniques; few studies evaluate both classical econometric models and modern machine learning approaches within the same empirical framework. This leaves open the question of which approaches are most robust across different data conditions and policy regimes.

Theoretical Framework

This study is grounded in the theoretical relationship between the yield curve and the business cycle. According to the *Expectations Hypothesis* of the term structure of interest rates, long-term interest rates reflect the average of expected future short-term interest rates. When investors anticipate economic slowdowns, they expect future short-term rates to fall. This shift in expectations can cause long-term yields to fall below short-term rates, resulting in an inverted yield curve. Historically, such inversions have preceded U.S. recessions, making the yield curve a widely recognized leading indicator of economic downturns (Estrella & Mishkin, 1998).

This study aims to use several modeling approaches; Logistic Regression a traditional econometric model. In addition, tree-based ensemble models such as Balanced Random Forest and Easy Ensemble Classifier are employed to account for non-linearities and imbalanced class distributions inherent in recession data. Finally, the study incorporates various architectures of Long Short-Term Memory (LSTM) networks.

Methodology

Approach

This study adopts a quantitative and experimental approach to investigate the predictive relationship between yield curve dynamics and U.S. recessionary periods. The methodology is grounded in empirical modeling and computational experimentation, aiming to quantify the extent to which various term spreads and machine learning models can accurately forecast recessions.

The quantitative aspect is reflected in the use of structured time-series data and statistical indicators derived from Treasury yields, processed to generate explanatory features. The experimental component involves the systematic development, training, and evaluation of multiple predictive models under controlled conditions. These include traditional econometric models and modern machine learning classifiers, all tested on consistent datasets with standardized performance metrics. This dual approach ensures both the analytical rigor of hypothesis-driven modeling and the adaptability of data-driven experimentation.

Data Collection

This study utilizes secondary quantitative data obtained from publicly available economic datasets. The primary source of financial indicators is the Federal Reserve Bank of St. Louis (FRED), from which three daily time series were extracted: the 10-Year Treasury Constant Maturity Rate (GS10), the 2-Year Treasury Constant Maturity Rate (DGS2), and the 3-Month Treasury Bill Rate (DGS3MO). These yield series were selected due to their widespread use in macro-financial research and their theoretical linkage to expectations about future economic activity.

To identify recessionary periods for labeling, the study uses official dates provided by the National Bureau of Economic Research (n.d.).

Tools and Techniques

Computational analysis for this research was conducted using Python3 within a Jupyter Notebook environment.

Data Analysis

Initial preprocessing of the time-series data involved using forward-fill to handle missing values. This technique was chosen to maintain temporal continuity without introducing artificial volatility or structural breaks. Subsequently, two features were engineered from the processed data: the 10-year minus 2-year Treasury spread (GS10–DGS2) and the 10-year minus 3-month Treasury spread (GS10–DGS3MO). A binary target variable was constructed based on the NBER-designated recession dates, labeling each daily observation as either within (1) or outside (0) a recession period. Following the creation of these three features, the dataset was further resampled into three distinct time frequencies: daily, weekly, and monthly. This would allow for the evaluation of model performance under varying levels of temporal aggregation and data granularity. The final feature engineered dataset spans 01-Jan-1982 to 18-Jul-2025and includes both recession and non-recession periods, facilitating robust training and evaluation of predictive models under realistic, imbalanced class distributions.

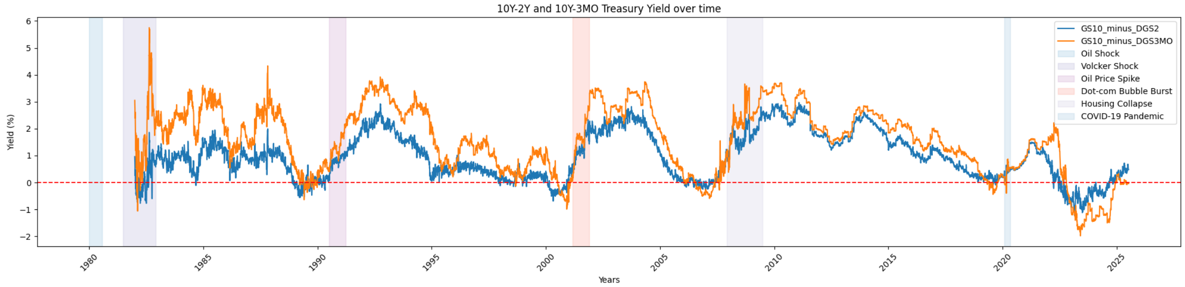

Figure 1

10Y-2Y and 10Y-3MO U.S. Treasury Yield Spreads Over Time.

Project Implementation

System Architecture

The overarching architecture of this project was systematically structured into four distinct phases: data preparation, model testing and validation, analysis of model results, and the implementation of forecasting methodologies. This sequential framework ensured a logical progression from raw data to actionable insights regarding recession indicators. Initially, the process focused on the meticulous acquisition, cleaning, and transformation of raw datasets, specifically those identified as crucial recession indicators through prior research and literature reviews. Subsequently, a selection of machine learning models, chosen for their theoretical applicability to time-series analysis and predictive analytics, underwent rigorous testing and validation against historical data. The performance metrics derived from these models were then meticulously recorded and analyzed to discern their efficacy and reliability. The final phase involved leveraging the insights from the best-performing models to generate future forecasts for economic recession, providing a forward-looking perspective based on empirical evidence.

Development Process

The development process for this study commenced with the critical steps of initial data cleaning and subsequent importation, establishing a robust foundation for the analytical procedures. This foundational stage ensured data quality and consistency, which are paramount for reliable model performance. The second, and most extensive, phase involved a multi-faceted analytical approach. This included monitoring key economic indicators such as the 10-year-minus-2-year Treasury yield spread and the 10-year-minus-3-month Treasury yield spread, with careful consideration given to varying average day intervals to capture different temporal dynamics. The Treasury yield data was systematically split into training, validation, and testing sets to ensure unbiased model evaluation. Furthermore, the process entailed generating detailed comparisons between actual historical data and the curves predicted by linear regression models, providing a visual and statistical assessment of model fit. Finally, various forecasting models were applied, and their respective results were thoroughly analyzed to identify patterns and predictive capabilities. To rigorously evaluate the performance of the diverse models employed, a comprehensive set of reference metrics was utilized. These included precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC$_{\text{test}}$ score). Each of these metrics provided a distinct perspective on model effectiveness, particularly in identifying and predicting recessionary periods. Precision assessed the accuracy of positive predictions, recall measured the model's ability to identify all actual positive cases, the F1-score offered a balanced measure of both, and the AUC-ROC$_{\text{test}}$ score provided an aggregate measure of performance across all possible classification thresholds. Ultimately, the model demonstrating the most superior performance across these critical evaluation metrics was selected and implemented for forecasting future recessionary periods within this project, serving as the primary predictive tool.

Technologies Used

All experimental procedures and analytical tasks within this project were conducted in an offline, notebook-based environment, chosen for its flexibility, reproducibility, and iterative development capabilities. The entire workflow, encompassing structured preprocessing, meticulous feature engineering, comprehensive model training, and rigorous evaluation steps, was implemented as modular components. The core programming language utilized for all computational tasks was Python, leveraging its extensive libraries for data manipulation, statistical analysis, and machine learning. Jupyter Notebook served as the primary interactive development environment, facilitating code execution, visualization, and documentation in an integrated format. For version control and collaborative development, GitHub was employed, ensuring efficient tracking of changes and seamless teamwork. Finally, the MS Office Suite was instrumental for generating comprehensive reports, creating professional presentations, and managing project documentation, thereby supporting the communication and dissemination of the project's findings.

LSTM Model Architecture

The Long Short-Term Memory (LSTM) models employed in this study are configured with 1...*n* layers, where the number of layers and their respective hidden units are determined by the model name. For example, LSTM_4 denotes a single-layer LSTM with 4 units, while LSTM_4_2 denotes a two-layer LSTM with 4 units in the first layer and 2 units in the second layer.

Each model terminates with a dense output layer using a sigmoid activation function to map the final hidden state to a binary classification. Dropout regularization with a rate of 0.3 is applied after each LSTM layer to mitigate overfitting. Training is performed using the Adam optimizer and a binary focal loss function (equation (**??**)), parameterized by $\gamma = 2.0$, α(equation (**??**)) and class weight calculations (equation (**??**)) to address class imbalance. Model performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The following LSTM variants were implemented:

- LSTM_4, LSTM_8: Single-layer LSTM models with 4 or 8 hidden units.
- LSTM_4_4, LSTM_8_4, LSTM_8_8: Two-layer LSTM models with corresponding hidden unit configurations.

Results/Findings

Key Results

We present the key empirical findings of our study, beginning with class distribution insights, followed by model performance based on AUC-ROC$_{\text{test}}$ (AUCs are calculated on the test set), and finally an analysis of the predicted recession probabilities over time.

Inherent Class Imbalance

A significant class imbalance is evident across all time frequencies, as shown in Figure 2. Recession periods are underrepresented in comparison to non-recession periods (approximately 19% recession) Such skewed distributions necessitate the use of class-balancing techniques (SMOTE, random undersampling, class weights and weighting factor (α)) to mitigate bias in model training.

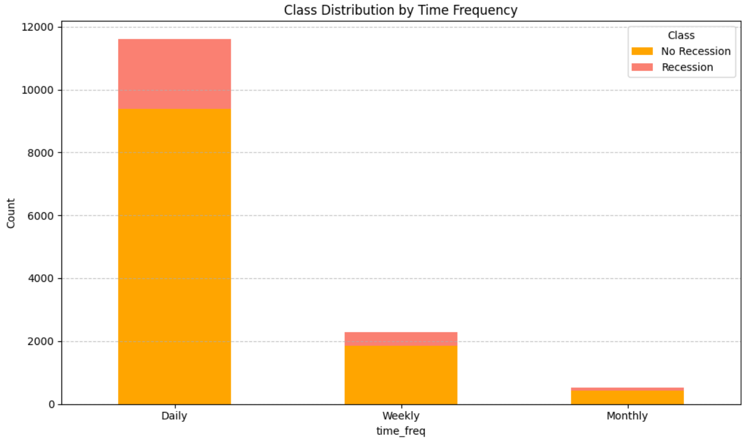

Figure 2
Stacked Bar Chart Showing Class Imbalance.

AUC-ROC$_{test}$ Performance and Model Selection

To evaluate model performance, we used AUC-ROC$_{\text{test}}$ scores across each time frequency(daily, weekly and monthly). Traditional models with AUC-ROC$_{\text{test}}$ scores below 0.5 (indicating worse-than-random performance) were excluded from further analysis, while all LSTM models were retained regardless of performance for comparison purposes. As shown in Table 1, For traditional models, Logistic Regression and Easy Ensemble Classifier consistently achieved the highest AUCs across all datasets. Logistic Regression performed the best with Easy

Ensemble performing slightly lower but still comparable. Other Traditional models, including XGBoost, Balanced Random Forest, and traditional Random Forest, failed to reach the 0.5 AUC-ROC$_{test}$ threshold in all scenarios.

Table 1

AUC-ROC$_{test}$ Scores by Model and Time Frequency (traditional models with AUC-ROC$_{test}$ < 0.5 excluded; all LSTM models retained).

Time Frequency	Model	AUC-ROC$_{test}$
Daily	Logistic Regression	0.7222
	Easy Ensemble Classifier	0.6918
	`LSTM_4`	0.7202
	`LSTM_4_4`	0.7330
	`LSTM_8`	0.2752
	`LSTM_8_4`	0.7291
	`LSTM_8_8`	0.7021
Weekly	Logistic Regression	0.7251
	Easy Ensemble Classifier	0.6853
	`LSTM_4`	0.7178
	`LSTM_4_4`	0.7263
	`LSTM_8`	0.7604
	`LSTM_8_4`	0.4078
	`LSTM_8_8`	0.6019
Monthly	Logistic Regression	0.7263
	Easy Ensemble Classifier	0.6675
	`LSTM_4`	0.5335
	`LSTM_4_4`	0.3880
	`LSTM_8`	0.3730
	`LSTM_8_4`	0.5952
	`LSTM_8_8`	0.4559

Recession Probability Forecasts

Figure 3 displays the predicted probabilities of recession from the Logistic Regression and Easy Ensemble models across daily, weekly, and monthly datasets. Both Logistic Regression and Easy Ensemble models produce elevated probability estimates preceding known recession periods. Logistic Regression forecasts appear smoother and more gradual, the peaks appear less than 52 weeks to a recession, while Easy Ensemble produces sharper peaks about 52 weeks before a recession, indicating a greater sensitivity to abrupt changes in macroeconomic conditions.

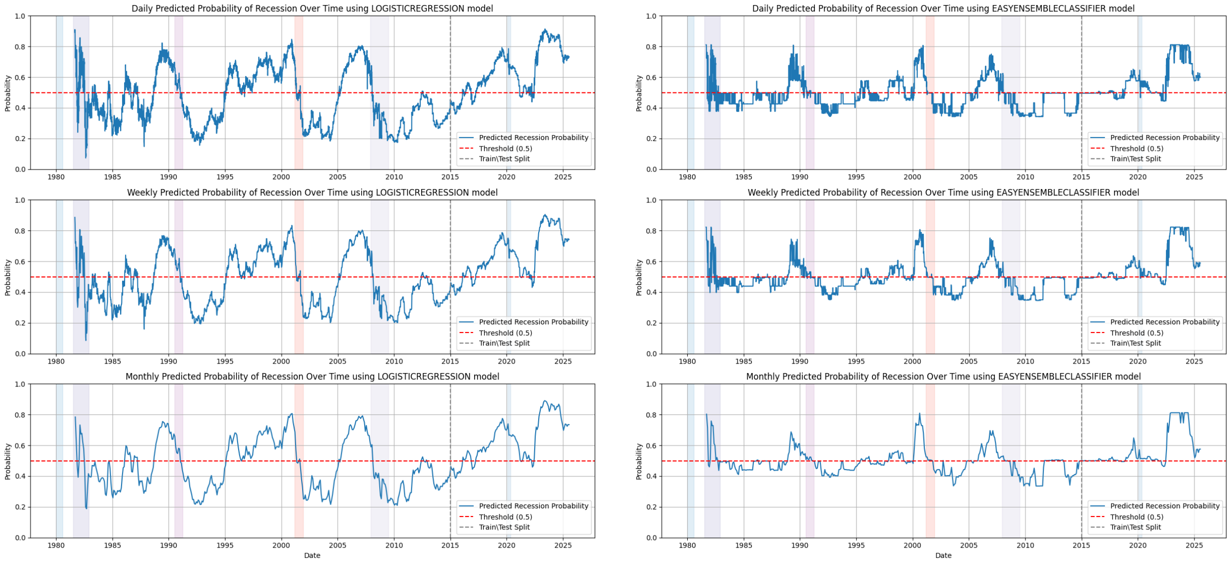

Figure 3

Predicted probability of recession using Logistic Regression (left) and Easy Ensemble (right).

To assess the forecasting capabilities of simpler deep learning architectures, we examined single-layer LSTM models in Figure 4 (`LSTM_4` and `LSTM_8`) across each time frequencies. At the daily forecasts, `LSTM_4` achieved moderate success (AUC-ROC$_{\text{test}}$ = 0.7202), offering balanced sensitivity to changing conditions, while `LSTM_8` performed poorly (AUC-ROC$_{\text{test}}$ = 0.2752), displaying extreme volatility and a high rate of false positives. For the weekly level, `LSTM_8` demonstrated the strongest performance (AUC-ROC$_{\text{test}}$ = 0.7604), producing smooth and timely signals ahead of recessions. `LSTM_4` (AUC-ROC$_{\text{test}}$ = 0.7178) showed sharper but more volatile responses, with better precision around recession windows, although it also exhibited random probability spikes between 1995–2000 and 2010–2015, similar to `LSTM_8` , but more pronounced. On monthly data, `LSTM_4` yielded stable yet uninformative predictions (AUC-ROC$_{\text{test}}$ = 0.5335), with probabilities remaining close to the 0.5 threshold, In contrast, `LSTM_8` (AUC-ROC$_{\text{test}}$ = 0.3730) tended to overreact to past recessions while downplaying periods of economic expansion. Single-layer LSTMs showed varying degrees of effectiveness across time resolutions, with `LSTM_4` offering more consistent performance and `LSTM_8` excelling only in the weekly setting but struggling at finer and coarser granularities.

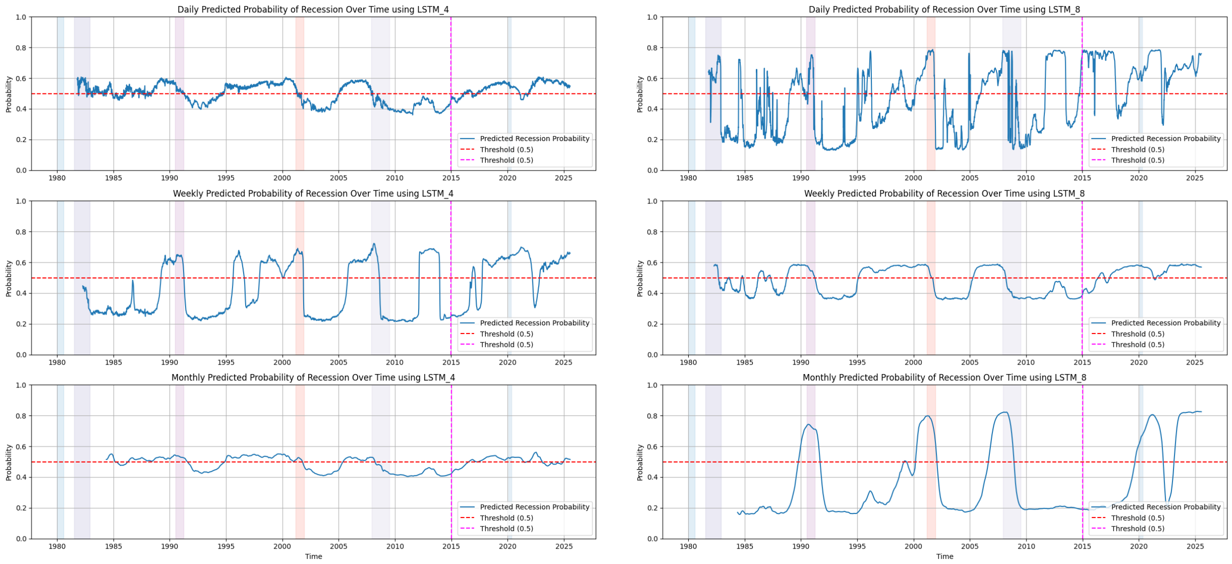

Figure 4

Predicted probability of recession using `LSTM_4` *(left) and* `LSTM_8` *(right).*

The double-layer LSTM models provided a broader view of how increased network depth influences recession probability forecasts as shown in Figure 5. On daily data, `LSTM_4_4` generated relatively stable probabilities with modest peaks around known recession periods, avoiding excessive noise and aligning well with economic downturns. Its predictions tracked closely with those of the single-layer `LSTM_4` but with slightly improved temporal precision. `LSTM_8_4` also performed steadily on daily data but showed occasional overprediction with sustained high probabilities outside recession windows. `LSTM_8_8` , however, exhibited increased volatility compared to the other double-layer models, with its prominent peaks occurring just before recession periods. It frequently spiked above the 0.5 probability threshold even during expansions. In the weekly setting, all models tended to produce signals with extended periods of high probabilities and frequent sharp transitions, traits indicative of a high false positive rate.

While the predicted probability plots indicate that models performed best on the monthly dataset and worst on the weekly dataset, the AUC-ROC$_{\text{test}}$ values suggest otherwise.

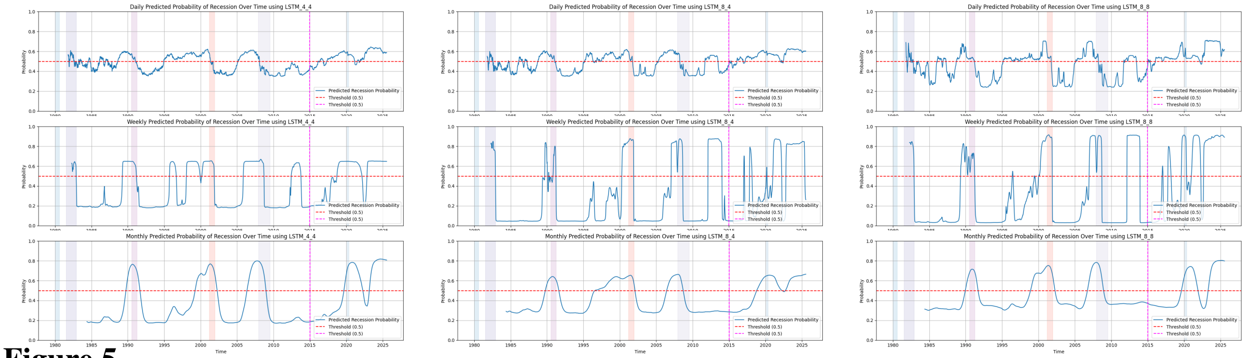

Figure 5

Predicted probability of recession using `LSTM_4_4` (left), `LSTM_8_4` (center) and `LSTM_8_8` (right).

Data Interpretation

Traditional Model Behavior and Interpretability.

The strong and consistent performance of Logistic Regression and the Easy Ensemble Classifier across daily, weekly, and monthly datasets highlights their robustness in separating recession signals from background noise. Logistic Regression, in particular, achieved higher AUC-ROC$_{\text{test}}$ values, demonstrating that even relatively simple, interpretable models can perform competitively in complex economic forecasting tasks. Analysis of the predicted probability trajectories further supports these findings: Logistic Regression tended to produce smoother, more gradual increases in predicted recession risk leading up to historically verified downturns, making it well-suited for early warning systems where trend clarity and stability are essential. In contrast, the Easy Ensemble Classifier exhibited sharper, more volatile probability spikes. This heightened sensitivity may offer advantages in rapidly evolving macroeconomic contexts, though it could also increase the risk of false positives.

LSTM Model Behavior and Interpretability.

The performance of LSTM models revealed both potential and pitfalls when applying deep learning to recession forecasting. Single-layer LSTMs such as `LSTM_4` and `LSTM_8` showed reasonable performance. With `LSTM_8` achieving the highest AUC among all models at the weekly level. These models were able to capture early signs of recessions with some degree of lead time, validating their ability to model temporal dependencies in sequential economic data.

However, they also displayed varying levels of volatility suggesting possible sensitivity to noise or overfitting. Double-layer LSTM models exhibited increased complexity, with `LSTM_4_4` and `LSTM_8_8` striking a favourable balance between signal detection and noise detection across all time frequencies. Although they all perfomred terribly on the weekly dataset (unstable and exaggerated signals, including elevated probabilities during expansionary periods).

Unexpected Outcomes

One of the key unexpected outcomes of this study was the underperformance of the LSTM models relative to the baseline Logistic Regression model. Given the LSTMs' theoretical advantage in capturing temporal dependencies in sequential data, we expected them to match or exceed the performance of traditional classifiers. However, across most time frequencies, Logistic Regression consistently outperformed LSTM variants in terms of AUC-ROC$_{\text{test}}$. This result challenges the assumption that model complexity necessarily yields superior predictive accuracy in economic time series. Additionally, our findings revealed that AUC-ROC$_{\text{test}}$ may not be a fully appropriate metric for evaluating LSTM models in this context. While some LSTM configurations produced informative probability trajectories that aligned well with recession timing, their AUC scores remained relatively low. This discrepancy highlights the limitations of using AUC-ROC$_{\text{test}}$ as the sole performance criterion for time-series classification, particularly when sequence shape and trend stability are critical for interpretability and practical use.

Discussion

Implications

The findings of this study offer valuable insights into the application of machine learning techniques for recession forecasting using yield curve data. The consistent performance of Logistic Regression and the Easy Ensemble Classifier across all time frequencies demonstrates that simple, interpretable models can effectively extract recession signals from macro-financial time series when paired with thoughtful preprocessing and class-balancing strategies. Furthermore, the ability of LSTM models to produce early warning signals highlights the potential of deep learning architectures to model sequential dependencies in economic indicators. These results support the growing role of data-driven models in economic forecasting and suggest that with proper tuning, both traditional and neural approaches can complement one another in real-time monitoring systems.

Comparison with Existing Work

Our findings align with long-established literature indicating that the yield curve, particularly the spread between the 10-year Treasury bond and the 3-month bill, contains strong predictive power for future U.S. recessions. The foundational work by Estrella and Mishkin (1998) and its ongoing application by the Federal Reserve Bank of St. Louisemphasize that a negative term spread is one of the most robust leading indicators of recessions, supported by decades of historical data.

Our use of Logistic Regression is in direct continuity with this empirical tradition, and its top performance in our study affirms that even under modern machine learning workflows, the logistic model remains well-suited to macroeconomic binary classification tasks. However, our work extends prior literature by incorporating LSTM-based time-series models and comparing them to ensemble classifiers under class imbalance constraints. Unlike the static probit/logit approaches used in classical studies, our sequential models attempt to forecast dynamic risk trajectories. While LSTM models showed potential (in probability curve and temporal alignment) their performance did not surpass simpler models on AUC-based metrics, possibly due to data sparsity or overfitting. This comparison reveals that while ML architectures provide modeling flexibility, their advantage is not guaranteed without rigorous regularization and domain-specific adaptation.

Limitations

Several limitations constrain the generalizability and operationalization of this study's results. First, the dataset is inherently imbalanced, with recession periods representing only a small fraction of the total observations. Although class-balancing techniques such as SMOTE, undersampling, and custom weighting schemes were applied, rare event modeling remains fundamentally difficult and can impact sensitivity. Second, while AUC-ROC was used as the primary model selection criterion, it may not fully reflect the forecasting utility of LSTM models. AUC evaluates rank-order classification performance but ignores the temporal structure and trajectory of predicted probabilities. Future work could incorporate metrics such as Precision-Recall AUC (PR-AUC), calibration plots, or time-based scoring mechanisms to better capture sequence quality. Finally, the study explored a limited set of LSTM configurations. A more exhaustive hyperparameter search; including deeper architectures, attention mechanisms, and bidirectional models could provide better insight into the upper bounds of deep learning performance in this domain.

Conclusion

Summary of Contributions

The findings of this study offer valuable insights into the application of machine learning techniques for recession forecasting using yield curve data. The consistent performance of Logistic Regression and the Easy Ensemble Classifier across all time frequencies demonstrates that simple, interpretable models can effectively extract recession signals from macro-financial time series when paired with thoughtful preprocessing and class-balancing strategies. Furthermore, the ability of LSTM models to produce early warning signals highlights the potential of deep learning architectures to model sequential dependencies in economic indicators. These results support the growing role of data-driven models in economic forecasting and suggest that with proper tuning, both traditional and neural approaches can complement one another in real-time monitoring systems. Our use of Logistic Regression is in direct continuity with this empirical tradition, and its top performance in our study affirms that even under modern machine learning workflows, the logistic model remains well-suited to macroeconomic binary classification tasks. However, our work extends prior literature by incorporating LSTM-based time-series models and comparing them to ensemble classifiers under class imbalance constraints. Unlike the static probit/logit approaches used in classical studies, our sequential models attempt to forecast dynamic risk trajectories. While LSTM models showed potential (in probability curve and temporal alignment) their performance did not surpass simpler models on AUC based metrics, possibly due to data sparsity or overfitting. This comparison reveals that while ML architectures provide modeling flexibility, their advantage is not guaranteed without rigorous regularization and domain-specific adaptation.

Future Work

One of the key unexpected outcomes of this study was the underperformance of the LSTM models relative to the baseline Logistic Regression model. Given the LSTMs' theoretical advantage in capturing temporal dependencies in sequential data, we expected them to match or exceed the performance of traditional classifiers. However, across most time frequencies, Logistic

Regression consistently outperformed LSTM variants in terms of AUC-ROC$_{test}$. This result challenges the assumption that model complexity necessarily yields superior predictive accuracy in economic time series. Additionally, our findings revealed that AUC-ROC$_{test}$ may not be a fully appropriate metric for evaluating LSTM models in this context. While some LSTM configurations produced informative probability trajectories that aligned well with recession timing, their AUC scores remained relatively low. This discrepancy highlights the limitations of using AUC-ROC$_{test}$ as the sole performance criterion for time-series classification, particularly when sequence shape and trend stability are critical for interpretability and practical use.

Takeaways

Several limitations constrain the generalizability and operationalization of this study's results. First, the dataset is inherently imbalanced, with recession periods representing only a small fraction of the total observations. Although class-balancing techniques such as SMOTE, undersampling, and custom weighting schemes were applied, rare event modeling remains fundamentally difficult and can impact sensitivity. Second, while AUC-ROC$_{test}$ was used as the primary model selection criterion, it may not fully reflect the forecasting utility of LSTM models. AUC evaluates rank-order classification performance but ignores the temporal structure and trajectory of predicted probabilities. Future work could incorporate metrics such as Precision-Recall AUC (PR-AUC), calibration plots, or time-based scoring mechanisms to better capture sequence quality. Finally, the study explored a limited set of LSTM configurations. A more exhaustive hyperparameter search; including deeper architectures, attention mechanisms, and bidirectional models could provide better insight into the upper bounds of deep learning performance in this domain.

References

Aramonte, S., & Xia, D. (2019). Yield curve inversion and recession risk. *BIS Quarterly Review*. https://www.bis.org/publ/qtrpdf/r_qt1909.htm

Bauer, M. D., & Mertens, T. M. (2018). Economic forecasts with the yield curve. *FRBSF Economic Letter*, (2018-07). https://www.frbsf.org/economic-research/publications/economic-letter/2018/march/economic-forecasts-with-yield-curve/

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

National Bureau of Economic Research. (n.d.). US Business Cycle Expansions and Contractions [Accessed: 2025-07-07].