Recession Forecasting with Logit and Random Forests

Macroeconometrics and Machine Learning Project

Maxime Coulet, Benjamin Lengereau, Kilian Guillon

January 2025

Abstract

This report explores ways to enhance the predictive power of a logit model in identifying U.S. recessions. We compare different sets of regressors: classic macroeconomic indicators (e.g., GDP, unemployment), factors from a PCA on the full FRED-MD dataset, and factors estimated using a Bayesian Sparse Factor Model. Finally, we examine how a Random Forest can further improve the model's predictions.

Introduction

Economies experience fluctuations characterized by alternating phases of growth and decline. These recurring patterns, known as business cycles, shape economic performance for multitude of sectors. The period in which the economics activity decline is known as Recession. Economists have put a great deal on definition, identification and prediction of such periods. While a common rule of thumb define recession periods from at least two consecutive quarters of negative GDP growth, the National Bureau of Economic Research (NBER) define recessions as a significant decline in economic activity spread across the economy, lasting more than a few months, visible in real GDP, real income, employment, industrial production, and wholesale-retail sales.

Machine learning methods such as Logit regression, random forests, and Bayesian classifiers have been extensively tested for predicting recession periods. In this study, we focus on the selection of key regressors using techniques such as Principal Component Analysis (PCA) and Bayesian Sparse Factor models, leveraging the FRED-MD dataset for analysis. With around 120 variables in the FRED-MD dataset, PCA and factor models extract key factors, summarizing information and mitigating the curse of dimensionality in Logit estimation. We compare and extend our approach by using random forests to select and average performance across different sets of regressors.

1 Methodology

1.1 Logit Model

We use a logit model to predict recession periods. The probability of a recession occurring at time t is modeled as:

$$P(\mathrm{y}_t = 1 \mid X_{t-1}) = \sigma(\omega_0 + X'_{t-1}\omega)$$

where y_t takes 1 if the period is a recession, X_t is set of regressors, ω the coefficient vector, ω_0 the intercept term, and $\sigma(\cdot)$ the sigmoid function.

We first estimate a standard logit model. To classify recession periods, we apply a threshold of 0.5 on the estimated probability. We then introduce a weighted loss function to deal with the imbalance in recession occurrences in our dataset. This approach attributes a higher loss to the misclassification of recession periods. We combine this loss function with a Lasso regularisation to perform variable selection. Given $y_1, \dots, y_N$ observations, the loss minimisation problem then becomes :

$$\min_{\omega} \quad -\sum_{t=1}^{T} w_{y_t} \left[y_t \log \sigma(\omega_0 + X_{t-1}'\omega) + (1 - y_t) \log(1 - \sigma(\omega_0 + X_{t-1}'\omega)) \right] + \lambda \sum_{j=1}^{p} |\omega_j|$$

where $w_{y_t} = \frac{N}{2N_{y_t}}$ with N_{y_t} the number of occurrences of y_t in the observations and λ is a penalty strength parameter Throughout this project, we will enhance the logit model's predictive power by estimating it on factors derived from Principal Component Analysis (PCA) and a Bayesian Sparse Factor Model.

1.2 Principal Component Analysis and Dimensionality Reduction

Principal Component Analysis is a statistical dimensionality reduction method that transforms a set of correlated variables into a new set of uncorrelated variables called principal components. Its goal is to extract the most relevant information from the data while reducing model complexity. PCA works by computing the eigenvectors and eigenvalues of the covariance matrix of the data, identifying the directions (principal components) along which the variance is maximized. The first principal component captures the highest variance, the second captures the next highest variance while being orthogonal to the first, and so on.
In our study, PCA helps condense the numerous macroeconomic variables from the FRED database into a smaller set of components that capture the most significant variance in the data. This approach offers several advantages: it reduces noise, limits multicollinearity issues among predictors, and, most importantly, mitigates the problem of high dimensionality. Working with too many variables can lead to overfitting, increased computational costs, and reduced interpretability. By lowering the dimensionality of the data, PCA helps make the model more robust and efficient.

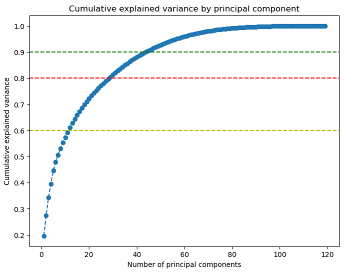

Figure 1: Cumulative explained variance by PCA factors

To assess the impact of the number of components chosen, we test different levels of explained variance, retaining 60%, 80%, and 90% of the total variance. For 60% of variance explained, there are 12 principal components, for 80% there are 29, and for 90% there are 44. This allows us to analyze how the amount of retained information affects the model's ability to predict recessions, striking a balance between complexity and performance. The relationship between the number of principal components and the explained variance can be observed in Figure 1.

1.3 Sparse factor

In macroeconomics, static factor models are less popular than dynamic factor models to identify common movements among variables. Initially, we considered dynamic factor models, but overcoming the high-dimensionality problem remains a significant challenge, both for us and in recent research. In contrast, the bayesian literature on static factor models offers a more developed framework. Given a set of N observations $Y = (y_1, \ldots, y_N)$, where $(y_i)_{i=1,\ldots,N}$ are G-dimensional vectors of variables, a bayesian formulation of a latent factor model is

$$\begin{aligned} \textbf{Likelihood:} \quad & y_i | \omega_i, B, \Sigma \overset{i.i.d.}{\sim} \mathcal{N}_G\left(B\omega_i, \Sigma\right), \\ \textbf{Latent Factors:} \quad & \omega_i \overset{i.i.d.}{\sim} \mathcal{N}_K\left(0_K, I_K\right) \end{aligned} \tag{1}$$

where B is the loading matrix, ω_i are latent factors with a normal prior. A major flaw of this formulation is that B is identifiable only up to a right orthogonal transformation. Ročková and George (2016) have proposed the use of Spike-and-slab prior on the loading matrix to enhance the identifiability of B.

$$\begin{aligned} B_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1 &\overset{\text{ind}}{\sim} (1-\gamma_{jk})\text{Laplace}(\lambda_0) + \gamma_{jk}\text{Laplace}(\lambda_1), \\ \gamma_{jk} &\overset{\text{ind}}{\sim} \text{Bernoulli}\left(\prod_{l=1}^{k} \nu_l\right), \\ \nu_l &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1), \end{aligned} \tag{2}$$

where $\lambda_0 \gg \lambda_1$ are scale parameters, and γ_{jk} is a latent binary indicator stored in the $G \times K$ feature allocation matrix Γ. Because $\lambda_0 \gg \lambda_1$, the first Laplace distribution has a significantly smaller variance than the second, making it correspond to the spike prior centered around 0, while the second represents the diffuse slab prior. Such priors impose constraints on the loadings to shrink coefficients corresponding to factors with negligible effects on a variable. The mode of the posterior can be found iteratively with an Expectation-Maximization algorithm. This involves solving a LASSO objective function to estimate the loadings, updating the latent factors, and repeating the process until convergence. Ročková and George (2016), propose an augmented parameter version of the basic Expectation-Maximization algorithm to enhance convergence speed and estimation performance with the downside of losing monotonicity in the iterated error.

1.4 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. Each tree is trained on a different bootstrap sample of the data, and at each split, a random subset of features is considered to introduce diversity among the trees. This randomness helps prevent individual trees from overfitting to the training data, making Random Forest a robust choice for classification tasks such as recession prediction. The final prediction is determined by majority voting among all trees in the forest.

To optimize model performance, we perform hyperparameter tuning using cross-validation. We focus on several key hyperparameters: the number of trees in the forest is set to 1000 to ensure stable predictions. The maximum depth of each tree is evaluated with values of None (unlimited depth), 10, and 30 to strike a balance between complexity and the risk of overfitting. Additionally, we consider the minimum number of samples required to split an internal node, varying this between 2, 5, and 10 to control how the trees grow. Lastly, we adjust the minimum number of samples required in a leaf node, testing values of 1 and 4 to prevent overly small leaf nodes that may capture noise.

By applying cross-validation, we ensure that our selected hyperparameters generalize well to unseen data rather than being overly optimized for a specific subset. This approach improves the model's reliability in predicting recessions based on macroeconomic indicators. Additionally, we will compare the performance of the Random Forest model with that of the Logit model to evaluate which method yields better predictions for economic recessions.

2 Results

2.1 FRED-MD dataset

We use the FRED-MD monthly macroeconomic database, last updated in December 2024, which contains 134 monthly U.S. macroeconomic indicators spanning the period from 1960 to 2024. These indicators cover output and income, labor market conditions, housing, consumption, orders, inventories, money and credit, interest and exchange rates, prices, and the stock market. Recession periods are identified using the NBER-based recession indicator which classifies each month as either a recession (1) or non-recession (0) period. The NBER business cycle dating is based on a subjective assessment of a variety of indicators. To ensure comparability across variables, the dataset applies various transformations to the raw time series [1].

2.2 Logit results

Our first logit model is estimated on an arbitrary set of 11 regressors: Real Personal Income, Civilian Unemployment Rate, S&P 500 Strock Price Index, Real Manufacturing and Trade Industries Sales, All Employees in Goods-Producing Industries, Average Weekly Hours in Manufacturing, 3-Month Treasury C Minus FEDFUNDS, Average Hourly Earnings in Goods-Producing industries, IP Index, and the Inventories to Sales Ratio in Total Businesses.

The results of the Logit model on this set of regressors, presented in Table 1, show a high accuracy of 94% . However the model has poor performances in predicting recession periods (class 1) with a recall of 53% showing this model misses 47% of the recession period.

[1] See the FRED-MD appendix

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.95	0.98	0.96	139
1 (Recession)	0.73	0.53	0.62	15
Accuracy			0.94	154
Macro Avg	0.84	0.76	0.79	154
Weighted Avg	0.93	0.94	0.93	154

Table 1: Classification report for the Logit model on a arbitrary set of regressors

The results of the Logit model with a weighted loss and Lasso penalty on the same set of regressors, presented in Table 2, show an almost similar overall accuracy of 93%, but with a significant improvement in recession recall to 87%. The F1-score for class 1 increases to 70%.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.98	0.94	0.96	139
1 (Recession)	0.59	0.87	0.70	15
Accuracy			0.94	154
Macro Avg	0.79	0.90	0.83	154
Weighted Avg	0.95	0.93	0.93	154

Table 2: Classification report for the Logit model on a arbitrary set of regressors with weighted loss and Lasso penalty

Next, we estimate a logit model using PCA factors, which reduce dimensionality while preserving key information. Table 3 presents the results for PCA factors explaining 60% of the variance. All 12 factors were preserved after applying L1 regularization. The accuracy remains high at 92%, and the model achieves a recall of 100% for recessions, meaning all actual recession periods were correctly identified. However, the precision for class 1 is lower, at 61%, indicating some false positives.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	1.00	0.91	0.95	135
1 (Recession)	0.61	1.00	0.76	19
Accuracy			0.92	154
Macro Avg	0.81	0.96	0.86	154
Weighted Avg	0.95	0.92	0.93	154

Table 3: Classification report for the Logit model with PCA (60% variance explained)

Increasing the number of PCA factors to explain 80% of the variance, the results in Table 4 show a slight drop in performance. Here, L1 regularization discarded one of the 29 factors, leaving 28 factors in the final model. The accuracy decreases to 90%, and the recall for class 1 remains high at 89%, but with a lower F1-score of 69%. This suggests that adding more factors may introduce noise rather than improving predictive power.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.98	0.90	0.94	135
1 (Recession)	0.57	0.89	0.69	19
Accuracy			0.90	154
Macro Avg	0.78	0.90	0.82	154
Weighted Avg	0.93	0.90	0.91	154

Table 4: Classification report for the Logit model with PCA (80% variance explained)

Finally, we estimate a logit model using sparse factors extracted via a Bayesian Sparse Factor Model. The L1 regularization retained all 10 factors. The results, shown in Table 5, indicate a slight decrease in the accuracy compared to the 80% PCA model, with an accuracy of 89% and an F1-score of 74% for class 1. The recall remains high at 89%.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.98	0.89	0.93	135
1 (Recession)	0.53	0.89	0.67	19
Accuracy			0.89	154
Macro Avg	0.76	0.89	0.80	154
Weighted Avg	0.93	0.89	0.90	154

Table 5: Classification report for the Logit model with sparse factors

2.3 Random Forest results

We will now compare the results of the logit model with a random forest. The random forest allows us to better process complex, non-linear relationships between macroeconomic variables and recession probabilities. Additionally, random forests can handle a large number of predictors without requiring strong assumptions about their distribution or functional form. However, the logit model remains a benchmark due to its interpretability. Comparing both models will help assess whether the added flexibility of the random forest improves predictive performance or if the simpler, more interpretable logit model remains sufficient.

The results of the Random Forest model on all covariates, presented in Table 6, show a high accuracy of 94%. While the model excels in predicting the absence of recession (class 0), with a precision of 0.95 and a recall of 0.98, it demonstrates more modest performance in predicting recession periods (class 1), with a precision of 0.80 and a recall of 0.63. This means that when the model predicts a recession, it is correct 80% of the time, but it does not always identify actual recession periods.

To improve performance in detecting recessions, it may be beneficial to reduce the dimensionality of the dataset or explore rebalancing techniques, such as oversampling recession periods. Adjusting class weights in the loss function could also help better capture recessions.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.95	0.98	0.96	135
1 (Recession)	0.80	0.63	0.71	19
Accuracy		0.94 (154 samples)		
Macro Avg	0.87	0.80	0.83	154
Weighted Avg	0.93	0.94	0.93	154

Table 6: Classification report for the Random Forest model

Applying Principal Component Analysis (PCA) with 60% variance explained, as shown in Table 7, leads to a slight increase in model performance compared to using all features. While the model classifies non-recession periods (class 0) with 94% precision and 99% recall, its performance on recession periods (class 1) shows improvement, with a precision of 92% and a recall of 58%. This indicates that when the model predicts a recession, it is correct 92% of the time, but it fails to detect 42% of actual recession periods.

Comparing this to the model trained without PCA, we observe that dimensionality reduction positively impacts predictive performance, particularly for recession detection. Reducing the number of features through PCA may eliminate some macroeconomic information that is not crucial for distinguishing between stable and recession periods. Among the tested PCA thresholds (60%, 80%, and 90% variance explained), the 60% threshold yields the best results, suggesting that retaining only the most significant components helps prevent overfitting while maintaining relevant economic patterns. Increasing the variance explained to 80% or 90% likely introduces unnecessary complexity without improving recession detection.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.94	0.99	0.97	135
1 (Recession)	0.92	0.58	0.71	19
Accuracy		0.94 (154 samples)		
Macro Avg	0.93	0.79	0.84	154
Weighted Avg	0.94	0.94	0.94	154

Table 7: Classification report for the Random Forest model with PCA (60% variance explained)

When using sparse factors, we obtain similar results to those achieved with the random forest on all covariates, slightly below those of the random forest with PCA. Sparse factors reduce dimensionality more than PCA, which limits the predictive capacity of the model here.

We therefore train a random forest using the sparse factors and covariates of the FRED dataset. We obtain slightly better results, table 8 but not as much as with PCA.

Class	Precision	Recall	F1-score	Support
0 (No Recession)	0.94	0.99	0.96	135
1 (Recession)	0.85	0.58	0.69	19
Accuracy		0.94 (154 samples)		
Macro Avg	0.89	0.78	0.83	154
Weighted Avg	0.93	0.94	0.93	154

Table 8: Classification report for the Random Forest model with sparse factors and covariates

Conclusion

We were able to develop two models to predict the state of the U.S. economy using macroeconomic indicators. Feature engineering methods, which reduce dimensionality, proved to be very useful in improving the results. Moreover, the two models—Logit and Random Forest—differ from each other, with the former achieving a higher recall and the latter a higher precision. Nonetheless, the present study faces several limitations that may affect the predictive performance of the models. First, the prediction framework currently relies on using data from t-1 to predict a recession in t. A more effective approach would be to incorporate all previous periods, allowing the model to capture a more comprehensive view of the economic indicators and enhancing its ability to predict recessions. Additionally, the application of Principal Component Analysis reduces dimensionality in a non-targeted manner, which contrasts with variable selection methods that can focus on the most relevant features for the task at hand. This can result in the loss of important information that is crucial for accurate predictions. Furthermore, there is a significant imbalance between recession and non-recession periods in the dataset, with insufficient instances of recessions to provide optimal training. This imbalance leads the models to favor predicting non-recession periods, thereby potentially compromising their ability to detect actual recession events effectively.

References

Ročková, V. and E. I. George (2016). “Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity”. In: *Journal of the American Statistical Association* 111.516, pp. 1608–1622. DOI: 10.1080/01621459.2015.1100620. eprint: https://doi.org/10.1080/01621459.2015.1100620. URL: https://doi.org/10.1080/01621459.2015.1100620.