Using an Ensemble of Machine Learning Algorithms to Predict Economic Recession

Leakey Omolo and Nguyet Nguyen *

Posted Date: 31 July 2024

doi: 10.20944/preprints202407.2620.v1

Keywords: Economic crisis; machine learning; ensemble model; recession probability

Article

Using an Ensemble of Machine Learning Algorithms to Predict Economic Recession

Leakey Omolo † and Nguyet Nguyen *,†

Department of Mathematics and Statistics, Youngstown State University, Cafaro Hall, Youngstown, OH, USA 44555

* Correspondence: ntnguyen01@ysu.edu; Tel.:+1-330-941-1805

† These authors contributed equally to this work.

Abstract: The Covid19 pandemic and the current wars in some countries have put incredible pressures on the global economy. Challenges for the U.S. include not only economic factors, major disruptions and reorganizations of the supply chains, but also those of national security and global geopolitics. This unprecedented situation makes predicting economic crisis for the coming years crucial yet challenging. In this paper, we propose a method based on various machine-learning models to predict the probability of a recession for the US economy in the next year. We collect the U.S' monthly macroeconomics indicators and recession data from January of 1983 to December of 2023 to predict the probability of an economic recession in 2024. The performance of the individual economics indicator for the coming year was predicted separately, and then all of the the predicted indicators were used to forecast a possible economic recession. Our results showed that the U.S. will face a high probability of being in a recession period in the last quarter of 2024.

Keywords: economic crisis; machine learning; ensemble model; recession probability

1. Introduction

1.1. Overview of Economic Cycle

From an economic perspective, modern economies operate between the trends of expansion, growth and decline. The declining trend in an economy is known as *recession phase*. An economic expansion phase is characterized by improved standards of living, better purchasing power, increase in growth rate, ability to afford better health care and education, among others. On the other hand, recession phases are generally characterized by poor performance of the economy that leads to a reduction in economic growth rate and activities. As a result, people are unable to afford good living standards, or better health care due to reduced income. The dynamics in the economic growth trends affect both societies, businesses and individuals in a society. Companies experience reduced sales and profits due to decline in demand. Governments also suffer the consequences as people and businesses put more pressure as they look up on their leaders to reverse the adverse effects of economic instability. Consequentially, it is becoming extremely necessary that companies, societies and governments have an information on when these shifts are likely to happen ahead of time.

In this paper, we intend to develop a robust machine learning model that studies the signals and patterns of the historical data on economic indicators, and predicts the possibility of a recession or expansion. Python is used as the main programming language and for the exploratory data analysis(EDA).

Many economists,scientists, engineers and mathematicians desire to predict economic recession. The evolution of computing and information technology has made the application of mathematical and economic models easy and more robust. As a result, there has been significant development in studies on ML, Data Mining (DM), Artificial Intelligence (AI) and Language Recognition (LR) prospects [1]. Such steps have led to answers to many real-life problems and questions in the field of economics. Consequentially, many scientist, have found it necessary to apply ML ideas to predict economic recessions.

Although application of ML in various fields has improved a lot in the recent past, there are still gaps, especially in areas where not every facet of information is available for interrogation. The gaps

 doi:10.20944/preprints202407.2620.v1

in existing studies provide study opportunities, hence, this paper provides an avenue for exploration. Therefore, in this section we use the available literature to explore the existing data, models, results, and metrics used in the current studies.

Spyridon, et. al. applied logistic regression, k-nn, and Bayesian generalized models to predict the probability of economic recession in the U.S. in the year 2022 [2]. The validation results showed that the ensemble model outperforms standard econometric techniques and the traditional logit/probit models, achieving a higher predictive accuracy across the long, short, and medium term scales. Another study by [3] applied logistic regression, logistic regression and xgboost to build an ensemble model for forecasting economic recession probabilities. Although the model was not applied on the out of sample forecasts, the validation results from the forecasts using historical data demonstrated a better AUC for the ensemble model (0.9) with xgboost, random forests and logistic regression producing 0.89 and 0.82 respectively.

According to Youngjin, the time varying logit models lead to a large improvement in forecast performance compared to the other econometric techniques, with an accuracy of 89% [4]. Using parsimoniously specified logistic regressions, [5] used data from the FRED website to produce forecasts for the chances of U.S. economic recession in 2021. The results showed a maximum probability of 40% which was below the threshold of 50%, showing no chances of a recession at that time. Such outcome was consistent with the results published by the NBER.

Andreas(2020) proposed a method to forecast economic recession in Australia, Germany, Japan, Mexico, UK and U.S. using machine learning algorithms [6]. Classification metrics were compared for the decision trees, random forests, logistic regression, k-nn and support vector machine(SVM).

Andreas Psimopoulos used macroeconomic indicators that focus on GDP and Yield Curve to produce forecasts for the recession in six different countries [6]. In that case, GDP was considered a better measure and comparison tool across various economies. Other scholars used indicators that lean towards stock prices, interest rates, and money stock to help companies like S & P500 understand their projected performance[3]. Travis Berge employed industrial production, unemployment, manufacturing index among other variables to address the problem of *'producing a simple model using the most relevant indicators'* that many scientists face [7].Many of these studies use between 10 to twenty variables, however, other organizations like IMF use upto 166 indicators generally categorized into: income and output, interest rates, employment rate, nominal prices and wages, and construction [3]. Wells Fargo used a more advanced approach involving upto about 6000 variables,obtained from FRED. For simplicity and proper prediction, Well Fargo reduced the columns of the data set to about 190 indicators [8].

2. Methodology

This section addresses the methods and implementation processes of the solution to the problem of predicting economic recession in the United States. The objective of this section is to describe the methods followed to achieve the final results. Each step has components that play a role in the entire process. The CSV files consists of the variables used for the analysis. Data for the macroeconomic indicators and recession history are retrieved from the FRED and NBER websites respectively. The information is sent to a single folder, for storage and further exploration. During the initial stages of data exploration, data cleaning, granularity analysis for the individual columns occur, with the main objective of identifying any unusual data patterns or behaviors in the data set. Correlation analysis is also done to identify relationship between the variables.

The next step is the **data transformation** where the data for the individual macro-indicators is reshaped to fit the objectives of the project. As a result, memory system, granularity, and the seasonality of the indicators undergo modification before forwarding to the next stage for further interrogation.

At the **forecasting stage**, the available historical data, selected from well defined dates (1983 to 2023) is used to produce 1 year forecast for all the selected indicators. Python's prophet and ARIMA models are used to produce the forecasts, based on their performance metrics for specific variables.

 doi:10.20944/preprints202407.2620.v1

Four machine learning algorithms namely: Logistic regression, Gaussian Naive Bayes, Xgboost and Random Forests are applied on the historical data to train an ensemble model, which is then used to produce predicted probability of the economic crisis in the year 2024 based on the 1-year forecast data.

Figure 1 shows the summary of the steps involved in predicting the economic crisis.

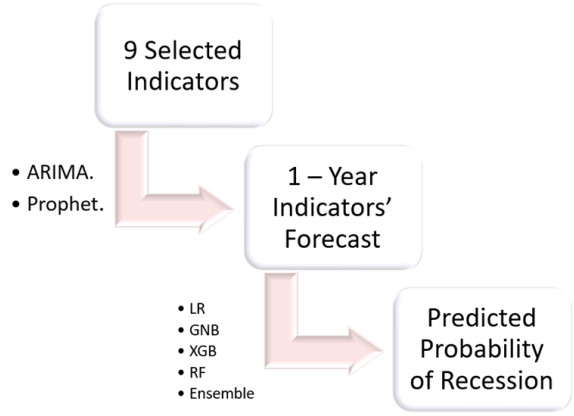


Figure 1. Economic crisis prediction process.

2.1. Data Selection and Preparation

This section focuses on the implementation of the structure and design as discussed in the introduction to this section. All the programming and software work in this project were developed and implemented in python language using Jupiter notebook (anaconda 3) as the IDE.

Selection of the Macroeconomic Indicators:

The macroeconomic indicators' data was retrieved from FRED website with the aid of **FRED Add-in** available in excel. It consisted of many other parameters which were considered irrelevant for the study, hence they were cleaned through a simple deletion process.

To address the problem of the differences between the parameters of the individual macroeconomic indicators, a metadata table was created as shown below.

From Table 1 above, it is clear that the granularity, units and seasonality components are different for the indicators. The choice of these 23 indicators was based on background studies and the results from the contemporary studies using machine learning algorithms. Such studies identify economic variables based on specific characteristics of the economy such as labor, availability of money, consumers' purchasing power, production (both manufacturing and industrial), and trusts in the state of the economy by the consumers and producers [9]. Some of the indicators affect the economy in more than one aspect. For instance, Industrial production may influence labor and and availability of money in the economy, both of which impacts economic performance of a country.

Table 1. Macroeconomic Indicators from the FRED website.

#	Code	Name of the Indicator	Granularity	Units
1	UNRATE	Unemployment Rate	Monthly	Percent, Seasonally Adjusted
2	AHETPI	Average Hourly Earnings of Production and Nonsupervisory Employees, Total Private	Monthly	Dollars per Hour, Seasonally Adjusted
3	PERMIT	New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Total Units	Monthly	Thousands of Units, Seasonally Adjusted Annual Rate
4	CSCICP03USM665S	Consumer Opinion Surveys:OECD Indicator for United States	Monthly	Normalised (Normal=100), Seasonally Adjusted
5	AAA10Y	Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity	Daily	Percent, Not Seasonally Adjusted
6	M2REAL	Real M2 Money Stock	Monthly	Billions of 1982-84 Dollars, Seasonally Adjusted
7	CPIAUCSL	Consumer Price Index for All Urban Consumers: All Items in U.S. City Average	Monthly	Index 1982-1984=100, Seasonally Adjusted
8	FEDFUNDS	Federal Funds Effective Rate	Monthly	Percent, Not Seasonally Adjusted
9	PAYEMS	All Employees, Total Nonfarm	Monthly	Thousands of Persons, Seasonally Adjusted
10	PCE	Personal Consumption Expenditures	Monthly	Billions of Dollars, Seasonally Adjusted Annual Rate
11	GS10	Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity.	Monthly	Missing Units
12	T10Y2Y	10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity	Daily	Percent, Not Seasonally Adjusted
13	DGS10	Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity	Monthly	Percent, Not Seasonally Adjusted
14	CPILFESL	Consumer Price Index for All Urban Consumers	Monthly	Index 1982-1984=100, Seasonally Adjusted
15	HOUST	New Privately-Owned Housing Units Started: Total Units	Monthly	Thousands of Units, Seasonally Adjusted Annual Rate
16	WPSFD49207	Producer Price Index by Commodity: Final Demand: Finished Goods	Monthly	Index 1982=100, Seasonally Adjusted
17	GS1	Market Yield on U.S. Treasury Securities at 1-Year Constant Maturity	Monthly	Percent, Not Seasonally Adjusted
18	DSPIC96	Real Disposable Personal Income	Monthly	Billions of Chained 2017 Dollars, Seasonally Adjusted Annual Rate
19	INDPRO	Industrial Production: Total Index	Monthly	Index 2017=100, Seasonally Adjusted
20	SPCPI	Consumer Price Index	Monthly	Missing Units
21	IC4WSA	4-Week Moving Average of Initial Claims	Weekly	Number, Seasonally Adjusted
22	IPMAN	Industrial Production: Manufacturing (NAICS)	Monthly	Index 2017=100, Seasonally Adjusted
23	WM2NS	Money Stock	Weekly	Billions of Dollars, Not Seasonally Adjusted

 doi:10.20944/preprints202407.2620.v1

The Recession Data. The US economic recession data and dates was identified as the outcome variable in this study. Eve though NBER provides data from as early as 1840's, data was downloaded in CSV format for the duration between January 1983 to December 2023. This exists in binary format in the form of 0's and 1's (0=no recession, 1 = recession). A quick exploration of the data set shows that there are more 'no recession' incidences than 'recession' events (imbalance), and this needed to be addressed. The data had monthly granularity, and the information was enough to create training and testing sets for the models.

2.2. The Individual Indicator Exploration

The individual signal analysis aims at providing an understanding of every indicator selected for further analysis and classification. By using *pandas library* to retrieve summary data about the indicators, we obtain information that helps identify possible necessary changes and transformations on the original data.

To identify trends and uncover underlying seasonality variations, moving averages is applied.The moving average for 12 months is plotted on the same axes as the US recession linear graph for each macroeconomic indicator. Figures 5 and 6 show the output for two of the indicators (Industrial production and Unemployment rate) that were used in the final predication.

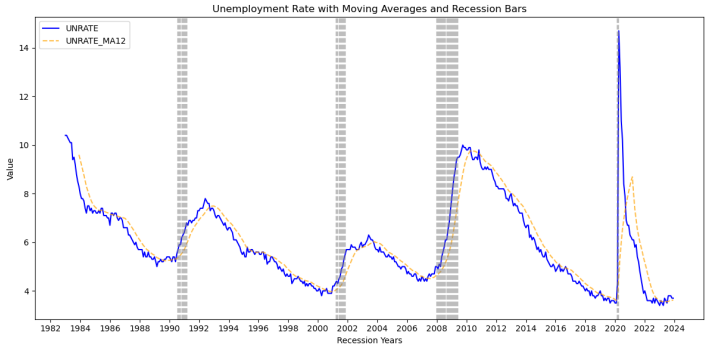


Figure 2. Moving average for the unemployment rate.

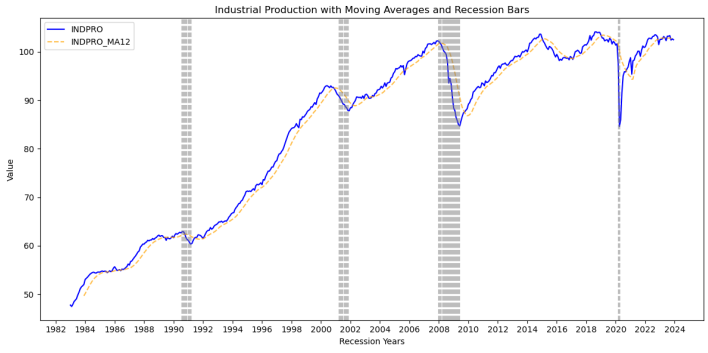


Figure 3. Moving average for the Industrial Production.

From the 5 and 6 above, the vertical gray bars represent the US economic recessions. The plots help the user visualize the evolution of the signals over time, and to understand the behavior of the signals before and after the incidents of economic crises. The original signal is represented in blue, while the 12-month moving average (MA) plot is in orange. Although the MA has a clear lag at the beginning, it shows the trend of the signal over the years that were considered in this study.

Next, we plotted the annual and monthly variations of all the indicator signals with the economic recession bars. The aim was to understand how the signals respond to recession events. Also the step was useful in identifying the most relevant indicators for the study, as some of the selected variables

 doi:10.20944/preprints202407.2620.v1

did not show any significant variations with respect to the recession bars over time. The figure 4 below shows the time series plot for the personal consumption expenditure.

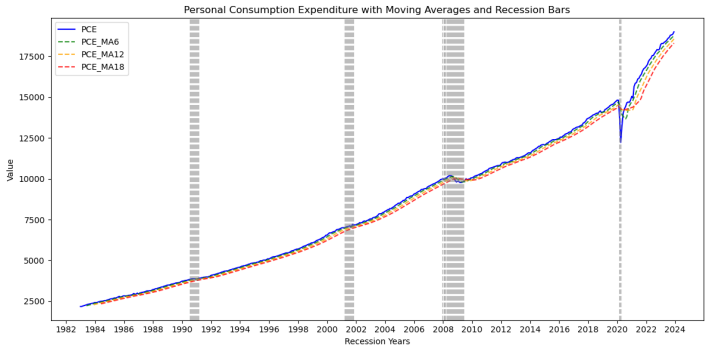


Figure 4. Personal Consumption Expenditure plot showing no significant variation.

Annual and monthly variation were plotted as shown in the figures below to provide a visual understanding of how the signals change over time. From the plots, we observe that the variations are clearer when data is aggregated on yearly basis.

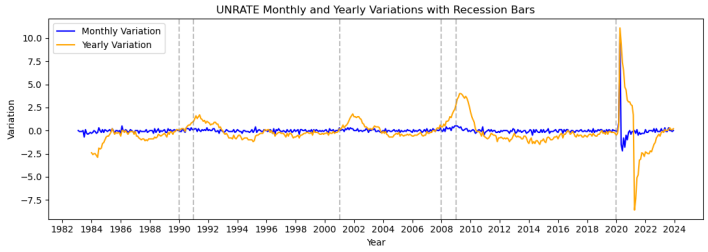


Figure 5. Annual and Monthly variations for Unemployment Rate.

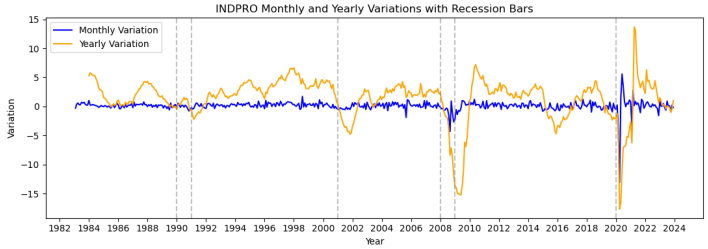


Figure 6. Annual and Monthly variations for the Industrial Production.

The final step of the individual signal exploration involved use of the *Facebook Prophet library* and *ARIMA* to produce forecasts for the individual signals. Although the tool does not produce very reliable forecasts for monthly granularity, it works so well with daily and weekly time frames, and produce good graphics for identifying trend and seasonality[6]. Therefore ARIMA was used to as a supplementary forecasting tool in instances when Prophet produced extremely unreliable results (in terms of accuracy).

2.3. General Signal Analysis

The general signal analysis involves evaluation of the signals in a more generalized manner. The process involves cross examination of the macroeconomic indicators on a single Jupiter notebook. To address the problem of differences in units, normalization is done to allow easy comparison. The *Scikit/sklearn* in python uses the standard scaler tool to remove the mean, and scaling of the unit

 doi:10.20944/preprints202407.2620.v1

variance for all the signals[7]. The *missingno library* is then used to produce the triangular correlation heat map as shown in Figure 7 below:

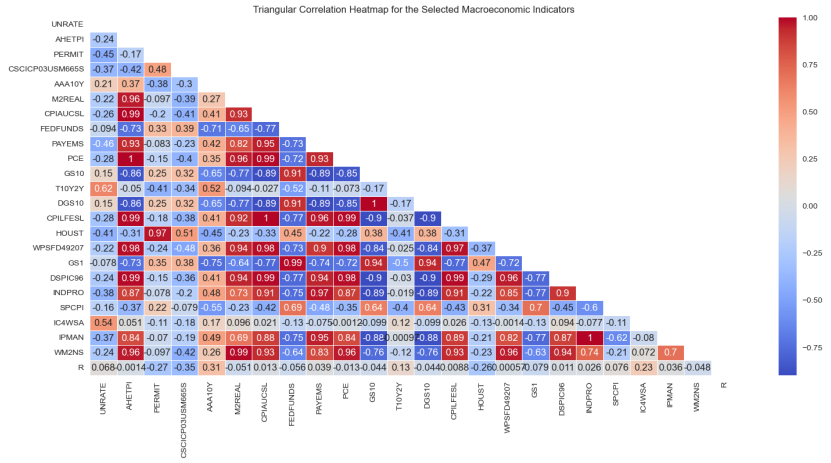


Figure 7. Correlation co-efficient between the 23 macroeconomic indicators.

The Figure 7 above illustrates the correlation strength between various macroeconomic indicators. The values range between -1 (highest negative correlation to +1(highest positive correlation), with 0 (no correlation) as the mean point. The tool helps identify auto-correlation, which would otherwise be one of the contributing factors to over-fitting of the classification model. For this paper, the correlation of the indicators, together with the outcomes from the individual signal analysis was considered while selecting the indicators to be used in the classification algorithm. The following section discusses the transformation in the system design.

2.4. Economic Indicator Selection

The next step involved selection of the 9 variables to be used in the economic crisis prediction model, and the process was based on the following factors:

- Minimizing correlation between the indicators: Indicators showing high correlation as observed from Figure 7 were either dropped to minimize redundancy or retained after normalization.
- Areas of interest in the economy: Even though some indicators showed high correlations, they were still selected to capture the economic areas of interest for the recession such as labor, availability of money, consumer purchasing ability and trust [10].
- Relationship with the target variable: The indicators with higher correlation coefficients with the recession were selected.

After the selection process, the macroeconomic indicators were reduced to 9 as shown in the Table 2 below.

Table 2. 9 Selected Macroeconomic Indicators for the Economic crisis prediction model.

Code	*Name of the Variable*	*Granularity*
UNRATE	Unemployment Rate	Monthly
AHETPI	Average Hourly Earnings of Production and Nonsupervisory Employees, Total Private	Monthly
PERMIT	New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Total Units	Monthly
CSCICP03USM665S	Consumer Opinion Surveys:OECD Indicator for United States	Monthly
AAA10Y	Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity	Daily
T10Y2Y	10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity	Daily
INDPRO	Industrial Production: Total Index	Monthly
SPCPI	Consumer Price Index	Monthly
IC4WSA	4-Week Moving Average of Initial Claims	Weekly

From the Table 2 above,, it is noted that the dataset is much more reduced, and transformation presents itself as a way to increase information. The frequency aggregation tool in excel was applied to

 doi:10.20944/preprints202407.2620.v1

address the problem of different granularities since it allows for conversation of higher frequency data series to lower frequency series such as daily to weekly, or weekly to monthly using the *mean(), max(), min() and standard deviation* [11].

In this project, a six-month lag was added to the US recession dates as a strategy to allow the algorithm detect the economic downturns before they actually happen, by 'dragging' the signals 6-month into the past. A simple shift technique was applied to add lag to the original signal.

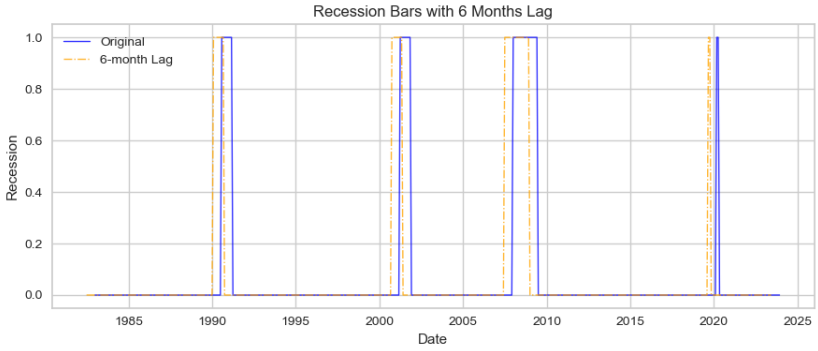


Figure 8. US Recession Dates with 6-Month Lag.

To deal with the problem of discontinuous bars, a more ingenous method had to be designed. The idea was to apply logical operations to discover the start date for the recession, choose a prior date according to the lag,then extend the signal through the end date of the recession. The $shift\_fill()$ tool in the *month_delta library* was applied together with the logical operations $logical\_and()$, $logical\_xor()$ and $logical\_or()$. Before the data set can be used for classification, re-sampling is performed to solve the problem of imbalance in the recession dataset. Two techniques were used: Random over sampling (ROS)(for the linear and probabilistic algorithms) and the Synthetic Minority Oversampling Technique (SMOTE) for the non-linear algorithms. ROS resamples by duplicating some of the original samples of the minority class, while SMOTE focuses on generating samples next to the wrongly classified outcomes using k-nearest neighbors (KNN).

3. Predictive Models

The U.S. economics crisis prediction consists of two steps. The first step is predicting the Economics indicators, and the second step is using the predicted indicators to forecast the economics crisis.

3.1. Predictive Models for Economic Indicator

Since the main aim of this project was to generate predictions for future recession, it was essential to produce forecasts for the individual indicators. In the first step, to enhance the accuracy and reliability of these forecasts, a Box-Cox transformation was applied to the macroeconomic indicators. This transformation helped stabilize variance, normalize the distribution of the data, and linearize relationships, making the underlying patterns in the data more pronounced and the forecasting models more effective. Consequently, a one year (12 months) forecast for every feature was generated using *prophet()* and *ARIMA()* tools in Python (January 2024 through December 2024). The initial objective was to use prophet for all the variables; however, it was observed that the tool produced very unreliable results for some of the indicators, especially those with more than weekly granularity. To determine which tool to apply to which indicator, performance metrics (MAE and RMSE) were generated from the transformed data for all the individual variables and compared.

 doi:10.20944/preprints202407.2620.v1

3.2. Predictive Models for Economic Crisis

At the second step, we train four binary classification models using historical data, and take the average of the results to build a more stable ensemble model. The models are: Naive Bayes (NB), XGBoost (XG), Random Forest (RF), and Logistic Regression (LR).

We use the models to fit the U.S. recessions with and without transformations then and compare the outcomes. Figure 9 represents the procedures of our predictive models to predict economics crisis by showing the training and testing, and forecasting data. The training was done through the forward chaining cross-validation process.

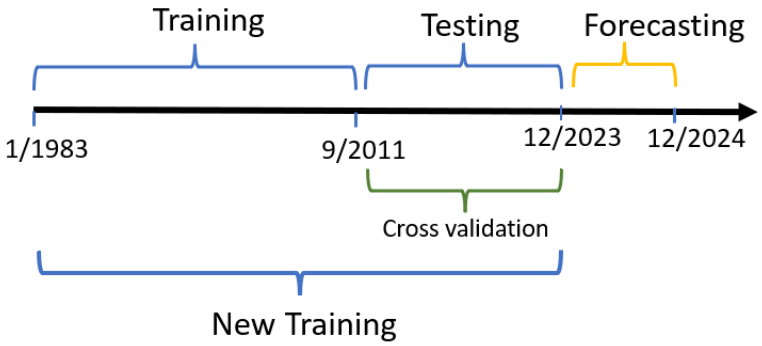


Figure 9. Procedures of predicting economics crisis using machine learning models.

3.2.1. Logistic Regression Algorithm (Linear)

The implementation of logistic regression, involves data importation from the 'selected and transformed' step and the lagged recession database. The process starts matching the end dates using the $equal\_dataframe()$ tool, then followed by re-sampling so that the the entire data-frame has the same size. the dataset was initially split in the ratio of 7:3 for the train and test sets respectively. Cross-validation is then used to fit the model and predict new values through an iterative process. once the iteration process is complete over the k-folds, a new data set is returned. $Predict\_proba()$ is applied to forecast and return return the probability that either of the classification outcomes (0 or 1) is true. Once the results are ready, $show\_forecast()$ from the $plotly\_library()$ is used to display the results. The confusion matrix and classification reports are used to visualize, evaluate and validate the outcomes of the process. More about these tools will be discussed in the results section. Since logistic regression is a linear model, we expect the algorithm to benefit from data normalization and transformation. The number of iterations in the cross-validation process depend on the number of folds, which was determined through trial and error.

3.2.2. Gaussian Naive Bayes Algorithm (Probabilistic)

Naive Bayes algorithm assumes that the macroeconomic indicators are normally distributed. The algorithm was useful for this paper since the features (indicators) were continuous and could be assumed to have a Gaussian (normal) distribution. The implementation of this algorithm involved calculating the mean and variance for each indicator within each class, and then using these parameters to compute the probability of given value, assuming a Gaussian distribution. This probability is then used to classify new instances based on the highest probability of the class given the feature value.

As was the case with the logistic regression, data is imported from the 'selected and transformed', and the 'lagged recession' databases. The 'date' column is initially dropped because it does not contribute to the model's learning process directly as a feature. Subsequently, the dataset is split into training and testing sets, with a 7:3 split ratio. This division ensures that the model can be trained on a

 doi:10.20944/preprints202407.2620.v1

large portion of the data while being evaluated on a separate, unseen portion to assess its predictive performance.

A Gaussian Naive Bayes classifier is instantiated and trained on the balanced training set. Naive Bayes classifiers are well-suited for this task due to their efficiency and effectiveness in handling categorical data and their independence assumption, which simplifies the computation. The trained model's performance is evaluated using a confusion matrix and classification report, providing insights into its accuracy, precision, recall, and F1 score across both classes (recession and non-recession). The accuracy score is specifically highlighted, offering a quick measure of how often the model predicts correctly.

3.2.3. Random Forests and Xgboost(Non-Linear)

The implementation of non-linear models is almost similar to the linear approach with slight differences. The Random forests and Xgboost are both non-linear algorithms whose implementation differ from the linear models in terms of creation. Therefore, explanation of the two non-linear models is typically the same.

The importation of the features and target variable datasets the same way as the linear and and probabilistic approaches. However, here, SMOTE is applied for re-sampling of the recession outcomes instead of the ROS. Also, *randomforestclassifier()* and *XGBclassifier()* are used respectively to create the classification models. The values of folds and estimators are determined through a trial and error approach, with the number of remaining the same as that used in the linear model. The number of estimators was taken o be 100, so as to produce a good prediction without relying entirely on the power of the computer. The results were then prepared for validation and plotted to visualize the outcomes.

Since the non-linear approaches have an embeded $feature\_importance()$ tool, it slightly differ from the linear model. The tool is useful in demonstrating the significance of each variable in the classification process [12]. The results are presented in a probabilistic manner, while the feature importance is visualized through bar graphs using *matplotlib* and *seaborn* libraries in python. Also, the feature importance could be used to eliminate predictor variables that do not attain a certain score [13]. However, for the purposes of this paper, that step was not undertaken as all the useful indicators had been identified through the feature selection and transformation process. The final step is to forward the classification outcomes to the validation where various metrics are used to evaluate the models performance.

3.2.4. Ensemble Model

The ensemble model is the combination of the four predictive models by averaging the predictions of the four models: Naive Bayes, XGBoost, Random Forest, and Logistic Regression. This model aims to leverage the strengths and mitigate the weaknesses of individual models, potentially leading to more robust and reliable predictions [6].

By taking the average results of the four prediction models, the ensemble model reduces variance and bias, drawing on the principle that a group of models can often make more accurate predictions collectively than any single model could individually.

Note that in the five of the models above, Since to obtain the binary outcomes of the economics crisis status: 0 and 1, the continuous probabilities obtained from the models were converted into binary predictions by applying a threshold of 0.5. The predicted probability above this threshold are classified as 1 (indicating a recession), and those below are classified as 0 (indicating no recession). The final results for the predictions are forwarded to the validation for evaluation.

4. Results

In this section, we report the results and observations from the two steps followed in section 3: the predictions of the economics indicators and the predictions of the U.S. economics crisis.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2024 doi:10.20944/preprints202407.2620.v1

4.1. Economic Indicator Forecasting Results

After applying the two models (*prophet()* and *ARIMA()*) on the individual economic indicators, forecast data frame was created as shown below.

	date	SPCPI	INDPRO	AAA10Y	AHETPI	CSCICP03USM665S	PERMIT	UNRATE	IC4WSA	T10Y2Y
0	2024-01-01	4.226885	101.379730	1.132066	28.654846	97.602266	1405.111958	3.719448	396194.6678	0.369628
1	2024-02-01	4.196419	101.507719	1.138630	28.738097	97.576642	1387.976533	3.741147	423648.2614	0.361406
2	2024-03-01	4.248352	101.712413	1.163932	28.805814	97.616403	1389.163157	3.477531	351894.6534	0.319486
3	2024-04-01	4.247027	100.897587	1.112810	28.956393	97.505362	1391.237244	4.114789	601412.4345	0.320666
4	2024-05-01	4.294448	100.967085	1.065371	29.016499	97.430135	1388.620782	3.974844	458320.8833	0.309522
5	2024-06-01	4.343839	101.380367	1.044801	29.094962	97.362673	1391.205613	3.807514	382734.3151	0.245728

Table 3. Forecasts from the raw data (cropped).

The entire data set was used to train the forecasting models model and plots for each indicator generated. The figures below shows the forecasting plots for two of the macroeconomic indicators. The rest of the plots are included in the appendix.

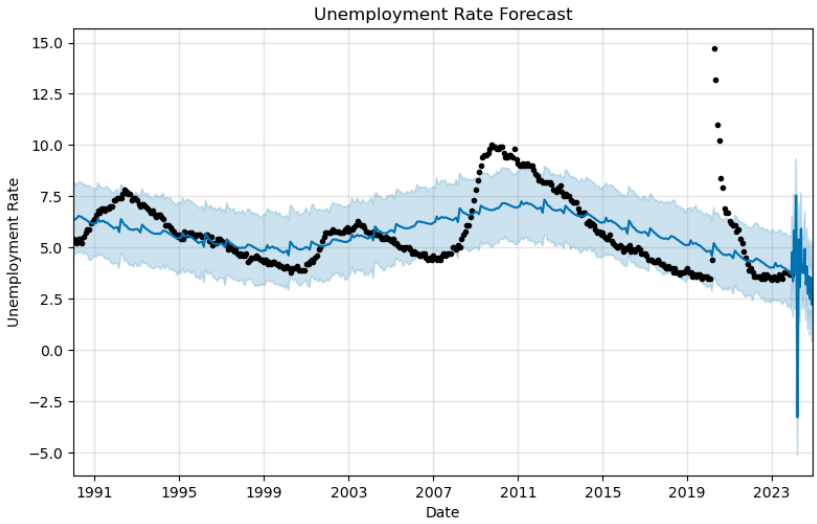


Figure 10. Unemployment Rate Forecasts as generated by the *prophet()*

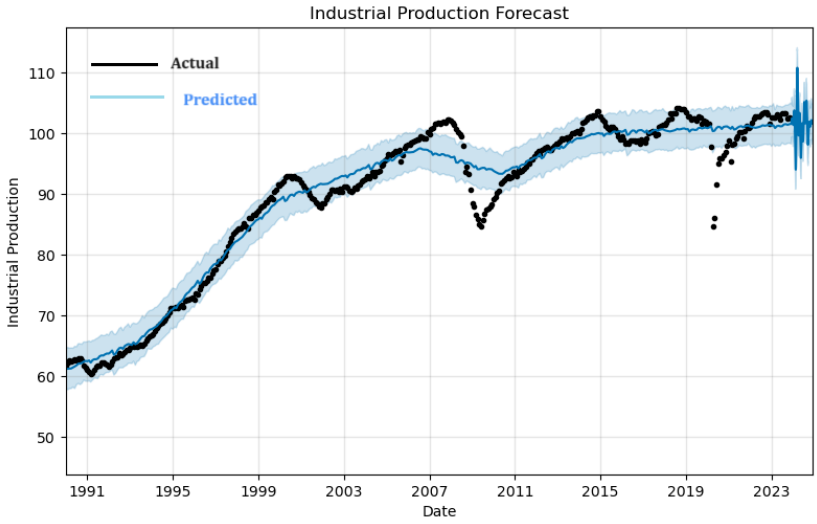


Figure 11. Industrial Production Forecasts as generated by the *prophet()*

 doi:10.20944/preprints202407.2620.v1

Industrial production forecast shows a general increasing trend from 1991 until the early 2020's, with a sharp uptick followed by significant volatility in the forecast period post-2023. The shaded area represents confidence intervals, and it widens substantially towards the end of the forecast period, indicating higher uncertainty or variability in the industrial production projections. Unemployment rate forecast, indicates a fluctuating trend around a stable mean until the early 2020s. However, in the forecast period, there is a pronounced spike in unemployment rates, which sharply reverses into a steep decline towards the end of the period. The confidence interval also widens dramatically, reflecting considerable uncertainty in future unemployment rates.

The two plots suggest that a significant economic event or change in policy might be anticipated around 2024, affecting both industrial production and unemployment rates with a more pronounced immediate impact on unemployment.

Before using the forecasts, a normalization process was necessary since the model had been built on a normalized data set. To ensure the conformity, the forecasts were transformed through a *min_max scaling* process to produce a new dataset that could now be applied in the model.

Table 4. Transformed Forecasts (first six months).

	date	SPCPI	INDPRO	AAA10Y	AHETPI	CSCICP03USM665S	PERMIT	UNRATE	IC4WSA	T10Y2Y
0	2024-01-01	0.109908	0.591713	0.752285	0.000000	0.976129	1.000000	0.379622	0.177543	1.000000
1	2024-02-01	0.000000	0.748789	0.803313	0.088882	0.932863	0.000000	0.413672	0.287569	0.959002
2	2024-03-01	0.187349	1.000000	1.000000	0.161180	1.000000	0.069250	0.000000	0.000000	0.749959
3	2024-04-01	0.182569	0.000000	0.602598	0.321944	0.812504	0.190291	1.000000	1.000000	0.755842
4	2024-05-01	0.353646	0.085293	0.233825	0.386116	0.685481	0.037597	0.780395	0.426528	0.700272
5	2024-06-01	0.531828	0.592495	0.073926	0.469887	0.571570	0.188445	0.517817	0.123597	0.382152

4.2. Model Validation Results

In the following subsections, we provide results for the individual models used in this paper, followed by the outcome of the overall ensemble algorithm.

4.3. Linear Model - Logistic Regression(LR)

A linear model was applied, the indicators' signal were normalized and lag added to the recession dates to enhance early detection of the recession. Upon comparing the confusion matrices before and after a transformation, we observe a substantial improvement in the model's classification ability. In the pre-transformation matrix, the model correctly predicted 114 instances of Class 0 (true negatives) and 9 instances of Class 1 (true positives), while incorrectly predicting 22 instances as Class 0 that were actually Class 1 (false negatives) and 3 instances as Class 1 that were actually Class 0 (false positives). Following the transformation, the true negatives increased to 126 and true positives to 11, indicating better accuracy in classifying both negative and positive cases. The false negatives decreased to 10, showing that fewer actual Class 1 instances were misclassified as Class 0. The false positives were reduced to just 1, significantly minimizing instances where Class 0 was incorrectly identified as Class 1. These changes demonstrate a marked increase in the model's predictive precision and recall, reflecting a significant enhancement in overall performance, particularly in reducing the mis-classification of Class 1, which is often the more critical metric in imbalanced datasets.

We also compared the classification reports for the two instances of classification and achieved the following outcomes.

 doi:10.20944/preprints202407.2620.v1

	precision	recall	f1-score	support
0	0.97	0.84	0.90	136
1	0.29	0.75	0.42	12
accuracy			0.83	148
macro avg	0.63	0.79	0.66	148
weighted avg	0.92	0.83	0.86	148

The Accuracy score: 83.11%
The AUC score: 0.712

Figure 12. Before Transformation.

	precision	recall	f1-score	support
0	0.99	0.93	0.96	136
1	0.52	0.92	0.67	12
accuracy			0.93	148
macro avg	0.76	0.92	0.81	148
weighted avg	0.95	0.93	0.93	148

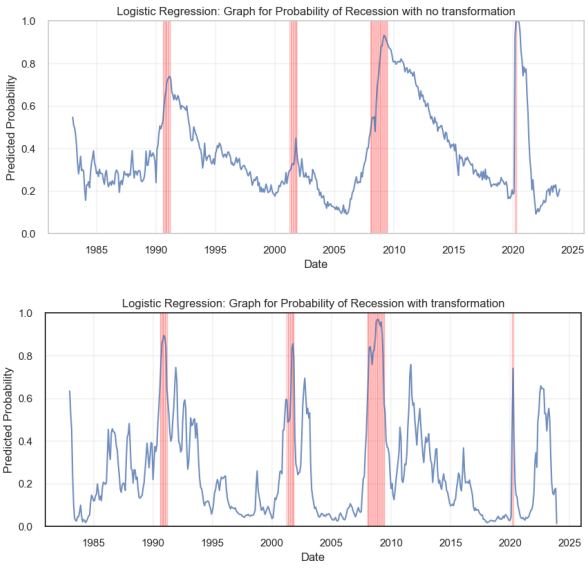
The Accuracy score: 92.57%
The AUC score: 0.794

Figure 13. After Transformation.

From the classification report, the accuracy rose from 0.83 to 0.93. The increase signifies a substantial enhancement in the model's general predictive abilities. Additionally, both the macro and weighted averages for precision, recall, and f1-score exhibited significant improvements. The macro average, which treats both classes equally, showed increases across all metrics: precision (0.63 to 0.76), recall (0.79 to 0.92), and f1-score (0.66 to 0.81). These improvements underscore a balanced boost in model performance for both classes.

Similarly, the weighted average, also showed increases: precision (0.92 to 0.95), recall (0.83 to 0.93), and f1-score (0.86 to 0.93). This comprehensive improvement indicates that the transformation's benefits extend to the model's ability to handle imbalanced data.

Next, we provide a graphical visualization of the probability happening. This was done and compared for the cases of with and without transformation.



Before transformation, Figure ?? displays a smoother curve with more gradual increases in predicted probabilities preceding the recession periods, highlighted by vertical red lines. After transfor-

 doi:10.20944/preprints202407.2620.v1

mation, Figure ?? exhibits sharper and more significant spikes in predicted probabilities, particularly during known recession periods like the early 1990s and the 2008-2009 financial crisis. These heightened peaks suggest that the transformed model may have a heightened sensitivity to the indicators of a recession, providing a more pronounced warning in advance. Overall, the transformation seems to have enhanced the model's ability to detect the likelihood of a recession.

4.3.1. Probabilistic Model - Gaussian Naive Bayes(GNB)

Here, a probabilistic model was applied, the signals were normalized and a lag added to the recession dates, just as was the case with the linear model. Ros resampling technique was applied on the imbalanced recession data with lag. The following results were obtained, and observations reported.

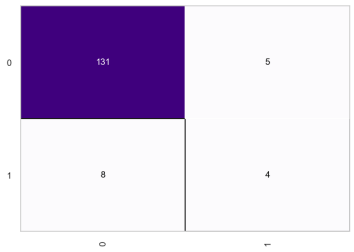


Figure 14. Before Transformation.

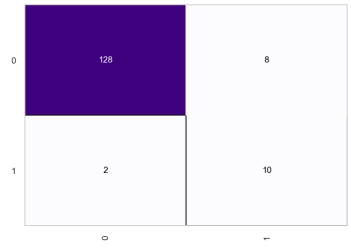


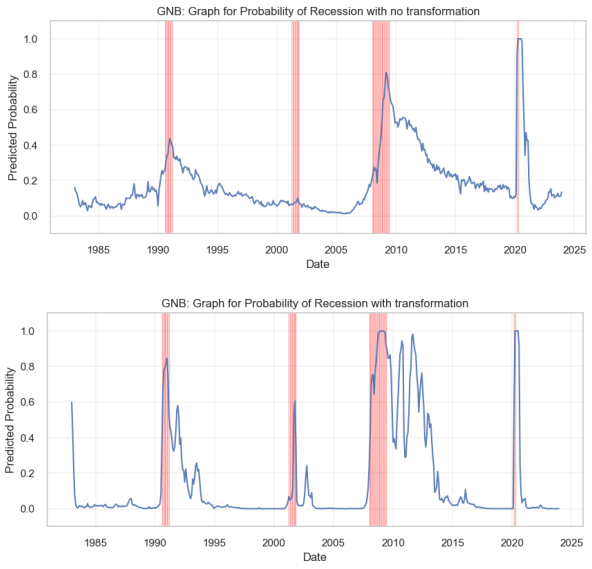
Figure 15. After Transformation.

From the confusion matrices, we observe a total of 131 true positives and 4 true negatives, with the number of false negatives being higher at 8 compared to the 5 false positives. After the transformation, the true negatives increase to 10, indicating an enhanced ability to correctly identify negative cases. The true positives slightly decrease to 128, which is a minimal change. However, there's a noteworthy reduction in false negatives to 2, showing the model's improved sensitivity in correctly predicting positive cases. Meanwhile, the false positives have risen to 8, suggesting a trade-off where the model is now mistaking some negatives for positives more often than before. Overall, the transformation seems to have favorably impacted the model's predictive accuracy, particularly by reducing false negatives, which is often crucial in scenarios where missing out on a positive prediction could have significant consequences.

The classification report of the Gaussian Naive Bayes model shows a marked improvement from the first outcome (no transformation) to the second (after transformation). Initially, the model struggled with the minority class, indicating a potential issue with class imbalance, as evidenced by the lower performance metrics for that class. However, after transformation, there is a significant enhancement in the model's ability to correctly identify the minority class, which is reflected in the improved recall and F1-score for that class. This suggests that an adjustment resampling has been successfully applied to address the imbalance and improve the overall performance of the model. Consequently, the macro average and weighted average scores across all metrics also show substantial improvement, confirming that the model's predictive power has become more balanced and effective across both classes.

 doi:10.20944/preprints202407.2620.v1

The graphical representation for the recession probabilities as given by the model is as shown below.



From the graphs, we observe a relatively smoother curve with fewer and less pronounced peaks in the pre-transformation plot. The observation suggests that the model is conservative in predicting the likelihood of a recession. Conversely, the post-transformation graph displays a more volatile behavior with numerous sharp peaks, indicating a model that is more sensitive to the underlying factors that may signal a recession. Particularly, the period around 2005 shows a significant change where the post-transformation model predicts several high-probability peaks that are absent in the pre-transformation model. This increased sensitivity could be beneficial if the model is capturing true positives (actual recessions) more effectively, but it also runs the risk of producing more false positives (predicting a recession when there isn't one). The model provides a trade-off between the cost of missing a recession versus the cost of a false alarm.

4.3.2. Non-Linear Models: Random Forests (RF) and eXtreme Gradient Boosting (Xgboost)

The non-linear models includes random forests and Xgboost. Similar formulations were used to set the models, and results recorded and reported. One of the most significant observations was that the two models did not rely so much on transformations of the indicators, however, addition of lag was still necessary at it served the purpose of early recession detection. The following confusion matrix shows the outcome for the random forests before and after transformation.

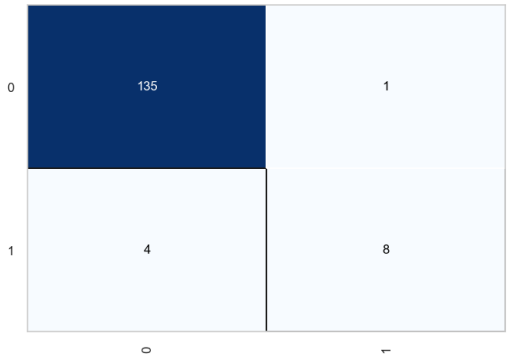


Figure 16. Classification matrix of the random forests before transformation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2024 doi:10.20944/preprints202407.2620.v1

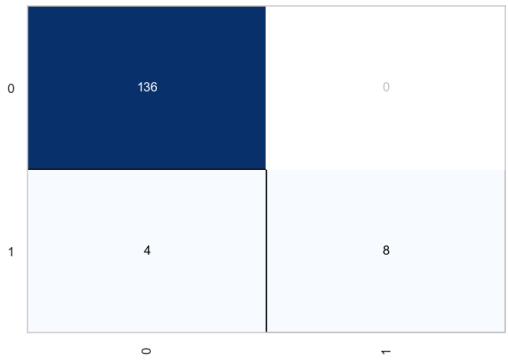
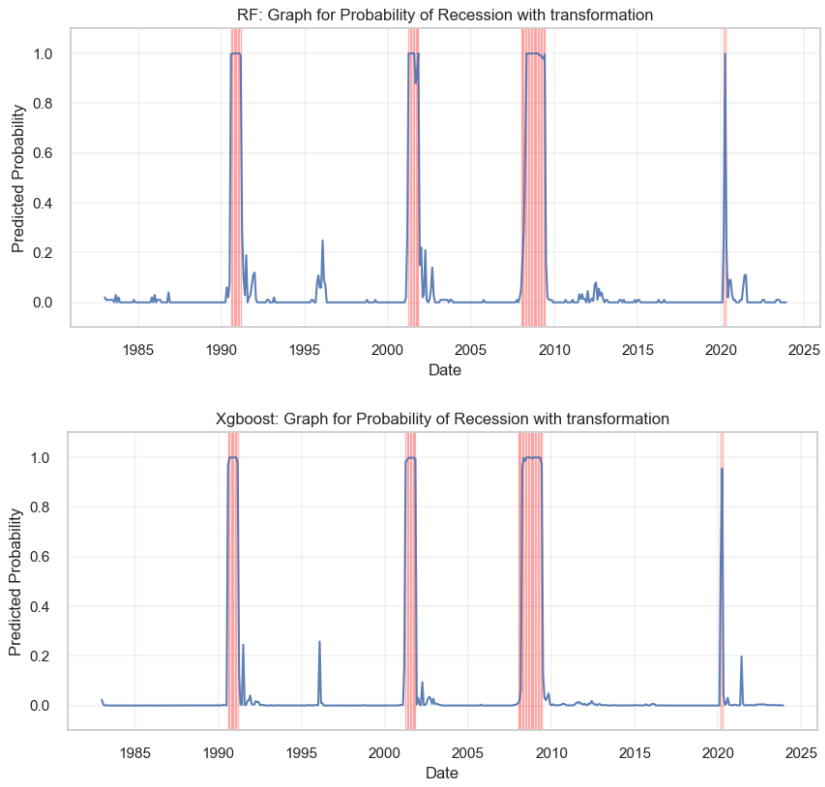


Figure 17. Classification matrix of the random forests after transformation.

From the above plots, we observe that transformation had very little impact in the classification process. The number of true negatives increased by 1, while the number of true positives remained the same for each case. From the classification report, the accuracy improved from 96.62 to 97.30 which is not a significant increase. Similar observations were noted in the case of Xgboost.

Next we plotted the probabilities and compared the results for the two non-linear models.



Comparing above probability plots for recession, we observe contrasting predictive behaviors. The RF model generates a graph with numerous peaks, some aligning well with the historical recessions as marked by the red vertical line, indicating a certain level of sensitivity to the conditions leading to a recession. The XGBoost model's plot is markedly smoother with fewer peaks, suggesting a more conservative approach to predicting recessions. Notably, the XGBoost plot shows fewer false positive peaks, but it also seems to miss some known recession periods. Overall, the RF model appears more sensitive, potentially at the cost of false alarms, while the XGBoost model seems to prioritize minimizing false positives, even if it risks missing some true positives. The two models were applied

 doi:10.20944/preprints202407.2620.v1

in the ensemble model because the bagging (for RF) and boosting (xgboost) techniques embeded in the algorithms an yield diverse perspectives on the data.

4.3.3. Feature Importance for the Non-linear Models

The non-linear models have a $feature\_importance()$ attribute which is not available in the other algorithms. We present these plots to demonstrate the contribution of each non-linear model towards the prediction of recession probabilities.

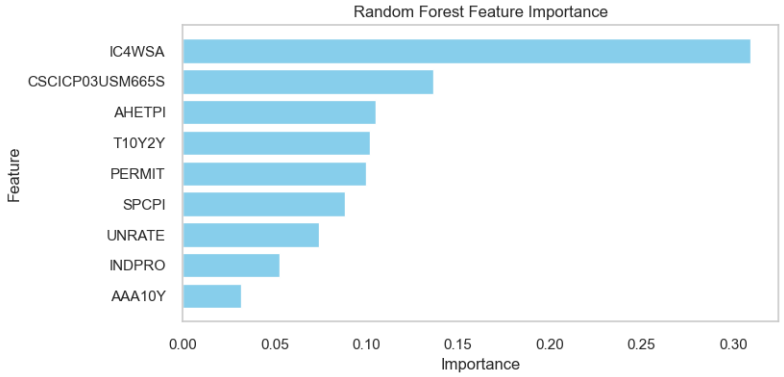


Figure 18. Random Forests Feature Importance Scores.

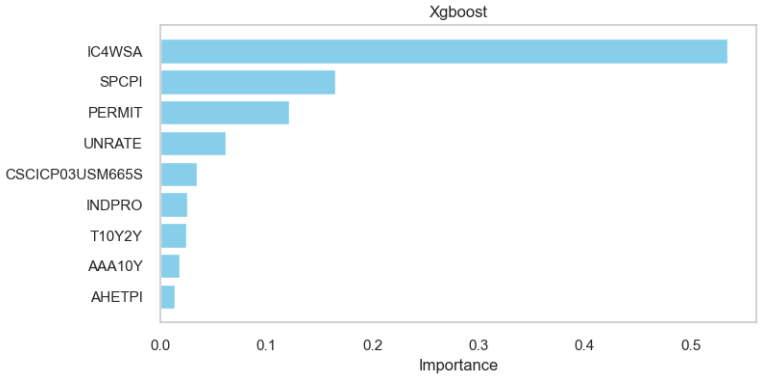


Figure 19. Xgboost Feature Importance Scores.

In the Random Forest model, the feature 'IC4WSA' stands out with the highest importance score, significantly more influential than the others. The features 'CSCICP03USM665S' and 'AHETPI' also show notable importance, though less than 'IC4WSA'. On the other hand, the XGBoost model assigns the greatest importance to 'IC4WSA' as well, but with an even more dominant weight compared to the other features. Features like 'SPCPI', 'PERMIT', and 'UNRATE' appear more prominent in the XGBoost model than in the Random Forest model. The XGBoost model also assigns substantial importance to 'AHETPI' and 'AAA10Y', whereas these features are less significant in the Random Forest model. These discrepancies underscore the differences in how each model processes.

4.4. Model Ensembling - Averaging Approach

The last stage of model configuration was the most different one as it did not rely on one algorithm but a combination of the four discussed above. Average was applied as an ensembling technique, to build a robust model that utilizes the strengths of the linear, probabilistic and non-linear approaches, to compensate for the weaknesses of each algorithm.

The outcomes for the predicted probabilities were obtained for the averaged model and plotted as shown in the following graph.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2024 doi:10.20944/preprints202407.2620.v1

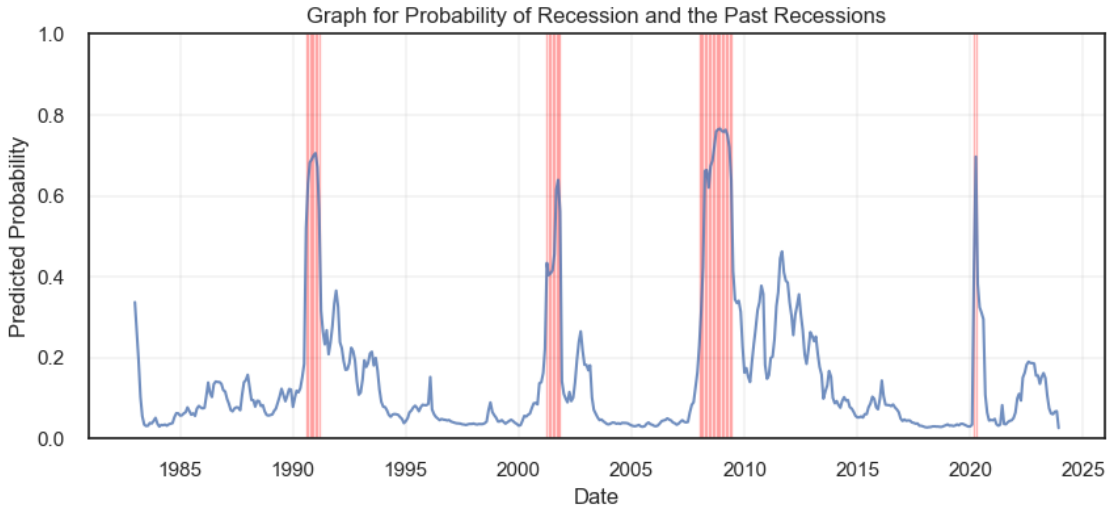


Figure 20. Training plots for the predicted probabilities of the US recession for the Ensembled Model.

4.4.1. Confusion Matrix and Classification Report

Results from the previous models were used to configure the algorithm, and generate the results. Consequentially, the following classification report was generated.

```
Ensemble Model Accuracy: 0.972972972972973

Confusion Matrix:
[[136   0]
 [  4   8]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       136
           1       1.00      0.67      0.80        12

    accuracy                           0.97       148
   macro avg       0.99      0.83      0.89       148
weighted avg       0.97      0.97      0.97       148
```

Figure 21. Classification report for the models' average.

The report provides an outline of the results in the form of a confusion matrix and classification report.

From the confusion matrix, we observe that the model achieved 136 true positives and 8 true negatives. This result indicates that the model is particularly adept at identifying the majority class (class 0) with a perfect recall rate. However, there are 4 instances where the model incorrectly predicted the minority class (class 1), signifying false negatives.

From the classification report, the precision for class 0 is 0.97, which is very high, and perfect for class 1, suggesting no false positives for the minority class. The recall rate is perfect for class 0 but stands at 0.67 for class 1, indicating that while the model captures all instances of class 0, it misses about a third of the instances of class 1. The f1-score, which combines precision and recall into a single measure, is excellent for class 0 at 0.99 and quite good for class 1 at 0.80, indicating a reasonably well-balanced model for both classes.

The model's overall accuracy is mirrored in the weighted average scores for precision, recall, and the f1-score, each standing at 0.97, aligning closely with the model's accuracy which stands at 0.973.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2024 doi:10.20944/preprints202407.2620.v1

The macro average scores, which treat both classes equally, are also high, with precision at 0.99, recall at 0.83, and the f1-score at 0.89.

4.4.2. ROC Curve and AUC

The ROC curve and AUC were used to illustrate the power of the ensembled model in predicting recession probabilities at different rates of TP and FP. The AUC – Area Under the ROC Curve – is a measure of the model's ability to distinguish between the two classes and is inherently tied to the ROC curve. The figure below shows the outcome:

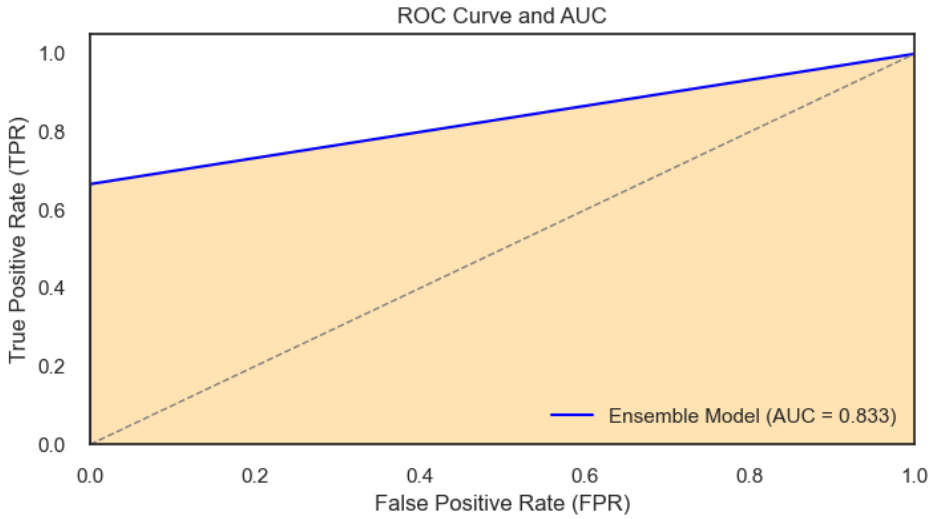


Figure 22. ROC curve with the AUC for the Ensembled Model.

The ROC curve for the ensemble model exhibits a good performance with an AUC of 0.83. An AUC value closer to 1 than to 0.5 indicates that the model has a strong capacity to measure of separability. Specifically, the AUC of 0.83 suggests that the model can almost correctly distinguish between the positive class and the negative class.

The ROC curve and AUC indicate that the ensemble model is effective for the task at hand, providing a good tool for decision-making processes where accurate classification is necessary. Further evaluation in a real-world setting or across multiple data sets would be beneficial to confirm the model's robustness and to fine-tune its threshold to maximize both precision and recall according to domain-specific requirements.

The table below shows the performance summary for the five models used in this project.

Table 5. Performance summary in terms of accuracy of the models.

	Model	Accuracy(%)	AUC(%)
0	LR	92.57	79.4
1	GNB	93.24	80.7
2	RF	97.30	83.0
3	XGB	97.97	82.9
4	Ensemble	97.30	83.3

 doi:10.20944/preprints202407.2620.v1

From the summary, we observe that logistic Regression (LR) shows relatively lower accuracy and AUC compared to other models, suggesting less predictive strength. The Gaussian Naive Bayes (GNB) has slightly better performance, with small increases in both accuracy and AUC. The Random Forest (RF) and Extreme Gradient Boosting (XGB) models achieve much higher accuracy (above 97%), with RF also having the highest AUC, indicating a strong ability to distinguish between classes. The Ensemble model, matches the RF's accuracy but with a slightly higher AUC, which might suggest better generalization. Interestingly, despite XGB having the highest accuracy, its AUC is lower than that of the RF and Ensemble models, indicating that while it may predict correctly most of the time, it may not rank the positive cases as consistently higher than the negative cases as the RF and Ensemble models do. The observations justify the choice of the ensemble model for performing the out-of-sample predictions in this paper.

4.5. Forecast Results for the Probability of Economic Recession

This section provides a detailed analysis of the potential likelihood of economic downturns in the forthcoming one year period using five machine learning models. First, we will present reports of the economic crisis prediction results for the four individual models. Then we present the results of using all of the four models called the ensemble model to predict economic crisis. We plot the predicted probability of the U.S. economic crisis for 2024 for each model in yellow line along with the real probability of economic crisis of 2022-2023 in blue line.

4.5.1. Logistic Regression

The results from the logistic regression were shown below.

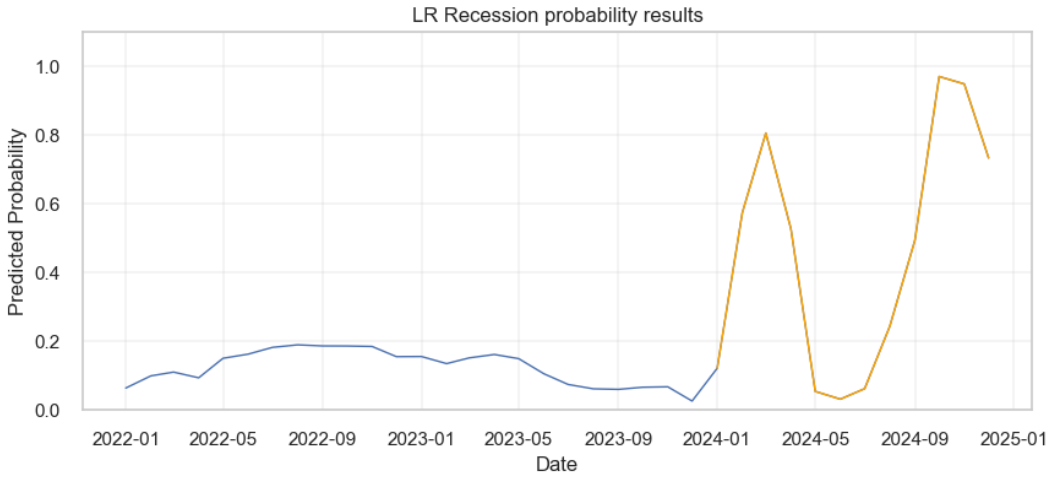


Figure 23. Probability of Economic crisis according to logistic regression.

The Figure 29 shows that from early 2022 to the end of 2023, the probability of a recession is consistently low, mostly under 0.2 as shown by the blue line. However, there are two prominent spikes in the model's predictions based on the one year forecast data as indicated by the orange line: one around March 2024 and another around October, 2024, with probabilities surging to above 0.8 and then dropping back down. This suggests that, according to the model, the risk of recession was significantly higher during these two periods.

Gaussian Naive Bayes

The results from the Gaussian Naive Bayes were as shown below.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2024 doi:10.20944/preprints202407.2620.v1

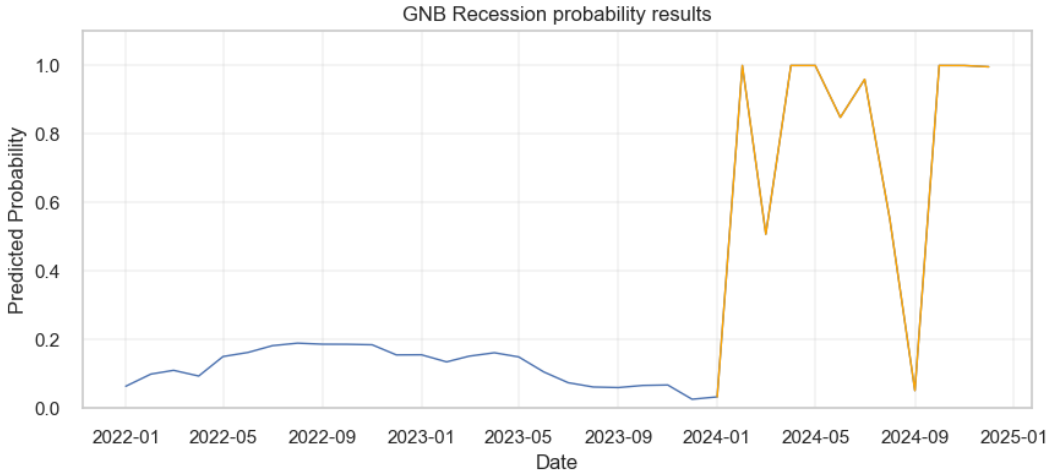


Figure 24. Probability of Economic crisis according to Gaussian Naive Bayes Model.

In Figure 29, the blue line shows the probabilities of recession from the actual data. In contrast to the logistic regression model, from early 2024 onward, the Gaussian Naive Bayes model displays a more volatile prediction pattern as shown by the orange line, with several sharp peaks and troughs indicating fluctuating probabilities of a recession. These probabilities reach as high as 1.0 at times, suggesting periods of very high predicted recession risk according to this model, but these spikes are brief with some predictions lasting less than a month.

III. Random Forests

The results from the Random Forests were as shown below.

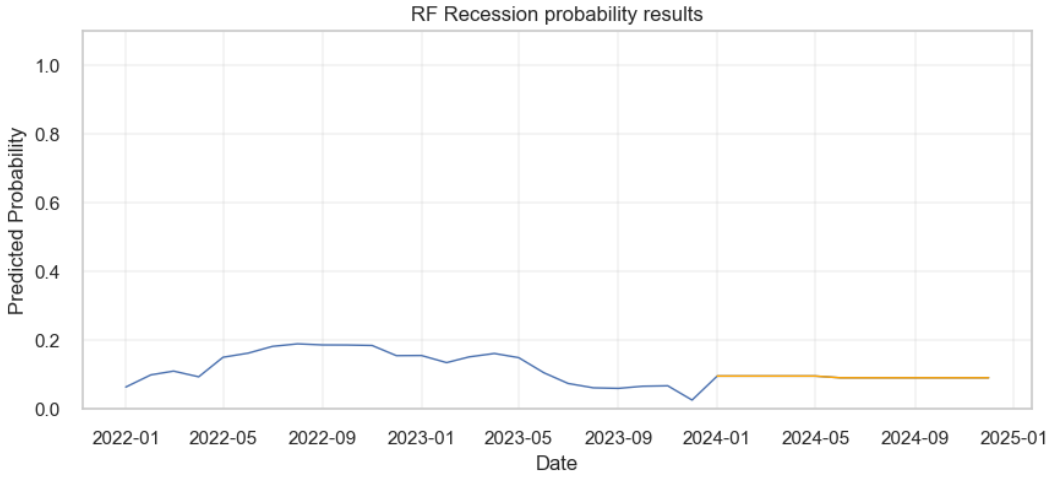


Figure 25. Probability of Economic crisis according to Random Forests Model.

The prediction results from the Random Forest (RF) model indicate a consistently low probability of recession from the beginning of 2022 through to early 2024, with the probability largely hovering under 0.2 as shown by the blue and orange lines. There's a slight increase observed around mid-2023, but it remains below 0.25. From early 2024 onward, the model shows a remarkably flat prediction, with the probability of recession being constant and minimal, close to zero. This suggests a stable outlook according to the Random Forest model, with no significant risk of recession predicted beyond early 2024.

IV. eXtreme Gradient Boosting

The results from the XGBoost were as shown below.

 doi:10.20944/preprints202407.2620.v1

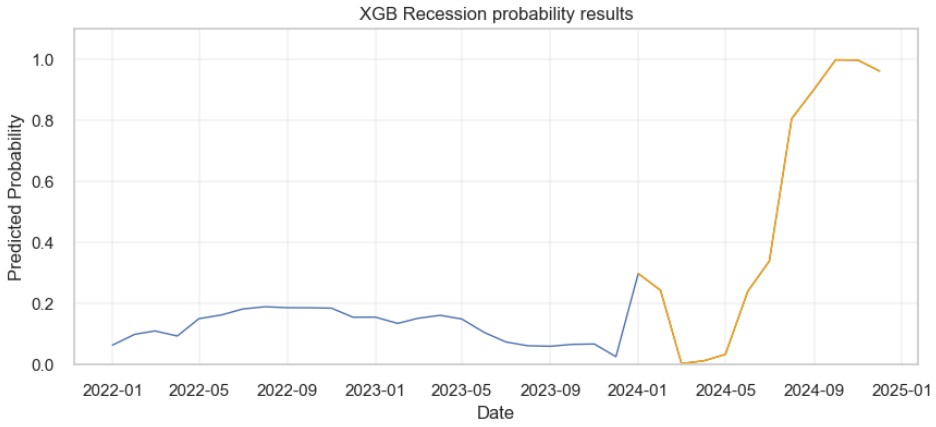


Figure 26. Probability of Economic crisis according to XGBoost Model.

According to the XGboost model's prediction results, the probability of a recession remains low and relatively stable from the beginning of 2022 up to early 2024, with a slight uptick around mid-2023 (blue line). From early 2024, the model predicts (as shown by the orange line), a sharp increase in the probability of a recession, reaching a peak just below 1.0 around October 2024. Following this peak, the model suggests a slight decrease but then maintains a high probability of a recession towards the end of 2024, indicating a period of high recession risk.

4.6. Applying the Ensemble Model to Predict U.S. Economic Recession

After building, training, testing and evaluating the ensemble model, the next step was to apply it on a real life situation. The forecasts from section 3, subsection 3.2.5 were used to generate the predictions in terms of probabilities. First, we plotted the general results for the entire data set including the historical data and one year forecast for the indicators . The results were as shown below:

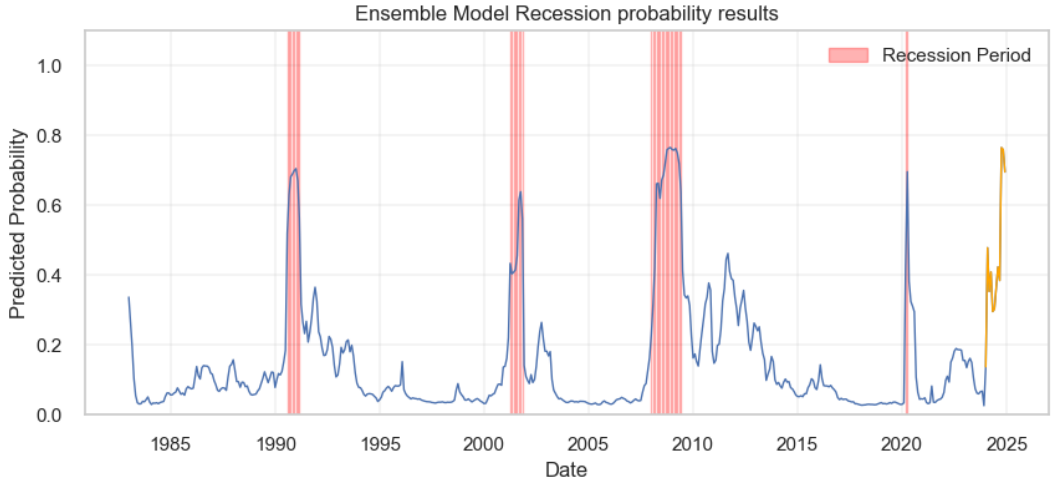


Figure 27. Prediction Results for the entire Data set.

The blue line represents the predicted likelihood of a recession occurring at different points in time, with several notable peaks corresponding to past recessions, as highlighted by the vertical red bands. The periods covered by these bands indicate times of economic downturn, as recognized by the model's historical data. Moving into the future, the forecast period is marked in orange, starting early 2024, and shows a significant spike in the recession probability at different months. This suggests that, based on the model's indicators, there is a considerable risk of a recession occurring in 2024 at specific months.

 doi:10.20944/preprints202407.2620.v1

To get a clear understanding of when the recessions were likely to happen according to the model, we plotted the prediction probabilities based on the last five years of the historical data and concatenated this with the 1-year out of sample forecasts. The outcome was as shown.

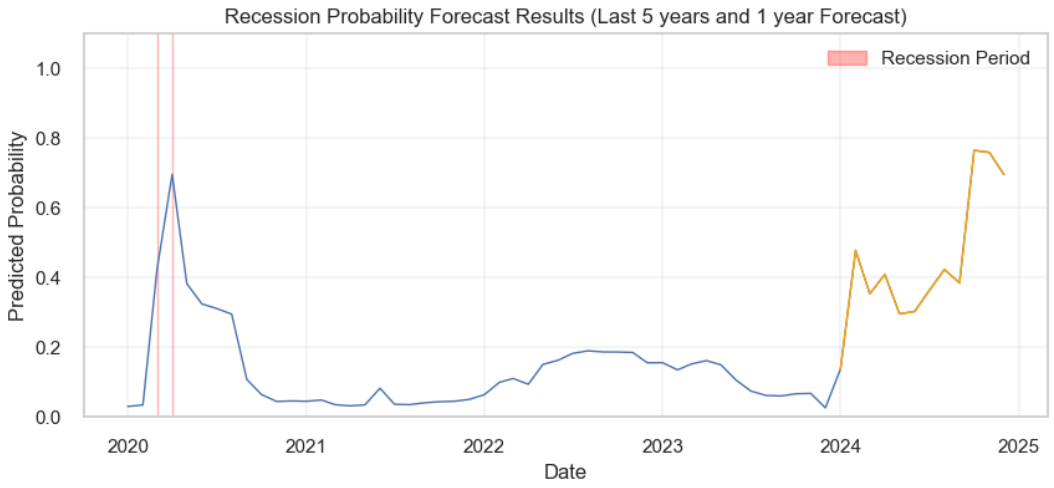


Figure 28. Results for the Past 5 years and 1-year Economic Recession Probability Prediction .

The graph shows the recession probabilities over the last five years, along with a 1-year forecast into the future, according to the ensemble predictive model. The historical data, represented by the blue line, shows a peak around 2020, likely indicating the economic impact of COVID-19. Following this, the recession probability declines and remains relatively low until the beginning of the forecast period, which is depicted in orange. The forecast predicts a volatile period with two pronounced peaks suggesting a heightened probability of recession occurring around January/February and October/November, 2024. The red bands highlight the actual recession periods, aligning with the peaks in predicted probabilities, thereby validating the model's effectiveness in historical data.

The figure below shows the specific months for the predicted recession dates for the year 2024.

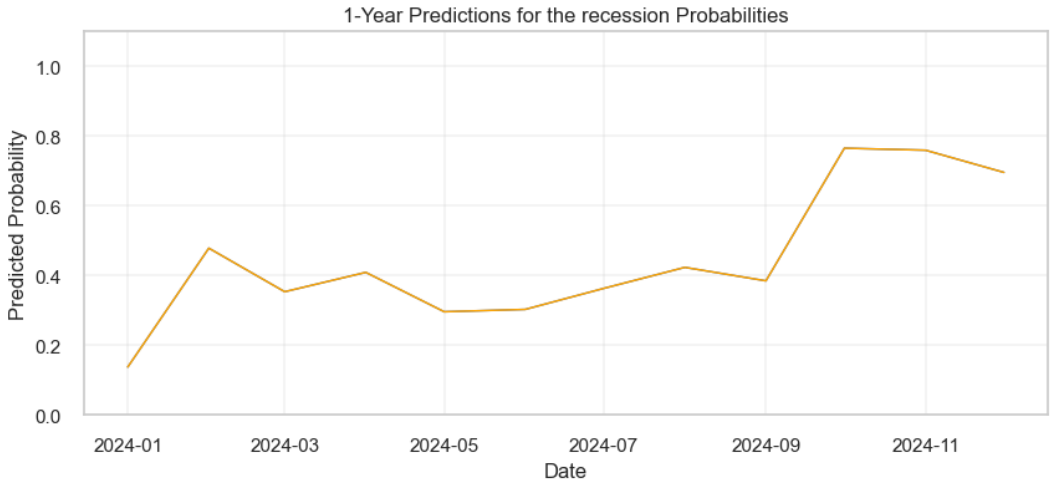


Figure 29. Results for the 1-year Economic Recession Probability Prediction.

5. Conclusions

The main objective of the study was to find, apply, and test machine learning models to predict the U.S. economic recessions. The results from section 4 indicated that the main objective of the study was achieved. In this paper, we used five mathematical models namely: logistic regression, naive bayes, random forests, xgboost, and ensemble model.

 doi:10.20944/preprints202407.2620.v1

All the models were diverse, bearing different strengths, weaknesses and capabilities. The logistic regression and the Naive Bayes performed better with transformed data compared with the other models. On the other hand, the non-linear models do not require a stationary input data, thus the data transformation is not needed. The ensemble model produced better predictions for the U.S. recession probabilities for the year 2024. From these results, we conclude that Model's average is the best to apply for recession detection, with AUC = 0.83. Compared to the results from the most recent studies, this value is good, and ranks among the best scores. A probability greater than 0.5 signifies an alert for a possible economic downturn.

Despite the good accuracy results reported from the averaged model, other metrics like recall and ROC curve suggests that accuracy alone should not be used as the ultimate measure of efficiency of a model. The averaged model has a recall of 0.67 and an accuracy of 97.3% on a sample size of 492 indicators. Although these values are good, each model need to be analyzed independently so as to avoid false conclusions. For instance, logistic regression is known to produce the best results when used with independent data. For effective application with the time series data, [14] suggests a proper choice of the macro indicators so as to avoid highly correlated variables. The non-linear models did not require any transformations, and this may raise concerns over the type of transformations required to achieve the best results. Besides, indicators were also predicted using other time series techniques that have their strengths and limitations. All these factors have influenced the final outcomes.

In the final stages of the training results analysis, the non-linear models presented a way to visualize the importance of the variables towards the predictions. Even though these outcomes could possibly change as we transition from one model to another, they offer insights that could be used to guide model tuning and feature engineering to enhance performance. Also, the differences observed in the performance scores may be crucial for making decisions about which model may perform better, or more reliably for a given prediction or set of prediction tasks. The outcomes may also be useful to the future users in understanding recessions.

The work in this paper was thoroughly evaluated using contemporary metrics in machine learning. The evaluation results offer insights into prediction of U.S. economic recessions with some degree of confidence. The outcome also shows that selection of macroeconomic indicators is a key step towards application of machine learning algorithms in economic recession detection. We also ascertain that publicly available data can be used to build useful models. The models could also be modified for use by multinational companies to evaluate their economic progress and make rightful decisions ahead of time before economic crises.

References

1. Mukhamediev, R.I.; Popova, Y.; Kuchin, Y.; Zaitseva, E.; Kalimoldayev, A.; Symagulov, A.; Levashenko, V.; Abdoldina, F.; Gopejenko, V.; Yakunin, K.; others. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. *Mathematics* **2022**, *10*, 2552.
2. Vrontos, S.D.; Galakis, J.; Vrontos, I.D. Modeling and predicting US recessions using machine learning techniques. *International Journal of Forecasting* **2021**, *37*, 647–671.
3. Barbosa, R. Ensemble of Machine Learning Algorithms for Economic Recession Detection. 2018.
4. Hwang, Y. Forecasting recessions with time-varying models. *Journal of Macroeconomics* **2019**, *62*, 103153.
5. Miller, D.S. Predicting Future Recessions **2019**.
6. Psimopoulos, A. Forecasting economic recessions using machine learning: an empirical study in six countries. *South-Eastern Europe Journal of Economics* **2020**, *18*, 40–99.
7. Berge, T.J. Predicting recessions with leading indicators: Model averaging and selection over the business cycle. *Journal of Forecasting* **2015**, *34*, 455–471.
8. Chikako, B.; Turgut, K. Predicting Recessions: A New Approach For Identifying Leading Indicators and Forecast Combinations. Technical report, IMF Working Paper, Monetary and Capital Markets Department, 2011.

 doi:10.20944/preprints202407.2620.v1

9. Ali, B.J.; Anwar, G. Marketing Strategy: Pricing strategies and its influence on consumer purchasing decision. *Ali, BJ, & Anwar, G.(2021). Marketing Strategy: Pricing strategies and its influence on consumer purchasing decision. International journal of Rural Development, Environment and Health Research* **2021**, *5*, 26–39.
10. Kumari, R.; Srivastava, S.K. Machine learning: A review on binary classification. *International Journal of Computer Applications* **2017**, *160*.
11. Chauvet, M.; Potter, S. Forecasting recessions using the yield curve. *Journal of forecasting* **2005**, *24*, 77–103.
12. Jiang, T.; Gradus, J.L.; Rosellini, A.J. Supervised machine learning: a brief primer. *Behavior Therapy* **2020**, *51*, 675–687.
13. Kakde, Y.; Agrawal, S. Predicting survival on titanic by applying exploratory data analytics and machine learning techniques. *International Journal of Computer Applications* **2018**, *179*, 32–38.
14. Ranganathan, P.; Pramesh, C.; Aggarwal, R. Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research* **2017**, *8*, 148–151.