International Journal of Computer Applications Technology and Research
Volume 14–Issue 12, 21 – 27, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1003

# Multi-Horizon US Recession Prediction Using Diverse Machine Learning Models

1st Narendra Lakshmana Gowda
Independent Researcher
Ashburn, VA, USA

2nd Vihar Manchala
Independent Researcher
Aldie, VA, USA

3rd Abdul Raheem Mohammed
Independent Researcher
McKinney, TX, USA

***Abstract*** This study compares several popular Machine Learning (ML) models for predicting US recessions. The models tested include gradient boosting, random forest, support vector classifier, feedforward neural network, and a custom consensus model. Predictions are made for 3, 6, 9, 12, and 18-month periods. The dataset spans from 1962 to 2023 and utilizes six leading indicators for training and testing the models. Two evaluation techniques are used to evaluate the models' performance. The first method employs a ranking system based on accuracy, precision, recall, F1, and AUROC scores. The second method uses bootstrapped AUROC scores to conduct a one-sided test on each model pair across all prediction horizons. The results indicate that gradient boosting, random forest, and consensus models outperform the logistic regression model in predicting recessions. These models show better classification capabilities across all time horizons. On the other hand, the neural network and support vector classifier models do not demonstrate as strong performance compared to the logistic regression model. Similarly, the findings from the consensus model highlight the advantage of combining predictions from multiple models. This approach results in more accurate recession forecasts than relying on the outputs of a single model.

***Index Terms*** Consensus Model, Feedforward Neural Network, Gradient Boosting, Machine Learning, Random Forest, Support Vector Classifier, US Recessions

## I. INTRODUCTION

When examining economic data it's easy to overlook the real people behind the numbers. Each data point represents the hopes and dreams of individuals and families. For instance, [1] found that losing a job increases the risk of mortality by 63%. Despite this connecting events like an inverted yield curve to a family losing their home can be challenging. However, these events are interlinked within the broader economy, and the implications of better forecasting affect real people in a significant way. This makes it imperative to find the best possible methods for modeling recessions. Policymakers and business leaders rely heavily on the work of researchers who study recession predictions and forecasting more generally. Their decisions are influenced by these forecasts which is why it is crucial for researchers to critically and objectively evaluate different recession forecasting techniques. The accuracy of these models can lead to better decision-making which can have profound effects on the economy and people's lives. This underscores the importance of research into recession forecasting. If predicting recessions were easy the need for extensive research would diminish. However, recessions often share common characteristics but are usually triggered by exogenous shocks which are difficult or impossible to foresee. For instance, the 2001 recession in the United States was caused by the bursting of the tech bubble and the tragic events of 9/11. On the other hand, the Great Recession of 2007-2009 was largely triggered by a collapse in the U.S. housing market. Despite these vastly different causes, various leading indicators can still give reliable signals of future economic conditions. This suggests that while each recession is unique there is enough consistency in various leading indicators to make predicting future economic conditions somewhat feasible. Nevertheless, no single indicator can predict recessions perfectly.

Historically, the methodologies used to identify relationships between economic indicators and recessions have remained relatively unchanged. Before the widespread adoption of Machine Learning (ML) models [2]–[7] logistic and probit regression models were the primary tools for recession prediction. However, the exponential increase in econometric data and the decreasing cost of computing are challenging this status quo in favor of more complex ML models. There is growing evidence suggesting that ML models may be superior to traditional logistic [8, 9] and probit models in forecasting recessions. Nonetheless, it is premature to assert a definitive consensus that ML models are universally better at recession prediction than traditional econometric techniques. One reason for this is the differing scoring methodologies used to evaluate these models. Several common methods are used to compare binary classification models such as the ones tested in this study. Popular scoring methodologies include metrics like the F1-score and significance tests using Receiver Operating Characteristic (ROC) curves. While there are some commonalities in these scoring methodologies across different research efforts there is no standardized approach. Consequently, the results can vary widely, partly because of these inconsistencies. Therefore, this study aims to contribute to the research on recession prediction by conducting a comprehensive comparison of several ML models against a traditional logistic regression model. Specifically, it tests gradient boosting, neural networks, support vector classifiers, random forests, and a consensus model against logistic

regression. The classification abilities of these models are evaluated over 3, 6, 9, 12, and 18-month time horizons. The models are then assessed using two different scoring methodologies. The first methodology introduces a novel ranking approach based on various metrics collected across different horizons. The second methodology employs bootstrapped Area under the Receiver Operating Characteristic (AUROC) scores to calculate *p*-values and perform one-sided tests between models. The approach proposed in this study aims to address a critical gap in existing research which often focuses disproportionately on specific forecasting horizons and uses inconsistent evaluation techniques. As a result, many studies lack a comprehensive assessment of the overall performance capabilities of ML models relative to logistic regression models in predicting recessions. By employing the novel score ranking and bootstrapped AUROC methodologies this study provides clear indications of general performance using both common scoring metrics and significance tests.

The study is as follows; related papers are shown in the following section. The materials and methods are provided in Section III. The experimental analysis is carried out in Section IV, and in Section V, we provide some conclusions and plans for future research.

## II. Related Works

Predicting economic cycles has long been a focus for policymakers and econometric researchers. [10] analyzed economic time series data to identify patterns in economic activity leading to the modern concept of a reference cycle. This idea evolved into what we now call the business cycle, describing the ongoing fluctuations in Gross Domestic Product (GDP). [10] work identified relationships between various indicators and the peaks and troughs of business cycles laying the foundation for future research on recession prediction. [11] advanced this research by using regression techniques to compare different economic indicators. They found that multivariate models which consider multiple indicators are more effective than single or bivariate models for predicting economic trends. Despite these advancements, predicting recessions remains challenging due to the scarcity of reliable leading indicators. [11, 12] study tested 45 econometric variables but found most were ineffective for predicting the 1990-1991 US recession. Even those that provided valuable insights offered inconsistent predictions for the 1970s and 1980s recessions. This inconsistency highlights the difficulty of finding reliable indicators for recession prediction in a constantly evolving economy. Recessions often share characteristics but have different causes complicating prediction efforts. For instance, the 2020 recession was triggered by the COVID-19 pandemic, a clear cause-and-effect scenario, while the Great Recession of 2008-2009 was driven by a complex interplay of factors, primarily the housing market collapse.

The definition of a recession is also debated. The common definition is two consecutive quarters of negative GDP growth, popularized by [13]. However, some argue this definition overemphasizes GDP and neglects other important economic factors. Today, the National Bureau of Economic Research (NBER) defines a recession as "a significant decline in economic activity spread across the economy, lasting more than a few months". [13] has identified certain financial variables as reliable indicators of US recessions with the term spread being particularly notable. The term spread is the difference between long-term and short-term government bond yields. Historically, a low or negative term spread often precedes economic downturns, while a high spread suggests economic expansion. Throughout the 1980s, further research validated the term spread's predictive power. [14, 15] found that the term spread was a reliable recession indicator but not effective for predicting precise changes in GDP. [14], followed by [16] demonstrated that financial and non-financial metrics could predict US recessions up to eight quarters in advance. Other financial indicators, such as stock prices and credit conditions have also proven useful in recession forecasting. For instance, changes in consumer spending which accounts for about 70% of US GDP have been linked to future economic conditions. Simple auto-regressive models have been effective for short-term GDP forecasting highlighting the value of past GDP changes as predictive information such as [17]. Consumer sentiment is another reliable non-financial indicator explaining significant variations in Gross National Product (GNP). [18] indicates that the yield curve changes in financial markets, and consumer sentiment are among the most consistent indicators for predicting recessions.

## III. Materials and Methods

### A. Data Analysis

As noted in the literature review there are two main challenges when collecting training and testing data to predict recessions. First, recessions are relatively rare in the US which makes it difficult for forecasters to gather enough data. Second, ML models usually need large datasets [19]–[28] but it's hard to find long econometric data series. Some metrics have data going back to the 1800s while others barely go back to the 1990s. To address the varying availability of data series, the focus was on creating a dataset that goes as far back in time as possible. This means prioritizing showing indicators with long data series. While this might neglect some potentially significant features that don't have long histories the goal is to create a dataset suitable for comparing models. With more data the relative performance of the models can be assessed more robustly. The data for this analysis comes mainly from two public sources. The first is the Federal Reserve Economic Data[1] (FRED) API which provides many economic time series mostly for the US but also for other countries. The second source is the Yahoo Finance API[2] which gives easy access to financial market data. The target variable is the NBER recession classification series, a monthly series from FRED that indicates

[1] https://fred.stlouisfed.org/

[2] https://developer.yahoo.com/api/

whether the US is in a recession. The first leading indicator is the quarterly log change in the S&P 500 which tracks the performance of the largest 500 US companies. This index is a good proxy for the overall performance of US financial markets and has long been used as a leading indicator of future economic conditions. Other leading indicators include the yield curve and a lagged version of it. The yield curve here refers to the spread between the 10-year treasury yield and the 3-month treasury bill rate, with the lagged version being the spread three months prior. The six-month log change in real GDP is also included. This data, originally from FRED was transformed using cubic spline interpolation to fit a monthly time frame. Annual changes in consumer spending another leading indicator is significant given the US's consumption-driven economy. The consumer spending data, also from FRED, was adjusted for inflation using the Consumer Price Index (CPI) before calculating the annual log change. Consumer sentiment is another key indicator. The sentiment data comes from the University of Michigan Consumer Sentiment series via FRED. Before 1978, this data was collected quarterly, so cubic spline interpolation was used to align it with the other indicators. To make the sentiment data more comparable, min-max normalization was applied, as it was not normally distributed. The final dataset spans from April 1962 to October 2023 with monthly values for six leading indicators and the NBER recession classifier. This dataset, with 739 rows is larger than those in many similar studies but still small for ML purposes. It includes 85 months classified as recessions, about 11.5% of the data. Similarly, multicollinearity, the correlation between independent variables, is an important consideration as shown in Fig. 1. High multicollinearity can lead to inaccurate coefficient estimates particularly in regression models, but it can also affect ML models. In this dataset, the yield curve and its lagged version, as well as the annual change in consumer spending and the scaled consumer sentiment, show a high correlation. The six-month change in real GDP and the annual change in consumer spending also correlate 0.5. Variance Inflation Factor (VIF) scores confirm moderate to high multicollinearity among some features. The primary goal of this analysis is to compare the models rather than correct for multicollinearity or perform feature selection. While multicollinearity can affect coefficient estimation it doesn't necessarily impact the models' predictive ability. However, it's important to consider multicollinearity when interpreting feature importance in the results.

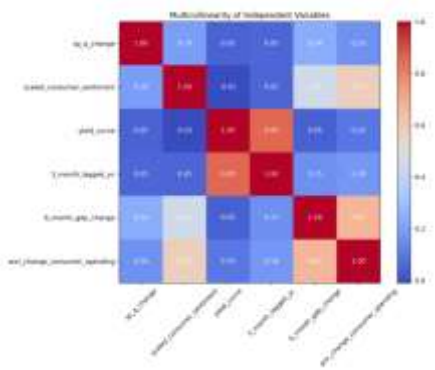

Fig. 1. Matrix of correlation for leading indicators

### *B. Model Analysis*

The traditional econometric model employed in this analysis is the logistic regression classifier. Logistic regression has a long history of use and was developed in 1944 as an alternative to the normal probability (probit) function. The logistic model maps the relationship between inputs and outputs using a sigmoid function which compresses predictions to values between 0 and 1. This transformation allows the output to be interpreted as the probability of a positive classification. Parameters for the logistic function are estimated using MLE. This involves taking the natural logarithm of the likelihood function and finding the parameters that maximize this log-likelihood function using optimization techniques like gradient descent. Once the optimization process converges the estimated parameters are used for predictions. Whereas, Artificial Neural Networks (ANNs) are inspired by the way the human brain processes information. These models have been in and out of vogue over the past 70 years but are currently experiencing resurgence. ANNs consist of three main parts the input layer, hidden layers, and output layer. Input layer this layer processes input data through various nodes and passes it to subsequent hidden layers. Hidden layers neural networks can have multiple hidden layers each applying an activation function to transform the data. In this analysis, a feedforward neural network with two hidden layers is utilized. The hidden layers use the Rectified Linear Unit (ReLU) activation function which introduces non-linearity and helps the model learn complex patterns. The ReLU function works by outputting the input value if it is greater than 0 and 0 otherwise. The final layer which often produces a single output, uses an activation function to generate the prediction. In this case, a sigmoid activation function is used to produce a classification probability between 0 and 1. The feedforward aspect means that data flows only in one direction through the network unlike feedback neural networks which can loop data back and forth. XGBoost, short for eXtreme Gradient Boosting, is a powerful and efficient implementation of the gradient-boosted trees algorithm. This ensemble learning model builds multiple decision tree models sequentially each correcting the errors of the previous one to reduce bias and improve accuracy. The optimization process in XGBoost involves the learning rate and step size. The learning rate controls how much the model's parameters are adjusted in the direction of the gradient at each iteration while the step size is the actual magnitude of these updates. Although the learning rate remains constant the effective step size decreases over iterations as the gradient is minimized leading to a more refined model. Random forest models are another type of ensemble method that combines the predictions of multiple decision trees. This model uses a technique called bootstrap aggregation, or bagging which involves training several trees on different subsets of the training data and features. Each tree makes predictions, and the final prediction is determined by a majority vote. This approach helps reduce overfitting and enhances the robustness of the predictions. The random forest model starts by randomly selecting subsets of the training data through bootstrapping. Each decision tree is trained on these subsets. At each node of the decision tree the algorithm selects the best split among a

International Journal of Computer Applications Technology and Research
Volume 14–Issue 12, 16 – 20, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1003

random subset of features. After training, the trees combine their predictions to vote for the final prediction. In classification tasks, the class label with the majority vote is selected. Support Vector Machines (SVM) are supervised models used for classification and regression tasks. The primary objective of an SVM model is to find the optimal hyperplane that divides the input space into regions representing different classes. The support vectors are the data points closest to the hyperplane and are crucial in defining it. When the data is not linearly separable SVM models use the kernel trick to transform the data into a higher dimension where a linear hyperplane can be established. The choice of kernel function a hyperparameter is specified manually. By using the kernel trick, SVM models can handle complex classification problems efficiently. The final model evaluated in this analysis is a custom consensus model. This model takes the probability of a positive classification from the random forest, SVM classifier, XGBoost, logistic regression, and neural network models. By averaging all the predictions the consensus model produces a final classification probability. This approach aims to test whether combining predictions from multiple models yields better results.

## IV. Experimental Analysis

The analysis is conducted in two main stages ensuring a thorough and systematic evaluation of the models. In the first stage, the models are tested at various forecast horizons. For each forecast horizon, a new model is created, and its hyperparameters are optimized through a tuning process. The data is split into training and testing sets after adjusting it to reflect the specific horizon. Once the optimized model is run on the testing data several metrics are collected and recorded. This consistent methodology ensures that the results are comparable across different models. In the context of this analysis the forecasting horizon is defined as the future time interval during which the model is trained and tested to predict recessions. Each model undergoes testing at five distinct monthly intervals 3, 6, 9, 12, and 18 months. For every iteration corresponding to a different forecasting horizon the target variable which is recession classification is adjusted to align with that horizon. The data is then split into 80% training data and 20% testing data. It is essential to note that during this splitting process; the data is shuffled, meaning its temporal order is not maintained in either the training or testing datasets. To address the class imbalance issue mentioned earlier the Synthetic Minority Oversampling Technique (SMOTE) is applied to the training data. SMOTE is a widely used methodology for generating synthetic data in highly imbalanced datasets to mitigate class imbalance. The SMOTE algorithm utilizes the *k*-nearest neighbors algorithm to create new samples of the minority class. In this analysis, the minority class consists of instances of recession. The transformed training data enhanced by SMOTE is then used for model training.

GridSearchCV is employed to identify the optimal combination of hyperparameters for each ML model. Hyperparameters are configuration settings explicitly defined by the model's creator and cannot be estimated directly from the training or testing data. These settings such as tree depth for tree-based models and the regularization technique for logistic regression models are determined before training and influence the model's behavior. The hyperparameters for each tested model vary to some extent. The second part of this process involves cross-validation, specifically *k*-fold cross-validation. K-fold cross-validation is a technique used to assess the performance and generalization ability of a model. The training data is divided into multiple subsets, known as folds which are of approximately equal size and randomly generated. The model is then trained on *k*-1 folds while the remaining fold serves as the validation set. The model's performance is assessed based on a specified metric which, in this analysis, is the F1 score. GridSearchCV combines hyperparameter tuning and cross-validation to identify the best-performing model. An object of hyperparameters, the grid, is passed in, and every unique combination of hyperparameters is tested using cross-validation. Once this process is completed, the best-performing model is selected and applied to the testing data.

### A. SHAP Analysis

The Shapley Additive exPlanations (SHAP) analysis results align well with previous studies on key features in recession prediction models. It's important to note that SHAP plots were not created for the consensus model since it combines predictions from other models. Feature importance was similar across models at each forecast horizon but there were some notable differences. For the 3 and 9-month forecast horizons as shown in Figs. 2 to 5 models heavily weighted the annual change in consumer spending and the six-month change in GDP. By the 9-month horizon, most models identified either the 3-month lagged term spread or the term spread itself as the most critical feature. This is consistent with existing literature suggesting that the term spread is a reliable predictor of US recessions.

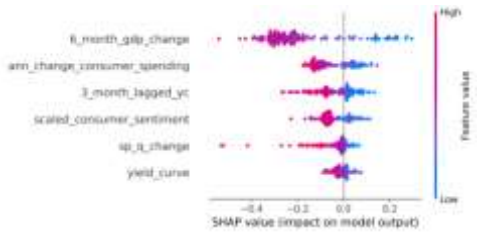


Fig. 2. Three-month horizon random forest classifier beeswarm plot

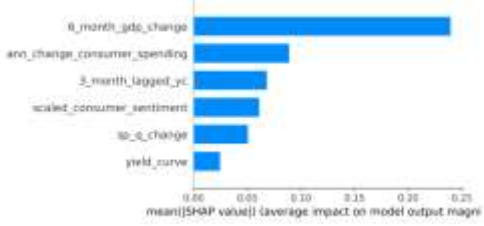


Fig. 3. Plot of the 3-month horizon summary for random forest classifier

International Journal of Computer Applications Technology and Research
Volume 14–Issue 12, 16 – 20, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1003

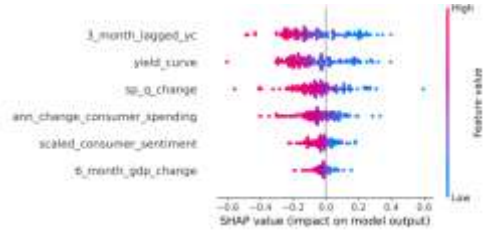


Fig. 4. A 9-month horizon logistic regression beeswarm plot

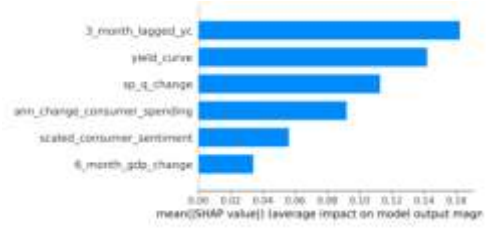


Fig. 5. Plot of the 9-month horizon summary of logistic regression

At the 12- and 18-month horizons as shown in Figs. 6 to 9, the importance of the term spread and its 3-month lagged version increases further aligning with previous research highlighting their significance. Interestingly, some models also started using low values of specific features to predict the absence of a future recession. For instance, the XGBoost model identified low six-month GDP changes as an arrow of recession. However, the 6-month GDP change was not as important as other features in the summary plots. Similar patterns were observed for the quarterly change in the S&P 500 and the annual change in consumer spending especially in models like the neural network at later horizons. It's worth noting that the neural network performed poorly at the 18-month horizon, suggesting that models might not always accurately identify economic troughs. This could be due to the average length of US recessions being around 11 months meaning that recessions could end by the 12-and 18-month forecasts. From the SHAP plots, some patterns emerge at shorter horizons, the 6-month GDP change and annual consumer spending change are crucial. By the 9-month horizon, the term spread and its lagged version become the most important. The quarterly change in the S&P 500 and consumer sentiment varied widely in importance across models and horizons. A notable observation is the dense distribution of feature importance at the 9-month horizon, indicating models have difficulty identifying a dominant indicator at this time frame. This could be due to the choice of leading indicators rather than the models themselves.

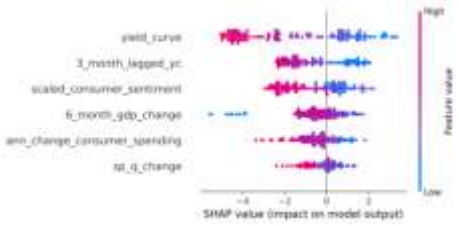


Fig. 6. XGBoost beeswarm plot for a duration of 12 months

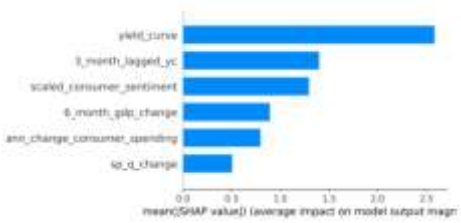


Fig. 7. XGBoost 12-month horizon summary plot

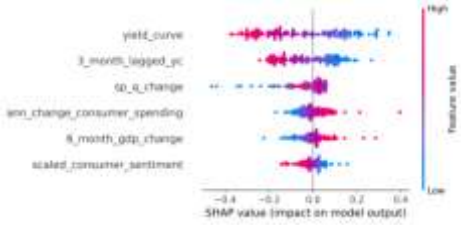


Fig. 8. Neural network swarm plot for a duration of 18 months

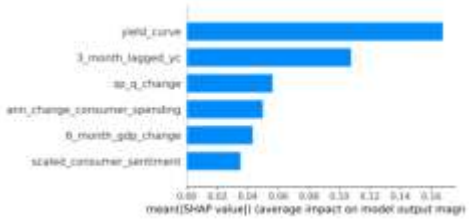


Fig. 9. Summary plot of neural networks for an 18-month period

### *B. Model Analysis*

The XGBoost model showed a fairly good performance across various time horizons as shown in Table I. For all time frames the F1 score was consistently above 0.5 indicating that the model performed better than random guessing and successfully captured underlying patterns in the data. The recall scores in particular were impressive as they revealed that at the 3- and 6-month horizons the XGBoost model could identify all recession periods. This ability to detect recessions was crucial in validating the model's effectiveness. However, a detailed look at the confusion matrix results showed that while XGBoost was proficient at spotting recessions, it also generated a considerable number of false positives. This tendency became more pronounced as the prediction horizon extended further into the future leading to a mix of false positives and false negatives. Although the XGBoost model performed reasonably well overall, its accuracy and reliability diminished over longer time horizons. In contrast, the random forest model faced challenges in identifying all recession periods at any given horizon as shown in Table I. Despite this shortcoming, the model's higher precision scores indicated that it produced fewer false positives overall compared to the XGBoost model. This higher precision suggested that the random forest model was more selective and cautious in its predictions reducing the number of false alarms. The F1 and AUROC scores further supported this showing that the model was generally adept at identifying recessions, especially at shorter horizons. While the random forest produced fewer false positives than the XGBoost

model, it unfortunately generated more false negatives, which could be detrimental when missing critical recession signals. The SVM emerged as one of the weakest models in the comparison as shown in Table I. It consistently failed to identify all recession periods at any horizon with none of its F1 scores surpassing 0.8. While it demonstrated some level of accuracy at the 3-month horizon, its performance declined sharply over longer periods. This decline was reflected in the high number of false positives and false negatives it produced, making it less reliable compared to the XGBoost and random forest models across all time horizons. The inability of the SVM to maintain consistent performance across different time frames highlighted its limitations in dealing with recession prediction tasks. The neural network model exhibited strong performance at the 3 and 6-month horizons but showed a significant drop in accuracy at longer horizons as shown in Table I. At the 6-month horizon, the neural network could identify all recession periods, but it also generated several false positives. From the 9-month horizon onwards, the F1 score decreased almost linearly, barely exceeding 0.5 at the 18-month horizon. This steady decline indicated that the neural network struggled to maintain its predictive power over extended periods. The neural network's propensity to produce numerous false negatives and positives further undermined its reliability. Although it was accurate at the 3 and 6-month horizons the stark drop in performance at later horizons made it less effective than both the XGBoost and random forest models. The consensus model, which combined predictions from multiple models, emerged as the top performer as shown in Table I. It achieved the highest F1 score at the 3-month horizon and maintained AUROC scores at or above 0.9 for all horizons. This model's ability to identify all recession periods at the 6-month horizon and sustain good precision across later horizons underscored its robustness. The consensus model produced false positives and negatives at various horizons, but the total number of errors was lower than those of the other models. This reduction in absolute errors highlighted the strength of combining multiple models to enhance predictive accuracy. Despite its strengths, the consensus model still faced challenges at later horizons, performing best at the nine-month horizon or earlier. The logistic regression model also showed commendable performance, achieving an F1 score above 0.8 at more than one horizon as shown in Table I. However, its performance dropped significantly after the 3 and 6-month horizons. The false positive rate increased dramatically past the six-month horizon, indicating that logistic regression was more prone to making incorrect predictions as the time frame extended. This characteristic made logistic regression a good choice for early predictions but less reliable beyond the six-month mark.

TABLE I
DIFFERENT MODELS ON VARIOUS HORIZONS

| Horizon | Model | Accuracy | Precision | Recall | F1 Score | AUROC |
|---|---|---|---|---|---|---|
| 3 | XGBoost | 0.959459 | 0.714286 | 1.000000 | 0.833333 | 0.988471 |
| 6 | XGBoost | 0.925170 | 0.633333 | 1.000000 | 0.775510 | 0.972039 |
| 9 | XGBoost | 0.890411 | 0.518519 | 0.823529 | 0.636364 | 0.906977 |
| 12 | XGBoost | 0.904110 | 0.666667 | 0.727273 | 0.695652 | 0.888196 |
| 18 | XGBoost | 0.875862 | 0.600000 | 0.750000 | 0.666667 | 0.914945 |
| 3 | Random Forest | 0.966216 | 0.857143 | 0.800000 | 0.827586 | 0.985213 |
| 6 | Random Forest | 0.938776 | 0.727273 | 0.842105 | 0.780488 | 0.967516 |
| 9 | Random Forest | 0.897260 | 0.545455 | 0.705882 | 0.615385 | 0.921341 |
| 12 | Random Forest | 0.917808 | 0.727273 | 0.727273 | 0.727273 | 0.924670 |
| 18 | Random Forest | 0.862069 | 0.566667 | 0.708333 | 0.629630 | 0.907025 |
| 3 | SVC | 0.959459 | 0.846154 | 0.733333 | 0.785714 | 0.980451 |
| 6 | SVC | 0.884354 | 0.541667 | 0.684211 | 0.604651 | 0.893503 |
| 9 | SVC | 0.924658 | 0.650000 | 0.764706 | 0.702703 | 0.898313 |
| 12 | SVC | 0.897260 | 0.620690 | 0.818182 | 0.705882 | 0.885447 |
| 18 | SVC | 0.855172 | 0.551724 | 0.666667 | 0.603774 | 0.852961 |
| 3 | Neural Network | 0.966216 | 0.812500 | 0.866667 | 0.838710 | 0.976942 |
| 6 | Neural Network | 0.918367 | 0.612903 | 1.000000 | 0.760000 | 0.972451 |
| 9 | Neural Network | 0.931507 | 0.705882 | 0.705882 | 0.705882 | 0.908345 |
| 12 | Neural Network | 0.869863 | 0.555556 | 0.681818 | 0.612245 | 0.852273 |
| 18 | Neural Network | 0.765517 | 0.400000 | 0.833333 | 0.540541 | 0.825069 |
| 3 | Consensus | 0.972973 | 0.823529 | 0.933333 | 0.875000 | 0.989474 |
| 6 | Consensus | 0.938776 | 0.678571 | 1.000000 | 0.808511 | 0.977385 |
| 9 | Consensus | 0.931507 | 0.733333 | 0.647059 | 0.687500 | 0.934793 |
| 12 | Consensus | 0.897260 | 0.620690 | 0.818182 | 0.705882 | 0.904692 |
| 18 | Consensus | 0.882759 | 0.620690 | 0.750000 | 0.679245 | 0.914256 |
| 3 | Logistic | 0.959459 | 0.764706 | 0.866667 | 0.812500 | 0.973935 |
| 6 | Logistic | 0.959184 | 0.882353 | 0.789474 | 0.833333 | 0.979441 |
| 9 | Logistic | 0.815068 | 0.380952 | 0.941176 | 0.542373 | 0.883265 |
| 12 | Logistic | 0.780822 | 0.395833 | 0.863636 | 0.542857 | 0.828446 |
| 18 | Logistic | 0.848276 | 0.527778 | 0.791667 | 0.633333 | 0.857438 |

## V. Conclusion and Future Works

Predicting recessions is crucial for policymakers and business leaders because accurate forecasts can significantly impact people's lives. Job losses during recessions can severely affect individuals' well-being, making it essential to improve recession predictions. Better forecasts enable smarter decisions, reducing economic losses. In 2000, only 25% of global data was digital but by 2007, this had risen to 94%. With the advent of

big data and ML it's important to explore whether these new technologies can enhance recession forecasting. This study aims to determine if ML models can predict recessions more accurately than traditional logistic regression models. The study used a dataset of leading economic indicators from 1962 to 2023, primarily sourced from FRED and Yahoo Finance APIs. The dataset had 739 rows, which is small for training ML models. The study tested various models, including logistic regression, neural networks, support vector classifiers, random forests, XGBoost, and consensus models, over forecasting periods of three to 18 months. Accuracy, precision, recall, AUROC, and F1 scores were used to evaluate the models. Two scoring methods were employed. The first ranked the models based on metrics, while the second compared bootstrapped AUROC scores. Results showed that consensus, random forest, and XGBoost models generally outperformed logistic regression. The study also found that combining multiple model predictions can yield better results than relying on a single model. This research highlights the potential of ML in recession forecasting and the benefits of combining model outputs for superior predictions.

## REFERENCES

[1] X. Chen, X. Ruan, and W. Zhang, "Dynamic portfolio choice and information trading with recursive utility," *Economic Modelling*, vol. 98, pp. 154–167, May 2021, doi: 10.1016/j.econmod.2021.02.020.

[2] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.

[3] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.

[4] Gowda, N. L. (2025, March 3). Federated Learning a Collaborative Machine Learning Across Countries with Data Privacy. https://ijritcc.org/index.php/ijritcc/article/view/11481

[5] Soni, A., Alla, S., Dodda, S., & Volikatla, H. (2024). Advancing Household Robotics: Deep Interactive Reinforcement Learning for Efficient Training and Enhanced Performance. *arXiv preprint arXiv:2405.18687*.

[6] Alla, S., & Pazos, P. (2019). Healthcare robotics: Key factors that impact robot adoption in healthcare.

[7] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1

[8] R. Rabemananjara and J. M. Zakoian, "Threshold arch models and asymmetries in volatility," *Journal of Applied Econometrics*, vol. 8, no. 1, pp. 31–49, Jan. 1993, doi: 10.1002/jae.3950080104.

[9] Alla, S., Bheesetty, N., Prakash, Y., Sunkara, M., Chidipudi, P., Chattu, V. K., & Velur, V. R. (2023). MACHINE-LEARNING ANALYSIS OF MORTALITY DUE TO COMORBIDITIES AND RESULTING MICROVASCULAR COMPLICATIONS IN COVID PATIENTS WITH TYPE-2 DIABETES MELLITUS. In *Proceedings of the International Annual Conference of the American Society for Engineering Management.* (pp. 1-9). American Society for Engineering Management (ASEM).

[10] W. F. Cascio, "Downsizing: What do we know? What have we learned?," *Academy of Management Perspectives*, vol. 7, no. 1, pp. 95–104, Feb. 1993, doi: 10.5465/ame.1993.9409142062.

[11] J. Fleming, "The quality of market volatility forecasts implied by S&P 100 index option prices," *Journal of Empirical Finance*, vol. 5, no. 4, pp. 317–345, Oct. 1998, doi: 10.1016/S0927-5398(98)00002-4.

[12] Alla, S. (2019). A Statistical Analysis of Surgeons' Preference on Robot-Assisted Surgeries. In *IIE Annual Conference. Proceedings* (pp. 1385-1390). Institute of Industrial and Systems Engineers (IISE).

[13] N. L. Gowda, B. S. Banjardar, V. Manchala, and A. R. Mohammed, "Bridging Classical Conditioning and Deep Reinforcement Learning: Advancements, Challenges, and Strategies for Autonomous Systems," May 04, 2024. Available: https://journal.esrgroups.org/jes/article/view/4121.

[14] B. B. Mandelbrot, "The variation of certain speculative prices," in *Fractals and Scaling in Finance*, Springer, New York, NY, 1997, pp. 371–418. doi: 10.1007/978-1-4757-2763-0_14.

[15] Alla, S., Soltanisehat, L., & Taylor, A. (2018, July). A Comparative Study of Various AI Based Breast Cancer DetectionTechniques. In *IX International Conference on Complex Systems* (p. 213).

[16] F. Alzazah, X. Cheng, and X. Gao, "Predict Market Fluctuations Based on the TSI and the Sentiment of Financial Video News Sites via Machine Learning," in *2023 15th International Conference on Computer and Automation Engineering, ICCAE 2023*, 2023, pp. 302–308. doi: 10.1109/ICCAE56788.2023.10111493.

[17] V. Bergen, M. Escobar, A. Rubtsov, and R. Zagst, "Robust multivariate portfolio choice with stochastic covariance in the presence of ambiguity," *Quantitative Finance*, vol. 18, no. 8, pp. 1265–1294, Aug. 2018, doi: 10.1080/14697688.2018.1429647.

[18] M. Smita, "Logistic Regression Model For Predicting Performance of S&P BSE30 Company Using IBM SPSS," *International Journal of Mathematics Trends and Technology*, vol. 67, no. 7, pp. 118–134, 2021, doi: 10.14445/22315373/ijmtt-v67i7p515.

[19] S. Rahman *et al.*, "Multi Perspectives Steganography Algorithm for Color Images on Multiple-Formats," *Sustainability (Switzerland)*, vol. 15, no. 5, p. 4252, Feb. 2023, doi: 10.3390/su15054252.

[20] Lakshminarayanan, V., Ravikumar, A., Sriraman, H., Alla, S., & Chattu, V. K. (2023). Health care equity through intelligent edge computing and augmented reality/virtual reality: a systematic review. *Journal of Multidisciplinary Healthcare*, 2839-2859.

[21] F. Alharbi and G. S. Kashyap, "Empowering Network Security through Advanced Analysis of Malware Samples: Leveraging System Metrics and Network Log Data for Informed Decision-Making," *International Journal of Networked and Distributed Computing*, pp. 1–15, Jun. 2024, doi: 10.1007/s44227-024-00032-1.

[22] Chattu, V. K., Alla, S., & Singh, B. (2024). Political prioritization for digital health and health equity through global health diplomacy.

[23] S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," *International Journal of Information Technology 2024*, pp. 1–10, Feb. 2024, doi: 10.1007/S41870-023-01721-W.

[24] Alla, S., Pazos, P., & DelAguila, R. (2017). The impact of requirements management documentation on software project outcomes in health care.

[25] Soltanisehat, L., & Alla, S. (2018). Centraized or distributed IT system?(blockchain concept). In *Proceedings of the International Annual Conference of the American Society for Engineering Management.* (pp. 1-7). American Society for Engineering Management (ASEM).

[26] Alla, S., Soltanisehat, L., Tatar, U., & Keskin, O. (2018). Blockchain technology in electronic healthcare systems. In *IIE Annual Conference. Proceedings* (pp. 901-906). Institute of Industrial and Systems Engineers (IISE).

[27] S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1

[28] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.